**Robustly Learning Multi-Index Models**

by

Nikos Zarifis

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2025

Date of final oral examination: 07/01/2025

The dissertation is approved by the following members of the Final Oral Committee:

Ilias Diakonikolas, Professor, Computer Sciences

Jelena Diakonikolas, Assistant Professor, Computer Sciences

Sebastien Roch, Professor, Mathematics

Jerry Zhu, Professor, Computer Sciences

# Acknowledgments

This dissertation would not have been possible without the tremendous support, encouragement, and inspiration of many people to whom I owe my deepest gratitude.

First and foremost, I want to express my deepest gratitude to my advisor Ilias Diakonikolas for his unwavering support throughout my Ph.D. journey. I feel incredibly fortunate to have had him as my mentor. Ilias has guided me with patience and taught me how to approach difficult research problems. His invaluable feedback and vision were important in my development as a researcher. I am leaving this program equipped with the tools to succeed thanks to him, and I am looking forward to continuing to work with him in the future.

I am also very grateful to work closely with Jelena Diakonikolas. Jelena devoted a lot of hours working with me and advising me. In many ways, Jelena became my second advisor, always pushing me to become a better researcher. I hope that we continue our collaboration in the future.

I would like to thank Sebastien Roch and Jerry Zhu for being on my committee and for their helpful comments and suggestions. I am grateful to Christos Tzamos, whose collaboration and support played a significant role during my Ph.D.

Furthermore, I wish to acknowledge all my collaborators: Christos, Daniel, Vasilis, Puqian, Giannis, Thanasis, Lisheng, Sushrut, Ankit, Jasper, Ali, Yuxin and Mingchen. I learned a lot from collaborating with each one of them and hope to continue working together.

During my Ph.D., I made several friends in Madison that made my life much better: Vasiliki, Vasilis, Elena, Takis, Andreas, Eleni, Vasilis, Eleana, Angeliki, Elisa, Thanasis, Giannis and Konstantinos. I want to thank all of them for their friendship and support. I want especially to thank Vasilis and Elena. They were always there for me and supported me and made the first years of my Ph.D. much better.

I also want to thank my friends in Greece who were always there for me and supported me during my Ph.D. journey (in no particular order): Nikos, Vasilis, Argyris, Dimitris K., Dimitris G., Petros, Angeliki, Thanasis, Katerina, Betty, Aspa, Sotiria, Nikos, Giorgos, Vasilis, Christoforos, Konstantina, Vardis, Grigoris, Anna, Evi and Kostas. Their encouragement helped me feel closer to home even when I was far away.

Most importantly, I am deeply grateful to my parents, Vasilis and Ritsa, my sibling,

Marialena for always supporting me, believing in me, and loving me unconditionally. Everything I have achieved comes from their sacrifices, support, and love. I will be forever thankful for all they have done for me.

# Contents

**1 Introduction** **1**
1.1 The Challenge of Learning in the Presence of Noise . . . 1
1.2 A Unifying Theme: Learning Low-Dimensional Structure in High Dimensions . . . 2
1.3 Main Contributions and Thesis Outline . . . 4
1.3.1 Learning Gaussian Halfspaces with General Massart Noise . . . 5
1.3.2 Learning Margin Halfspaces with Massart Noise . . . 15
1.3.3 Agnostic Proper Learning Halfspaces under Gaussian Distribution 21
1.3.4 Learning Generalized Linear Models under Well-Behaved Distributions . . . 27
1.3.5 Robustly Learning Single-Index Models under Well-Behaved Distributions . . . 34
1.3.6 Learning Generalized Linear Models with Gaussian Marginals . 42
1.3.7 Learning Single-Index Models with Gaussian Marginals . . . 49
1.3.8 Agnostically Learning Multi-Index Models with Queries . . . 57
1.3.9 Robustly PAC Learning Multi-Index Models under the Gaussian Distribution . . . 76
1.3.10 SQ Lower Bounds Using Linear Programming . . . 92
1.4 Preliminaries . . . 106
1.4.1 Miscellaneous Notation . . . 106
1.4.2 Probability Basics . . . 107
1.4.3 Gaussian Space and Hermite Polynomials . . . 108
1.4.4 Preliminaries: Multilinear Algebra . . . 111
1.4.5 Statistical Query Framework . . . 112

**I Learning Halfspaces with Noise** **117**

**2 Learning Halfspaces with Semi-Random Noise** **118**
2.1 Introduction . . . 118

2.2 Detailed Technical Overview . . . 124
2.2.1 Learning General Halfspaces with Constant-Bounded Massart Noise: Theorem 2.1.2 . . . 124
2.2.2 Learning Halfspaces with General Massart Noise: Theorem 2.1.5 130
2.2.3 SQ Lower Bounds: Theorems 2.1.3 and 2.1.6 . . . 136
2.3 A Certificate-Based Algorithm For Learning General Halfspaces with Constant-Bounded Massart Noise . . . 140
2.3.1 The Low-Degree Polynomial Certificate . . . 141
2.3.2 Efficiently Computing the Certificate via SDP . . . 167
2.3.3 Learning a Near-Optimal Halfspace via Online Convex Optimization 173
2.4 Learning Halfspaces with General Massart Noise . . . 180
2.4.1 The Sign-Matching Polynomial . . . 182
2.4.2 Projecting onto $\mathbf{w}^{\perp}$ . . . 187
2.4.3 Using the Low-Order Chow Tensors . . . 192

**3 Learning Margin Halfspaces with Massart Noise 202**
3.1 Introduction . . . 202
3.2 Our Algorithm and its Analysis: Proof of Theorem 3.1.2 . . . 205
3.3 Conclusions and Open Problems . . . 218

**4 Learning Agnostically Proper Halfspaces 220**
4.1 Introduction . . . 220
4.2 Proper Agnostic Learning Algorithm . . . 224
4.2.1 Analysis of Algorithm 5: Proof of Theorem 4.1.1 . . . 225
4.2.2 Proof of Proposition 4.2.1 . . . 234
4.3 Agnostic Proper PTAS for Homogeneous Halfspaces . . . 240
4.3.1 Analysis of Algorithm 6: Proof of Theorem 4.1.2 . . . 241
4.3.2 Proof of Proposition 4.3.3 . . . 246
4.4 Agnostic Proper Learning of ReLUs . . . 250

**II Robustly Learning Monotone Single-Index-Models 263**

**5 Robustly Learning Monotone GLMs under Well-Behaved-Distributions 264**
5.1 Introduction . . . 264
5.2 The Landscape of the $L_2^2$ Loss . . . 270
5.2.1 Proof of Proposition 5.2.6 . . . 294
5.2.2 Proof of Theorem 5.2.4 . . . 300
5.3 The Landscape of the $L_2^2$-Loss for Unbounded Activations . . . 304
5.4 Optimizing the Empirical $L_2^2$-Loss . . . 309
5.4.1 Sigmoidal Activations . . . 310

5.4.2 Unbounded Activations . . . 324

**6 Robustly Learning Monotone SIMs under Well-Behaved-Distributions 333**
6.1 Introduction . . . 333
6.1.1 Local Error Bounds and Alignment Sharpness . . . 335
6.2 Main Structural Result: Alignment Sharpness of Surrogate Loss . . . 337
6.2.1 $L_2^2$ Error and Misalignment . . . 339
6.2.2 Closeness of Idealized and Attainable Activations . . . 347
6.2.3 Proof of Proposition 6.2.1 . . . 351
6.3 Robust SIM Learning via Alignment Sharpness . . . 375
6.3.1 Initialization . . . 376
6.3.2 Optimization . . . 380
6.3.3 Testing . . . 397

**7 Robustly Learning Monotone GLMs under Gaussian Distribution 401**
7.1 Introduction . . . 401
7.2 Data Augmentation and Its Effect on the $L_2^2$ Loss Landscape . . . 408
7.2.1 Augmenting the Data: Connection to Ornstein–Uhlenbeck Semigroup . . . 408
7.2.2 Alignment of the Gradients of the Augmented Loss . . . 410
7.2.3 Critical Points and Their Connection to the $L_2^2$ Loss . . . 413
7.3 Learning GLMs via Variable Augmentation . . . 422
7.3.1 Finding the Best Parameter . . . 432
7.3.2 Proof of Lemma 7.3.4 . . . 437
7.4 SGD – VA Efficiently Learns Monotone GLMs . . . 443
7.4.1 Bounding Higher Order Hermite Coefficients of Monotone Activations . . . 447
7.4.2 Bounding the Augmentation Error . . . 450
7.4.3 Initialization Algorithm for Monotone Activations . . . 473
7.5 Discussion on Regular Activations . . . 479
7.5.1 Rescaling to the Unit Sphere . . . 480
7.5.2 Truncating the Regular Activations . . . 483
7.5.3 Examples of Regular Activations . . . 485
7.5.4 Required Assumptions on Activation . . . 491

**8 Robustly Learning Monotone SIMs under Gaussian Distribution 494**
8.1 Introduction . . . 494
8.2 Agnostic Learning SIMs . . . 496
8.2.1 Initialization . . . 501
8.2.2 The Spectral Subroutine . . . 504

8.2.3 Determining the Sample Complexity . . . 515
8.2.4 Proof of Proposition 8.2.12 . . . 522
8.3 Testing Subroutine . . . 528
8.4 Conclusion . . . 533

**III Robustly Learning Multi-Index Models 535**

**9 Learning Multi-Index Models with Queries 536**
9.1 Introduction . . . 536
9.1.1 Our Results . . . 537
9.1.2 Detailed Technical Overview . . . 539
9.1.3 Learning Bounded Variation Functions via Influence PCA . . . 543
9.1.4 Hardness of Proper Agnostic Query Learning for ReLUs and Halfspaces . . . 548
9.1.5 Related Work . . . 549
9.1.6 Roadmap . . . 552
9.2 From Zero- to First-Order: Derivative Queries via Oracle Queries . . . 552
9.2.1 Gradient Queries via Oracle Queries . . . 553
9.2.2 Smoothing the Labels for Learning Real-valued Functions . . . 555
9.2.3 Smoothing Labels for Learning Boolean Concepts . . . 559
9.3 Agnostically Learning Real-valued Multi-index Models . . . 562
9.3.1 Influence PCA for Learning in $L_2^2$ . . . 568
9.3.2 Proof of Theorem 9.3.2 . . . 573
9.3.3 Applications of Theorem 9.3.2 . . . 575
9.3.4 Proper Agnostic Query Learner for ReLUs . . . 581
9.3.5 Reducing the Dimension via Influence PCA . . . 582
9.4 Agnostically Learning Boolean Multi-index Models . . . 586
9.4.1 Influence PCA for Learning in $L_1$-norm . . . 587
9.4.2 Proof of Theorem 9.4.2 . . . 591
9.4.3 Corollaries for Intersections of Halfspaces and PTFs . . . 593
9.4.4 Proper Agnostic Query Learner for LTFs . . . 595
9.4.5 Reducing the Dimension via Influence PCA . . . 595
9.5 Hardness of Agnostic Proper Learning of Halfspaces and ReLUs with Queries . . . 601

**10 Agnostic PAC Learning Multi-Index Models 607**
10.1 Introduction . . . 607
10.1.1 Prior Work . . . 611
10.1.2 Technical Preliminaries . . . 613

10.1.3 Organization . . . 617
10.2 Learning Multiclass Linear Classifiers with Adversarial Label Noise . . . 617
10.2.1 Structural Result for Multiclass Linear Classification . . . 622
10.2.2 Finding a Relevant Direction . . . 631
10.2.3 Proof of Theorem 10.2.1 . . . 643
10.2.4 Proof of Lemma 10.2.7 . . . 648
10.3 Learning Multi-Index Models with Agnostic Noise . . . 652
10.3.1 Finding a Relevant Direction . . . 662
10.3.2 Proof of Theorem 10.3.3 . . . 670
10.4 Algorithmic Applications of General MIM Learner . . . 672
10.4.1 Learning Intersections of Halfspaces with Adversarial Label Noise 673
10.5 Omitted Content from Sections 10.2 to 10.4 . . . 682
10.5.1 Related Complexity Measures . . . 685
10.5.2 Near-equivalence to Generative Exponent for Single-Index Case 687
10.6 Lower Bound For Single-Shot Optimization . . . 691
10.6.1 Moment-Matching Construction . . . 694
10.6.2 Proof of Lemma 10.6.4 . . . 695
10.6.3 Proof of Claim 10.6.6 . . . 698
10.6.4 Reducing the Number of Intervals . . . 699

**IV Statistical-Computational Tradeoffs via SQ Complexity 704**

**11 SQ-Lower Bounds using Linear Programming 705**
11.1 Introduction . . . 705
11.1.1 Comparison to Prior Work . . . 710
11.2 SQ Lower Bound for Boolean-Valued Concepts: Proof of Theorem 11.1.1 712
11.3 Explicit SQ Lower Bounds for Boolean Concept Classes . . . 720
11.3.1 LTFs and Degree-$k$ PTFs . . . 720
11.3.2 Proof of Lemma 11.3.3 . . . 723
11.3.3 Intersections of Halfspaces: Degree Lower Bound via Gaussian Noise Sensitivity . . . 725
11.3.4 Proof of Theorem 11.1.2 . . . 727
11.4 Lower Bound for Real-Valued Functions . . . 731
11.4.1 CSQ Lower Bound: Proof of Theorem 11.1.3 . . . 731
11.4.2 SQ Lower Bound: Proof of Theorem 11.1.4 . . . 733
11.5 Applications for Classes of Real-Valued Functions . . . 736
11.5.1 ReLU Activation . . . 736
11.5.2 Sigmoid Activation . . . 738
11.6 Duality in Infinite-Dimensional LP . . . 740

11.7 Statistical Query Lower Bounds for Learning Massart Halfspaces . . . . 743
11.7.1 SQ Lower Bound for Learning Halfspaces with Constant-Bounded Massart Noise . . . . . 743
11.7.2 SQ Lower Bound for Learning Homogeneous Halfspaces with General Massart Noise . . . . . 756

**12 SQ Lower Bounds for Learning Multi-Index Models** **763**
12.1 Introduction . . . . . 763
12.1.1 SQ Lower Bound for General Multi-Index Models . . . . . 763
12.1.2 SQ Lower Bounds for Learning Multiclass Linear Classifiers with RCN . . . . . 767

**A Appendix to Chapter 2** **797**
A.1 "Massart" Noise with $\eta > 1/2$ is Equivalent to Agnostic Learning . . . . 797
A.2 Benign Noise is Equivalent to 1/2-Massart Noise . . . . . 798
A.3 A Non-Continuous Certificate . . . . . 799
A.4 Learning General Halfspaces with General Massart Noise . . . . . 800

**B Appendix to Chapter 3** **804**
B.1 Learning Margin Massart Halfspaces via Cutting Planes . . . . . 804

**C Appendix to Chapter 5** **807**
C.1 Learning Halfspaces using the Ramp Activation . . . . . 807

**D Appendix to Chapter 6** **809**
D.1 Remarks about the Distribution Class and the Activation Class . . . . 809
D.2 Efficiently Computing the Optimal Empirical Activation . . . . . 810
D.3 Uniform Convergence of Activations . . . . . 813

# List of Tables

1.1 Learning Real-Valued Functions using Queries: Running time comparisons of the best known PAC algorithms with our PAC+Queries technique (Influence PCA). . . . . 66
1.2 Learning Boolean Concepts using Queries: Running time comparisons of the best known agnostic learners (using random samples) with our Influence PCA technique (using queries). . . . . 73
1.3 Comparison of runtime and error guarantees between state-of-the-art efficient algorithms (see [KOS08, DKS18] for intersections of $K$ halfspaces) and our subspace approximation procedure. . . . . 79
1.4 Bounds on the degree $d$ of $\epsilon$-approximating polynomials in $L^1$-error under the Gaussian measure. For each concept class, we obtain an SQ lower bound of $n^{\Omega(d)}$. . . . . 98

9.1 Learning Real-Valued Functions using Queries: Running time comparisons of the best known PAC algorithms with our PAC+Queries technique (Influence PCA). For a summary of our results for the above classes, we refer to Section (1.3.8). . . . . 538
9.2 Learning Boolean Concepts using Queries: Running time comparisons of the best known agnostic learners (using random samples) with our Influence PCA technique (using queries). For a summary of our results for the above classes, we refer to Section (1.3.8). . . . . 540

10.1 Comparison of runtime and error guarantees between state-of-the-art efficient algorithms (see [KOS08, DKS18] for intersections of $K$ halfspaces) and our subspace approximation procedure. . . . . 610

11.1 Bounds on the degree $d$ of $\epsilon$-approximating polynomials in $L^1$-error under the Gaussian measure. For each concept class, we obtain an SQ lower bound of $n^{\Omega(d)}$. . . . . 707

11.2 Bounds on the degree $d$ of $\epsilon$-approximating polynomials in $L^1$ and $L^2$-error under the Gaussian measure. For each concept class, we obtain a CSQ (resp. SQ) lower bound of $n^{\Omega(d)}$, where $d$ is the $L^2$ degree (resp. $L^1$ degree). . . . . 709

**ROBUSTLY LEARNING MULTI-INDEX MODELS**

Nikos Zarifis

Under the supervision of Professor Ilias Diakonikolas
At the University of Wisconsin-Madison

This dissertation confronts a central challenge in modern machine learning: reliably learning from high-dimensional data that is inevitably corrupted by noise. While data often possesses a simpler, low-dimensional structure, identifying this structure remains a formidable task as datasets contain noise from measurement errors and mislabeled examples.

The core of this thesis is unified by a single question: Can we exploit the hidden low-dimensional subspace of a model to design algorithms that are provably robust to noise? This work answers this in the affirmative by developing novel techniques for learning in the presence of various label noise models, including the agnostic learning model, where no assumptions are made about the labels.

This dissertation first characterizes the complexity of learning halfspaces, the simplest single-index models, under semi-random Massart noise. We provide the first efficient algorithms for learning general halfspaces under Massart noise, even in the challenging regime where the noise rate is $1/2$.

The dissertation then addresses the problem of agnostically learning generalized linear and single-index models. We develop the first constant-factor approximate learners for a wide range of activation functions under well-behaved distributions. Under the Gaussian distribution, we design the first efficient constant-factor approximate learner for all monotone single-index models. This is achieved by developing a novel framework that bypasses the limitations of standard gradient-based methods and by using data augmentation techniques to provably improve the loss landscape.

The final part of the thesis extends the analysis to multi-index models. We introduce a framework that uses query access to overcome the standard computational barriers inherent in the standard agnostic PAC learning. For learning from random examples, we develop an iterative dimensionality reduction technique that yields the first constant-factor approximate learner for a broad class of multi-index models, including the special case of multiclass linear classification.

To establish the near-optimality of these algorithms, the dissertation concludes by developing a general framework based on linear programming duality to prove strong Statistical Query (SQ) lower bounds for agnostic learning under the Gaussian distribution.

Ilias Diakonikolas

# Chapter 1

# Introduction

## 1.1 The Challenge of Learning in the Presence of Noise

In an ideal world, machine learning algorithms would learn from perfectly clean data, where every example is a flawless representation of the underlying concept. The reality, however, is far from this ideal. Real-world data is invariably noisy, corrupted by measurement errors, mislabeled examples, or inherent ambiguity. This discrepancy between the idealized models of learning and the noisy reality of data has been a central challenge in machine learning since its inception. Valiant's Probably Approximately Correct (PAC) learning model [Val84], while a cornerstone of the field, does not explicitly account for this noise. To bridge this gap, the agnostic framework was introduced [Hau92, KSS94], a more realistic model that makes no assumptions about the nature of the noise and aims to develop algorithms that are robust by design.

Agnostic learning, however, is computationally hard in the general, distribution-free

setting, even for simple concept classes [GR06, FGKP06, Dan16]. This hardness has motivated a focused line of research into learning under specific, structured distributions, where efficient learning is sometimes possible. Of particular interest are high-dimensional distributions like the standard Gaussian, which often admits efficient noise-tolerant algorithms [KKMS08, ABL17, DKS18]. This thesis is situated at this frontier, seeking to precisely delineate the boundary of efficient agnostic learning for fundamental machine learning models under structured distributions.

## 1.2 A Unifying Theme: Learning Low-Dimensional Structure in High Dimensions

A central premise in high-dimensional data analysis is that despite the ambient dimensionality, the data often possesses a simpler, low-dimensional structure. This thesis focuses on learning problems where this structure is captured by single and multi-index models. In a single-index model, the target function depends on a single unknown direction, $f(\mathbf{x}) = \sigma(\mathbf{w} \cdot \mathbf{x})$, while in a multi-index model, it depends on a $k$-dimensional projection of the data, $f(\mathbf{x}) = g(\mathbf{W}\mathbf{x})$. These models are fundamental, forming the building blocks of modern neural networks.

Learning these models, particularly in the agnostic setting is a challenging regime. A unifying theme of this thesis is the exploration of how to leverage the underlying structure of the problem to design efficient algorithms.

> *Can one exploit the hidden $k$-dimensional subspace to robustly learn multi-index models?*

This demonstrates that by identifying and exploiting this hidden structure, it is

possible to design robust learners that can handle noise, often with nearly optimal guarantees. This thesis is based on the following works.

- **Optimal Algorithms for Learning Halfspaces:** In Chapter 2, we provide the first efficient algorithms for learning general (non-homogeneous) halfspaces under Massart noise, even for the case where $\eta = 1/2$. Our algorithmic results are complemented by nearly-matching computational lower bounds. It is based on [DKK+22]. In Chapter 3, we study the computational complexity of learning halfspaces under Massart noise for margin halfspaces, providing the first sample efficient algorithms for learning margin halfspaces, based on [DZ24]. In Chapter 4, we develop a framework for proper agnostic learning under Gaussian distribution, which returns a hypothesis from the same target class. It is based on [DKK+21a].

- **Agnostically Learning Single-Index Models:** In Chapter 5, we develop the first constant-factor approximate learners for a broad range of known sigmoidal and unbounded activation functions under well-behaved distributions, where in Chapter 7 we design the first constant-factor approximate learner for all monotone single-index models under the Gaussian distribution. It is based onthe works [DKTZ22a, ZWDD25]. In Chapters 6 and 8, we study the more challenging setting where the monotone activation function is unknown and we provide the first efficient constant-factor approximate learner for this problem. It is based onthe works: [ZWDD24, WZDD25].

- **Agnostically Learning Multi-Index Models:** In Chapter 9, we introduce a novel framework that uses query access to overcome the computational barriers of sample-based agnostic learning multi-index models. It is based on [DKK+24]. In

Chapter 10 we develop an iterative dimensionality reduction technique that learns from random examples multi-index models, providing the first constant-factor approximate learner for multi-index models and for the special case of multiclass linear classification. It is based on [DIKZ25].

- **Statistical Computational Tradeoffs:** Finally, in Chapter 11, we study the fundamental statistical-computational limits of the problems studied in this thesis. We develop a general framework based on linear programming duality to prove strong Statistical Query (SQ) lower bounds for agnostic learning under the Gaussian distribution. It is based on [DKPZ21].

## 1.3 Main Contributions and Thesis Outline

This thesis is composed of a series of papers that collectively advance our understanding of robust learning. We provide a suite of efficient algorithms for fundamental learning tasks, complemented by nearly-matching computational lower bounds that delineate the limits of what is possible.

### Part I: Characterizing the Complexity of Learning Halfspaces

The first part of this thesis delves into the foundational problem of learning halfspaces, the simplest single-index models, under random and adversarial noise.

A *halfspace* or Linear Threshold Function (LTF) is any Boolean-valued function $f : \mathbb{R}^d \mapsto \{\pm 1\}$ of the form $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} - t^*)$, for a vector $\mathbf{w}^* \in \mathbb{R}^d$ (known as the weight vector) and a scalar $t^* \in \mathbb{R}$ (known as the threshold). Halfspaces are a central class of Boolean functions in several areas of computer science, including complexity

theory, learning theory, and optimization [Ros58, Nov62, MP68, Yao90, GHR92, FS97, Vap98, STC00, O'D14]. In this thesis, we focus on the algorithmic problem of learning halfspaces from labeled examples, arguably one of the most extensively studied and influential problems in machine learning.

### 1.3.1 Learning Gaussian Halfspaces with General Massart Noise

The computational problem of PAC learning halfspaces is known to be efficiently solvable without noise (see, e.g., [MT94]) in the distribution-independent setting. The complexity of the problem in the presence of noisy data crucially depends on the noise model and the underlying distributional assumptions. In this thesis, we study the problem of distribution-specific PAC learning of halfspaces in the presence of Massart and adversarial noise. Formally, we have the following definitions

**Definition 1.3.1** (Distribution-specific PAC Learning with Massart Noise)**.** *Let $\mathcal{C}$ be a concept class of Boolean functions over $X = \mathbb{R}^d$, $\mathcal{F}$ be a* known family *of structured distributions on $X$, $0 \leq \eta \leq 1/2$, and $0 < \epsilon < 1$. Let $f$ be an unknown target function in $\mathcal{C}$. A* noisy example oracle*, $\mathrm{EX}^{\mathrm{Mas}}(f, \mathcal{F}, \eta)$, works as follows: Each time $\mathrm{EX}^{\mathrm{Mas}}(f, \mathcal{F}, \eta)$ is invoked, it returns a labeled example $(\mathbf{x}, y)$, such that: (a) $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$, where $\mathcal{D}_{\mathbf{x}}$ is a fixed distribution in $\mathcal{F}$, and (b) $y = f(\mathbf{x})$ with probability $1 - \eta(\mathbf{x})$ and $y = -f(\mathbf{x})$ with probability $\eta(\mathbf{x})$, for an* unknown *function $\eta(\mathbf{x}) \leq \eta$. Let $\mathcal{D}$ denote the joint distribution on $(\mathbf{x}, y)$ generated by the above oracle. A learning algorithm is given i.i.d. samples from $\mathcal{D}$ and its goal is to output a hypothesis $h$ such that with high probability it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, where $\mathrm{OPT} = \min_{c\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[c(\mathbf{x}) \neq y]$.*

The Massart noise model in the above form was defined in [MN06]. A very similar

noise model had been defined in the 80s by Sloan and Rivest [Slo88, Slo92, RS94, Slo96], and a related definition had been considered even earlier by Vapnik [Vap82]. The Massart model is a generalization of the Random Classification Noise (RCN) model [AL88] (where the flipping probability is uniform) and is a special case of the agnostic model (where the label noise is fully adversarial) [Hau92, KSS94].

The Massart model is a natural semi-random noise model that is more realistic than RCN. Specifically, label noise can reflect computational difficulty, ambiguity, or random factors. For example, a cursive "e" might be substantially more likely to be misclassified as "a" than an upper case Roman letter. Massart noise allows for such variations in misclassification rates without knowledge of which instances are more likely to be misclassified. That is, Massart noise-tolerant learners are less brittle than RCN tolerant learners. Agnostic learning is, of course, even more robust; unfortunately, agnostic learning is known to be computationally intractable in many settings of interest [GR06, FGKP06, Dan16, DKZ20, GGK20, DKPZ21].

## Motivation for this Work

The algorithmic task of PAC learning with Massart noise is a classical problem in computational learning theory. In the distribution-independent setting, known efficient algorithms [DGT19, CKMY20] achieve error $\eta+\epsilon$ and we now know [DK20a] that this error bound is close to best possible for efficient Statistical Query (SQ) algorithms [Kea98], even if $\mathrm{OPT} = \mathbf{E}[\eta(\mathbf{x})]$ is very small. This lower bound further motivates the study of the distribution-specific setting (the focus of the current work).

The work of [ABHU15] initiated the algorithmic study of learning Massart halfspaces under structured distributions. This work focused on the class of *homogeneous*

halfspaces, i.e., functions of the form $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$, and gave a polynomial-time learning algorithm with $\mathrm{OPT} + \epsilon$ error under the uniform distribution on the unit sphere. The [ABHU15] algorithm succeeds when the parameter $\eta$ is smaller than a sufficiently small constant ($\approx 10^{-6}$). A sequence of subsequent works [ABHZ16, YZ17, ZLC17, BZ17, MV19, DKTZ20a, ZSA20, ZL21] have given efficient learning algorithms for homogeneous Massart halfspaces that succeed for all $\eta < 1/2$ and under weaker distributional assumptions. The current state-of-the-art algorithms [DKTZ20a, ZSA20, ZL21] have sample and computational complexity $\mathrm{poly}(d, 1/\epsilon, 1/(1-2\eta))$ and succeed for all $\eta < 1/2$ under a class of distributions including isotropic log-concave distributions.

To summarize the preceding discussion, *known efficient algorithms for Massart halfspaces in the distribution-specific setting succeed under two crucial assumptions:*

*(i) The target halfspace is homogeneous (i.e., has zero threshold), and*

*(ii) The upper bound parameter of the Massart noise satisfies $\eta < 1/2$.*

Perhaps surprisingly, prior to the current work, no non-trivial bounds were known for the *general case* of this learning problem, where either of these two assumptions is removed. This represents a fundamental gap in our algorithmic understanding of learning halfspaces in the Massart model and serves as the main motivation of the current work.

In this work, we study the general version of this problem for the prototypical setting that the examples are drawn from the Gaussian distribution. As our main contribution, *we essentially characterize the complexity of the problem by giving the first efficient learning algorithms coupled with qualitatively matching SQ lower bounds.*

In the following paragraphs, we provide a more detailed technical motivation of the regimes we study followed by a detailed description of our results.

**Massart Learning of *General* Halfspaces** Suppose that the Massart noise rate $\eta$ is a constant *strictly smaller* than $1/2$. Even for this "low-noise" regime, all previous efficient learning algorithms that achieve error $\mathrm{OPT} + \epsilon$ require that the unknown halfspace is *homogeneous*. Superficially, this might seem like an innocuous assumption. After all, it seems straightforward to reduce a general halfspace to a homogeneous one by adding an extra constant coordinate to every sample. Given this, one could use a learner for the homogeneous case on the modified instance. It turns out that this intuition is fundamentally flawed. While such a reduction is valid in the distribution-independent setting, it does not work in the distribution-specific setting because it alters the marginal distribution on the examples. In light of this state of affairs, it is natural to ask whether an efficient Massart learner exists for *general* halfspaces in the low-noise noise regime where previous algorithms succeeded.

**Question 1.3.2.** What is the complexity of learning *general* halfspaces in the *constant-bounded* Massart noise setting, i.e., when $\eta = 1/2 - c$ for some universal constant $c > 0$?

As we will show, the complexity of the problem in this regime is characterized by the "bias" $\gamma$ of the target halfspace (see Definition 1.3.4) and (inherently) scales *quasi-polynomially* with $1/\gamma$ (Theorems 2.1.2 and 2.1.3). We note that previous algorithms only handle the special case of $\gamma = 1/2$.

**Learning Halfspaces with *General* Massart Noise** The original definition of the Massart noise model [MN06] allows the upper bound on the noise rate to be *equal* to $1/2$. On the other hand, all known Massart learning algorithms crucially rely on the assumption that $\eta < 1/2$. The latter assumption might have been motivated by

the random classification noise model, where the $\eta = 1/2$ regime is not meaningful. In the Massart model, however, it may well be the case that $\eta(\mathbf{x}) = 1/2$, for a small probability subset of the domain and $\eta(\mathbf{x})$ is small otherwise, in which case the optimal misclassification error OPT will be similarly small.

Understanding the complexity of learning halfspaces with *general* Massart noise, i.e., in the regime where $\eta = 1/2$, is both of theoretical and practical significance. From the theoretical standpoint, the "high-noise" regime subsumes the well-studied Tsybakov model [MT99, Tsy04] and can be viewed as the strongest known non-adversarial noise model in the literature. Interestingly, the general Massart noise model has also been previously studied in the statistical learning theory literature under the name *benign noise*, see, e.g., [Han09, HY15]. In the "benign noise" model, the only assumption made about the label noise is that the *Bayes optimal classifier* lies in the target class. (It is an easy exercise, see Section A.2, that the benign noise model is equivalent to the general Massart model.) In addition to its theoretical interest, the general Massart model naturally arises in a number of practical applications. A concrete example is that of *human annotator noise* [CM84, KB09, BK09, KB10], where it has been shown [KB10] that human annotators (especially non-experts) often flip coins (corresponding to $\eta(\mathbf{x}) = 1/2$) when presented with a hard-to-classify example. The preceding discussion motivates the following question:

**Question 1.3.3.** What is the complexity of learning halfspaces with *general* Massart noise?

Interestingly, prior to this work, this question remained wide-open even for the special case of homogeneous halfspaces. Specifically, previous Massart learning algorithms (for homogeneous halfspaces) provide *vacuous guarantees* for $\eta = 1/2$, because they require

sample complexity scaling polynomially with the parameter $1/\beta$, where $\beta := 1 - 2\eta$. It is worth noting that a dependence on $\beta$ is not information-theoretically required for the problem. Specifically, $O(d/\epsilon^2)$ samples suffice to achieve error $\mathrm{OPT} + \epsilon$ (see, e.g., [MN06]), alas with an exponential time algorithm.

At a high-level, the $\beta$-dependence in previous approaches is due to the fact that previous algorithms solve the (harder) *parameter recovery* problem, i.e., they approximate the hidden weight vector $\mathbf{w}^*$ (e.g., within small angle). While the sample complexity of PAC learning is independent of $\beta$, parameter learning requires at least $1/\beta$ samples. Consequently, a genuinely new approach is required to handle the $\beta = 0$ regime. That is, to answer Question 1.3.3, we need to understand to what extent it is possible to PAC learn without relying on parameter recovery.

As we will show, handling the $\beta = 0$ case comes at a cost. Our results (Theorems 2.1.5 and 2.1.6) establish an information-computation tradeoff for this problem (that persists even for homogeneous halfspaces) scaling *quasi-polynomially* with the parameter $1/\epsilon$.

## Main Results

In this work, we answer Questions 1.3.2 and 1.3.3 by providing both efficient learning algorithms and nearly-matching lower bounds in the Statistical Query (SQ) model. Perhaps surprisingly, we show that the complexity of our learning problem scales quasi-polynomially with $1/\epsilon$, where $\epsilon$ is the excess error. A conceptual implication of our findings is that Questions 1.3.2 and 1.3.3, while seemingly orthogonal, are in fact intimately connected at a technical level.

**Definition 1.3.4** ($(1-\gamma)$-biased Halfspace)**.** *For $\gamma \in [0, 1/2]$, we say that the halfspace $f$ is at most $(1-\gamma)$-biased (with respect to $\mathcal{D}$) if* $\min(\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) = +1], \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) =$

$-1]) \geq \gamma$.

The bias of a halfspace, as captured by the parameter $\gamma$, proves to be a crucial factor in determining the computational complexity of this learning problem. While homogeneous halfspaces correspond to the most balanced case ($\gamma = 1/2$) and are known to be efficiently learnable in polynomial time, the problem becomes significantly harder as the halfspace becomes more biased (i.e., as $\gamma$ approaches 0). Our work provides a precise characterization of this relationship, establishing nearly-matching upper and lower bounds that reveal a quasi-polynomial dependence on $1/\gamma$.

**Theorem 1.3.5** (Learning General Massart Halfspaces, Informal, from Chapter 2)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal that satisfies the Massart condition with respect to to an at most $(1-\gamma)$-biased halfspace.*

- *If the noise is bounded by $\eta \leq 1/2 - c$ for a constant $c > 0$, there is an algorithm that learns a hypothesis with error $\mathrm{OPT} + \epsilon$ in time $d^{O(\log(1/\gamma))}\mathrm{poly}(1/\epsilon)$, where $\gamma$ is the bias of the target halfspace. This is matched by an SQ lower bound of $d^{\Omega(\log(1/\gamma))}$.*

- *If the noise rate is $\eta = 1/2$ (benign noise), there is an algorithm with runtime $d^{O(\log(1/\epsilon))}2^{\mathrm{poly}(1/\epsilon)}$. This is matched by an SQ lower bound of $d^{\Omega(\log(1/\epsilon))}$, even for homogeneous halfspaces.*

## Brief Overview of Techniques

**Leveraging Low-Degree Moments** The unifying theme of both our upper and lower bound techniques is the use of low-degree moments. For our purposes, the

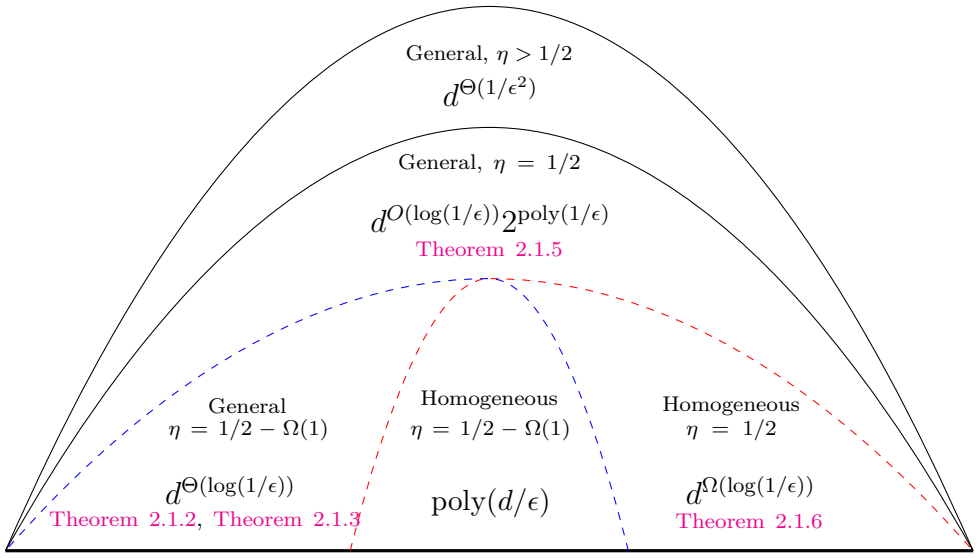



Figure 1.1: Overview of the (SQ) complexity of learning halfspaces with Massart noise. (1) For homogeneous halfspaces and $\eta < 1/2$, efficient algorithms were previously known under Gaussian or isotropic log-concave marginals. (2) For $\eta > 1/2$, the problem is equivalent to agnostic learning, for which tight upper and lower bounds were previously known. (3) The remaining regimes (general halfspaces and/or $\eta = 1/2$) are characterized in the current work.

low-degree moments of a distribution correspond to the values of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x})y]$ for low-degree polynomials $p$. It is clear that one can compute approximations of the (up to) $k$-degree moments of such a distribution with $\mathrm{poly}(d^k/\epsilon)$ samples and time. At a high-level, we are looking for a moment that will provide us with information about the correlation between $\mathbf{x}$ and $y$. In more detail, we would like to find polynomials $p$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x})y] \neq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{x})]\,\mathbf{E}_{y\sim\mathcal{D}_y}[y]$, or equivalently find polynomials $p$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{x})] = 0$ and $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x})y] \neq 0$. If no such polynomial exists, we can leverage this fact to prove SQ lower bounds; if such polynomials exist, we can hope to use them as a starting point for an algorithm.

**Learning with Constant-Bounded Massart Noise: Theorem 2.1.2** At a high-level, our algorithm fits in the certificate-based framework developed in [DKTZ20b, DKK+20, DKK+21b]. (A similar framework was developed independently in [CKMY20] to properly learn Massart halfspaces in the distribution-free setting, matching the error of the [DGT19] algorithm.) The framework relies on the following fact: if the true halfspace is given by $f(\mathbf{x}) = \mathrm{sign}(\ell^*(\mathbf{x}))$, where $\ell^*(\mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x} - t^*$, then for any affine function $\ell(\mathbf{x})$ we have that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})yT(\mathbf{x})] \geq 0$ for all non-negative functions $T(\mathbf{x})$ if and only if $\ell = \ell^*$. We can think of this as an (infinite) linear program that can be used to solve for $\mathbf{w}^*$ and $t^*$. In order to solve this program, we need a separation oracle. In particular, for any hypothesis $\mathbf{w}$ and $t$ that is too far from the truth, i.e., $\mathrm{sign}(\ell(\mathbf{x}))$ has error greater than $\mathrm{OPT} + \epsilon$, we need to be able to find an explicit non-negative function $T(\mathbf{x})$ such that the above constraint is violated. This essentially amounts to finding a function $T(\mathbf{x})$ that concentrates on the values of $\mathbf{x}$ for which our current hypothesis $\ell(\mathbf{x})$ is incorrect.

Our main new idea here is that for any sub-optimal halfspace guess $h$, there exists a certificate function of the form $T(\mathbf{x}) = \mathbb{1}_S(\mathbf{x})e^{\mathbf{v}\cdot\mathbf{x}}$ (*exponential-shift certificate*), where $S$ is a thin strip around the current (known) halfspace $h$ and $\mathbf{v}$ is an appropriate weight vector. However, finding a good weight vector $\mathbf{v}$ is a non-convex (hard) optimization problem in general. We simulate the behavior of the exponential-shift certificate by taking $T(\mathbf{x})$ to be the square of a low-degree polynomial $q(\mathbf{x})$ restricted to an appropriately chosen thin strip $S$. To construct our polynomial certificate, we essentially need to prove that for any at most $(1-\gamma)$-biased one-dimensional threshold function $f$, there exists a polynomial $q$ of degree $O(\log(1/\gamma))$ such that $q^2(\mathbf{x})$ concentrates on the positive values of $f$; see Lemma 2.2.3. To prove the existence of such a polynomial, we establish

that considering the $O(\log(1/\gamma))$-degree Taylor approximation of the exponential-shift $e^{\mathbf{v}\cdot\mathbf{x}}$ suffices; see Lemma 2.3.5. We can efficiently compute a polynomial certificate since, once we have a fixed strip $S$, we can compute the low-degree moments of $y$ restricted to $S$. From there, finding an appropriate polynomial $q$ amounts to finding a negative eigenvalue of a quadratic form. For more details, we refer the reader to Subsection 2.2.1 and Section 2.3.

**Learning with *General* Massart Noise: Theorem 2.1.5** In the case of learning with *general* Massart noise ($\eta = 1/2$), the certificate approach fails, as it requires polynomials of large degree — polynomial in $1/\epsilon$ — to get concentration in the disagreement region. This is prohibitively large as it would result in a runtime of $d^{\mathrm{poly}(1/\epsilon)}$, which we can readily get by simply running the agnostic learner of [KKMS08]. The main idea to obtain an improved bound is to bypass the limitations of the certificate approach by relying on *correlation* properties rather than *concentration*. As we show, there is always some polynomial $p$ of low degree, logarithmic in $1/\epsilon$, that correlates with the label $y$, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x})y] > 0$. Our algorithm exploits this weaker property and turns it into an iterative process that constantly improves the current guess.

Since low-degree polynomials are able to get non-trivial correlation, if we compute the low-order moment tensors of $y$, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{x}^{\otimes k}y]$, we can use them to compute a low-dimensional subspace $V$ onto which $\mathbf{w}^*$ has non-trivial ($\mathrm{poly}(\epsilon)$) projection; see Proposition 2.4.9 and Lemma 2.4.10. This means that picking a random element $\mathbf{v} \in V$ will, with reasonable probability (say $1/3$), have non-trivial, i.e., $\mathrm{poly}(\epsilon)$, correlation with $\mathbf{w}^*$. Our algorithm improves an initial guess $\mathbf{w}$ of the optimal weight vector as follows: by applying the above technique to a thin strip perpendicular to our current guess $\mathbf{w}$, we can – with some non-trivial probability – find a vector that correlates

non-trivially with the projection of $\mathbf{w}^*$ on the orthogonal complement of $\mathbf{w}$. This, in turn, will allow us to compute a new guess $\mathbf{w}'$ with a slightly better correlation with $\mathbf{w}^*$; see Lemma 2.4.14. Repeating this process $\mathrm{poly}(1/\epsilon)$ times will produce a vector with at most $\mathrm{OPT} + \epsilon$ error; see Section (2.2.2). Each iteration of the above algorithm requires $d^{O(\log(1/\epsilon))}$ samples and time to compute the moments of order $O(\log(1/\epsilon))$. We then need to run the algorithm many times to find a trial in which we get lucky for $\mathrm{poly}(1/\epsilon)$ rounds in a row. This latter operation increases the complexity by a $2^{\mathrm{poly}(1/\epsilon)}$ factor.

The crucial structural result that we exploit is that for any $\epsilon$-biased halfspace $f(\mathbf{x})$ we can construct a *mean-zero* polynomial $p$, which is a function of $\mathbf{w}^* \cdot \mathbf{x}$, and matches the sign of $f(\mathbf{x})$ everywhere, see (Informal) Proposition 2.2.8 and Subsection 2.4.1. We then show that, even when Massart noise with $\eta = 1/2$ is applied to $f$, this polynomial $p$ will achieve non-trivial correlation with $f$, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x})y] > 0$. This implies that some low-order moment of our distribution must have a non-trivial component in the $\mathbf{w}^*$-direction; see Subsection 2.4.3. Flattening the moment tensors and performing SVD on the resulting matrices, we can efficiently construct a subspace (spanned by the top eigenvectors) onto which $\mathbf{w}^*$ has non-trivial projection.

### 1.3.2 Learning Margin Halfspaces with Massart Noise

Moving beyond the Gaussian distribution, another common assumption is the *margin* condition: the data are separated by some unknown vector $\mathbf{w}^*$ with slack at least $\gamma > 0$. Such geometric structure allows learning for more general distribution, i.e., bounded distributions. The next theorem, proved in Chapter 3, gives the first algorithm whose sample complexity is simultaneously optimal in its dependence on $\epsilon$ and on the margin parameter $\gamma$. The concept class of halfspaces with a margin is defined as follows.

**Definition 1.3.6** ($\gamma$-Margin Halfspaces)**.** *Let $\mathcal{D}$ be a distribution over $\mathbb{S}^{d-1} \times \{\pm 1\}$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. Let $\mathbf{w}^* \in \mathbb{S}^{d-1}$ and $\gamma \in (0,1)$. We say that the distribution $\mathcal{D}$ satisfies the $\gamma$-margin condition with respect the halfspace* $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$[1]*, if (i) for $(\mathbf{x}, y) \sim D$, we have that $y = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$, and (ii) $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{w}^* \cdot \mathbf{x}| < \gamma] = 0$. The parameter $\gamma$ is called the margin of the halfspace* $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$.

Information-theoretically, the best possible 0-1 error attainable for learning a concept class with Massart noise is $\mathrm{OPT} := \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})]$. Since $\eta(\mathbf{x})$ is uniformly bounded above by $\eta$, it follows that $\mathrm{OPT} \leq \eta$; also note that it may well be the case that $\mathrm{OPT} \ll \eta$. Focusing on the class of $\gamma$-margin halfspaces, it follows from [MN06] that there exists a (computationally inefficient) estimator achieving error $\mathrm{OPT} + \epsilon$ with sample complexity $\widetilde{O}(1/((1-2\eta)\gamma^2\epsilon))$; and moreover that this sample upper bound is nearly best possible (within a logarithmic factor) for any estimator. (That is, the sample complexity of the Massart learning problem is essentially the same as in the realizable case, as long as $\eta$ is bounded from 1/2.)

Taking computational considerations into account, the feasibility landscape of the problem changes. Prior work [DK20a, NT22, DKMR22a] has provided strong evidence that achieving error better than $\eta + \epsilon$ is not possible in polynomial time. Consequently, algorithmic research has been focusing on achieving the qualitatively weaker error guarantee of $\eta + \epsilon$. We note that efficiently obtaining any non-trivial guarantee had remained open since the 80s. The first algorithmic progress for this problem is due to [DGT19], who gave a polynomial-time algorithm achieving error of $\eta + \epsilon$ with sample complexity $\mathrm{poly}(1/\gamma, 1/\epsilon)$. Subsequent work [CKMY20] gave an efficient algorithm with improved sample complexity of $\tilde{O}(1/(\gamma^4\epsilon^3))$. Prior to the current work, this remained

[1]We will henceforth assume that the threshold is $\theta = 0$, which is well-known to be no loss of generality.

the best known sample upper bound for efficient algorithms.

In summary, known computationally efficient algorithms for learning margin halfspaces with Massart noise require significantly more samples—namely, $\tilde{\Omega}(1/(\gamma^4\epsilon^3))$—than the information-theoretic minimum of $\widetilde{\Theta}_\eta(1/(\gamma^2\epsilon))$. It is thus natural to ask whether a polynomial-time algorithm with optimal (or near-optimal, i.e., within logarithmic factors) sample complexity exists. Recall that the answer to this question is affirmative in the realizable setting, where the Perceptron algorithm is optimal. Perhaps surprisingly, recent work [DDK+23a] (see also [DDK+23b]) gave evidence for the existence of inherent *information-computation tradeoffs* in the Massart noise model—in fact, even in the simpler model of Random Classification Noise (RCN) [AL88][2]. Specifically, they showed that any efficient Statistical Query (SQ) algorithm or low-degree polynomial tasks requires $\Omega(1/\epsilon^2)$ samples—a near quadratic blow-up compared to the $\tilde{O}(1/\epsilon)$ information-theoretic upper bound. This discussion serves as the motivation for the following question:

**Question 1.3.7.** What is the optimal *computational sample complexity* of the problem of learning $\gamma$-margin halfspaces with Massart noise?

By the term "computational sample complexity" above, we mean the sample complexity of polynomial-time algorithms for the problem. Given the fundamental nature of this learning problem, we believe that a fine-grained sample complexity versus computational complexity analysis is interesting on its own merits. *In this work, we develop a computationally efficient algorithm with sample complexity of* $\tilde{O}(1/(\gamma^2\epsilon^2))$. Given the aforementioned information-computation tradeoffs, there is evidence that this upper bound is close to best possible. As a bonus, our algorithm is also simple and practical,

[2]The RCN model is the special case of Massart noise, where $\eta(\mathbf{x}) = \eta$ for all points $\mathbf{x}$ in the domain.

relying on online SGD. (In fact, our algorithm runs in sample linear time, excluding a final testing step that slightly increases the runtime.)

## Main Result and Techniques

The main result of this part is the following:

**Theorem 1.3.8** (Main Result, Informal)**.** *Let $D$ be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ that satisfies the $\eta$-Massart noise condition with respect to an unknown $\gamma$-margin halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. There is algorithm that draws $n = \tilde{O}(1/(\epsilon^2\gamma^2))$ samples from $D$, runs in time $\tilde{O}(dn/\epsilon)$, and with probability at least $9/10$ returns a vector $\hat{\mathbf{w}}$ such that $\mathrm{err}_{\mathcal{D}}(\hat{\mathbf{w}}) \leq \eta + \epsilon$.*

The sample upper bound of Theorem 1.3.8 nearly matches the computational sample complexity of the problem (for SQ algorithms and low-degree polynomial tests), which was shown to be $\Omega(1/(\epsilon^2\gamma) + 1/(\epsilon\gamma^2))$ [MN06, DDK+23a, DDK+23b]. That is, Theorem 1.3.8 comes close to resolving the fine-grained complexity of this basic task. Moreover, it matches known algorithmic guarantees for the easier case of Random Classification Noise [DDK+23a, KIT+23].

**Brief Overview of Techniques** Here we provide a brief summary of our approach in tandem with a comparison to prior work. The algorithm of [DGT19] adaptively partitions the space into polyhedral regions and uses a different linear classifier in each region, each achieving error $\eta + \epsilon$ within the corresponding region. Their approach leverages the LeakyReLU loss (see (3.1)) as a convex proxy to the 0-1 loss. At a high-level, their approach reweights the samples in order to accurately classify a non-trivial fraction of points. [CKMY20] uses the LeakyReLU loss to efficiently identify

a region where the value of the loss conditioned on this region is sub-optimal; they then use this procedure as a separation oracle along with online convex optimization (see also [DKTZ20b, DKK+21b]) to output a linear classifier with 0-1 error at most $\eta + \epsilon$. Both of these approaches inherently require $\Omega(1/\epsilon^3)$ samples for the following reason: they both need to condition on a region where the probability mass of the distribution can be as small as $\Theta(\epsilon)$. Thus, even estimating the error of the loss would require at least $\Omega(1/\epsilon^2)$ conditional samples. Beyond the dependence on $1/\epsilon$, the sample complexity achieved in these prior works is also suboptimal in the margin parameter $\gamma$; namely, $\Omega(1/\gamma^4)$. This dependence follows from the facts that both of these works require estimating the loss in each iteration within error of at most $\gamma\epsilon$, and that their algorithmic approaches require $\Omega(1/\gamma^2)$ iterations.

To circumvent these issues, novel ideas are required. At a high-level, we design a uniform approach to decrease the "global" error, as opposed to the local error (as was done in prior work). Specifically, we construct a different sequence of convex loss functions, each of which attempts to accurately simulate the 0-1 objective. We note that a similar sequence of loss functions was used in the recent work [DKTZ24] in a related, but significantly different, adversarial online setting. Interestingly, a similar reweighting scheme was used in [CKMY20] for learning general Massart halfspaces. Beyond this similarity, these works have no implications for the sample complexity of our problem. Via this approach, we obtain an iterative algorithm which uses only $O_\gamma(1/\epsilon^2)$ samples in order to estimate the loss in each iterative step.

In more detail, note that the 0-1 loss can be written in the form $-\mathbf{E}[y\frac{\mathbf{w}\cdot\mathbf{x}}{|\mathbf{w}\cdot\mathbf{x}|}]$. We convexify this objective by considering, in each step, the loss $\ell(\mathbf{w},\mathbf{u}) = -\mathbf{E}[y\frac{\mathbf{w}\cdot\mathbf{x}}{|\mathbf{u}\cdot\mathbf{x}|}]$, where $\mathbf{u}$ is independent of $\mathbf{w}$; this loss is convex with respect to $\mathbf{w}$. Observe that $\ell(\mathbf{w},\mathbf{w})$

is proportional to the zero-one loss of $\mathbf{w}$. Unfortunately, it is possible that no optimal vector $\mathbf{w}^*$ (under 0-1 loss) minimizes $\ell(\mathbf{w}^*, \mathbf{w})$. For this reason, we consider the objective $\ell_\eta(\mathbf{w}, \mathbf{u}) = \mathbf{E}[(\mathbb{1}\{y \neq \mathrm{sign}(\mathbf{w} \cdot \mathbf{x})\} - \eta - \epsilon)|\mathbf{w} \cdot \mathbf{x}|/|\mathbf{u} \cdot \mathbf{x}|]$. This new objective satisfies the following: $\ell_\eta(\mathbf{w}^*, \mathbf{u}) < -\epsilon\gamma$ for any vector $\mathbf{u}$ and any $\mathbf{w}^*$ that minimizes the 0-1 objective; and $\ell_\eta(\mathbf{w}, \mathbf{w}) \geq \epsilon$ as long as $\mathbf{w}$ incurs 0-1 error at least $\eta + \epsilon$. By the convexity of $\ell_\eta(\mathbf{w}, \mathbf{u})$, this allows us to construct a separation oracle. Namely, we draw enough samples so that $\widehat{\ell}_\eta(\mathbf{w}, \mathbf{w}) - \widehat{\ell}_\eta(\mathbf{w}^*, \mathbf{w}) \geq \epsilon/2$, where $\widehat{\ell}$ is the empirical version of the loss. Due to the nature of these objectives, $O_\gamma(1/\epsilon^2)$ samples per iteration suffice for this purpose. This in turn implies that the cutting planes method efficiently finds a near-optimal weight vector after $O(\log(1/\epsilon)/\gamma^2)$ iterations. Overall, this approach leads to an efficient algorithm with sample complexity $\tilde{O}_\gamma(1/\epsilon^2)$. To get the desired sample complexity of $\tilde{O}(1/(\epsilon^2\gamma^2))$, more ideas are needed.

In the previous paragraph, we hid an obstacle that makes the above approach fail. Specifically, it may be possible that, for many points $\mathbf{x}$, the value of $|\mathbf{u} \cdot \mathbf{x}|$ is arbitrarily small. To fix this issue, we consider a clipped reweighting as follows: $\ell'_\eta(\mathbf{w}, \mathbf{u}) = \mathbf{E}[(\mathbb{1}\{y \neq \mathrm{sign}(\mathbf{w} \cdot \mathbf{x})\} - \eta - \epsilon)\frac{|\mathbf{w} \cdot \mathbf{x}|}{\max(|\mathbf{u} \cdot \mathbf{x}|, \gamma)}]$. This clipping step is not a problem for us, because the target halfspace $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$ was assumed to have margin $\gamma$. This guarantees that the difference between the expected (over $y$) pointwise losses at $(\mathbf{w}, \mathbf{w})$ and $(\mathbf{w}^*, \mathbf{w})$ is at least $\epsilon$ on the points $\mathbf{x}$ where $|\mathbf{u} \cdot \mathbf{x}| \leq \gamma$. Indeed, when this is the case, then $|\mathbf{w}^* \cdot \mathbf{x}|/|\mathbf{u} \cdot \mathbf{x}| \geq 1$. Overall, this suffices to guarantee that $\ell'_\eta(\mathbf{w}, \mathbf{w}) - \ell'_\eta(\mathbf{w}^*, \mathbf{w}) \geq \epsilon$.

### 1.3.3 Agnostic Proper Learning Halfspaces under Gaussian Distribution

Here we study the task of learning halfspaces in the *agnostic framework* [Hau92, KSS94], which models the phenomenon of learning from adversarially labeled data. While halfspaces are efficiently learnable in the presence of consistently labeled examples (see, e.g., [MT94]) — i.e., in Valiant's original PAC model [Val84] — even *weak* agnostic learning is computationally hard without distributional assumptions [GR06, FGKP06, Dan16]. To circumvent this computational intractability, a line of work has focused on the *distribution-specific* agnostic PAC model — where the learner has a priori information about the distribution on examples. In this setting, computationally efficient noise-tolerant learning algorithms are known [KKMS08, KLS09, ABL17, Dan15, DKS18, DKTZ20c] with various time-accuracy tradeoffs.

**Definition 1.3.9** (Distribution-Specific Agnostic Learning)**.** *Let $\mathcal{C}$ be a class of Boolean-valued functions on $\mathbb{R}^d$. Given i.i.d. labeled examples $(\mathbf{x}, y)$ from a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{\pm 1\}$, such that the marginal distribution $\mathcal{D}_{\mathbf{x}}$ is promised to lie in a known distribution family $\mathcal{F}$ and no assumptions are made on the labels, the goal of the learner is to output a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ with small misclassification error, $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \stackrel{\mathrm{def}}{=} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]$, as compared to the optimal misclassification error, $\mathrm{OPT} \stackrel{\mathrm{def}}{=} \inf_{g\in\mathcal{C}} \mathrm{err}^{\mathcal{D}}_{0-1}(g)$, by any function in the class.*

Some additional comments are in order on Definition 1.3.9. In *improper* learning, the only assumption about the hypothesis $h$ is that it is *polynomially evaluable*. In other words, we assume that $h \in \mathcal{H}$, where $\mathcal{H}$ is a (potentially complex) class of polynomially evaluable functions. In contrast, in *proper* learning we have the additional

requirement that the hypothesis $h$ is proper, i.e., $h \in \mathcal{C}$. These notions of learning are essentially equivalent in terms of sample complexity, but not always equivalent in terms of computational complexity. In particular, there exist concept classes that are efficiently improperly learnable, while proper learning is computationally hard.

The classical $L_1$-polynomial regression algorithm of [KKMS08] agnostically learns halfspaces under the Gaussian distribution, within error OPT$+\epsilon$, with sample complexity and runtime of $d^{\mathrm{poly}(1/\epsilon)}$. On the lower bound side, recent work has provided evidence that this complexity cannot be improved. Specifically, [DKZ20, GGK20, DKPZ21] obtained Statistical Query (SQ) lower bounds of $d^{\mathrm{poly}(1/\epsilon)}$ for this problem. That is, the complexity of this learning problem is well-understood.

The polynomial regression algorithm [KKMS08] is the only known agnostic learner for halfspaces and is inherently *improper*: instead of a halfspace, its output hypothesis is a degree-$k$ polynomial threshold function (PTF), i.e., the sign of a degree-$k$ polynomial, where $k = \mathrm{poly}(1/\epsilon)$. For the corresponding proper learning problem, prior to the present work, no non-trivial computational upper bound was known.

**Importance of Proper Learning.** While an improper hypothesis suffices for the purpose of prediction, an improper learner comes with some disadvantages. In our context, having such a complex output hypothesis requires spending $d^{\mathrm{poly}(1/\epsilon)}$ time for even evaluating the hypothesis on a single example. Moreover, storing the hypothesis function requires keeping track of the $d^{\mathrm{poly}(1/\epsilon)}$ coefficients defining the corresponding polynomial. In contrast, a proper hypothesis is easy to interpret and provides the most succinct representation. Specifically, a halfspace hypothesis would require only $O(d)$ time for evaluation and $O(d)$ storage space. Even though it is known that $d^{\mathrm{poly}(1/\epsilon)}$ time is required for identifying a good hypothesis during training, prior to this work, it was

not clear whether one can learn a succinct hypothesis that is more efficient at test time. The preceding discussion motivates the following natural question:

**Question 1.3.10.** Is there an efficient *proper* agnostic learner for halfspaces under Gaussian marginals?

The main result of this part (Theorem 1.3.12) is the first agnostic proper learner for this problem whose complexity qualitatively matches that of the known improper learner [KKMS08].

**Faster Runtime via Approximate Learning.** In view of the known SQ lower bounds for our problem [DKZ20, GGK20, DKPZ21], it is unlikely that the $d^{\mathrm{poly}(1/\epsilon)}$ runtime for agnostically learning halfspaces can be improved, even under the Gaussian distribution. A line of work [KLS09, ABL17, Dan15, DKS18, DKTZ20c] has focused on obtaining faster learning algorithms with relaxed error guarantees. Specifically, [ABL17] gave the first $\mathrm{poly}(d/\epsilon)$ time *constant-factor* approximation algorithm – i.e., an algorithm with misclassification error of $C \cdot \mathrm{OPT} + \epsilon$, for some universal constant $C > 1$ – for *homogeneous* halfspaces under the Gaussian, and, more generally, under any isotropic log-concave distribution. More recently, [Dan15] obtained a polynomial time approximation scheme (PTAS), i.e., an algorithm with error $(1+\gamma) \cdot \mathrm{OPT} + \epsilon$ and runtime $d^{\mathrm{poly}(1/\gamma)}/\mathrm{poly}(\epsilon)$, under the uniform distribution on the sphere (and, effectively, under the Gaussian distribution).

Interestingly, the constant factor approximation algorithm of [ABL17] is proper. On the other hand, the PTAS of [Dan15] is inherently improper, in part because it relies on the combination of the localization method [ABL17] and the (improper) polynomial regression algorithm [KKMS08]. It is thus natural to ask the following question:

**Question 1.3.11.** Is there a *proper* PTAS for agnostically learning halfspaces under Gaussian marginals?

As our second main contribution (Theorem 1.3.13), we give such a proper PTAS qualitatively matching the complexity of the known improper PTAS [Dan15].

## Main Results

In this part, we initiate a systematic algorithmic investigation of proper learning in the agnostic distribution-specific PAC model. Our main result is the first proper agnostic learner for the class of halfspaces under the Gaussian distribution, whose sample complexity and runtime qualitatively match the performance of the previously known improper algorithm.

**Theorem 1.3.12** (Proper Agnostic Learning of Halfspaces)**.** *Let $\mathcal{D}$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. There exists an algorithm that, given $\epsilon, \delta > 0$, and $N = d^{O(1/\epsilon^4)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ i.i.d. samples from $\mathcal{D}$, the algorithm runs in time $\mathrm{poly}(N) + (1/\epsilon)^{O(1/\epsilon^6)}\log(1/\delta)$, and computes a halfspace hypothesis $h$ such that, with probability at least $1-\delta$, it holds $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \leq \mathrm{OPT} + \epsilon$.*

Theorem 1.3.12 gives the first non-trivial agnostic proper learner for the class of halfspaces under natural distributional assumptions. The runtime of our algorithm is $d^{\mathrm{poly}(1/\epsilon)}$, which *qualitatively* matches the complexity of the improper polynomial regression algorithm [KKMS08] and is known to be qualitatively best possible in the SQ model [DKZ20, GGK20, DKPZ21].

The analysis of [KKMS08] established an upper bound of $d^{O(1/\epsilon^4)}$ on the complexity of polynomial regression for our setting. This bound was later improved to $d^{O(1/\epsilon^2)}$, using

optimal bounds on the underlying polynomial approximations [DKN10]. Designing a proper learner that *quantitatively* matches this upper bound is left as an interesting open question.

Our second main contribution is the first proper polynomial-time approximation scheme (PTAS) for the agnostic learning problem. In our context, a PTAS is an algorithm that, for any $\gamma, \epsilon > 0$, runs in time $d^{\mathrm{poly}(1/\gamma)}/\mathrm{poly}(\epsilon)$ and outputs a hypothesis $h$ satisfying $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \leq (1+\gamma)\mathrm{OPT} + \epsilon$. The parameter $\gamma > 0$ quantifies the approximation ratio of the algorithm. Prior work [Dan15] gave an improper PTAS for agnostically learning *homogeneous* halfspaces, i.e., halfspaces whose separating hyperplane goes through the origin. We give a proper algorithm for this problem.

**Theorem 1.3.13** (Proper PTAS for Agnostically Learning Halfspaces)**.** *Let $\mathcal{D}$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. There exists an algorithm that, given $\gamma, \epsilon, \delta > 0$ and $N = d^{\mathrm{poly}(1/\gamma)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ i.i.d. samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and computes a halfspace $h$ such that, with probability $1-\delta$, it holds $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \leq (1+\gamma)\mathrm{OPT} + \epsilon$, where OPT is the optimal misclassification error of any homogeneous halfspace.*

Theorem 1.3.13 gives the first proper PTAS for agnostically learning homogeneous halfspaces under any natural distributional assumptions and qualitatively matches the complexity of the improper PTAS by [Dan15]. We note that the homogeneity assumption is needed for technical reasons and is also required in the known improper learning algorithm. Obtaining a PTAS for agnostically learning arbitrary halfspaces remains an open problem (even for improper learners).

**Remark 1.3.14** (Extension to Other Non-Linear Activations)**.** While the focus of the current part is on the class of halfspaces, our algorithmic techniques are sufficiently

robust and naturally generalize to other activation functions, i.e., functions of the form $f(\mathbf{x}) = \sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is a well-behaved activation function. Specifically, in Section 4.4, we use our methods to develop the first proper agnostic learner for ReLU regression [DGK+20].

## Part II: Robustly Learning Single-Index Models

The second part of this thesis generalizes our investigation to broader classes of single-index models. Single-index models (SIMs) [Ich93, HJS01, HMS+04, DJS08, KS09, KKSK11, DH18] are a classical supervised learning model extensively studied in statistics and machine learning. SIMs capture the common assumption that the target function $f$ depends on an unknown direction $\mathbf{w}$, i.e., $f(\mathbf{x}) = u(\mathbf{w} \cdot \mathbf{x})$ for some link (a.k.a. activation) function $u : \mathbb{R} \mapsto \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$. Our focus is on *agnostic* learning of SIMs under mild distributional assumptions. Unlike the realizable setting, we allow adversarial noise, measuring contamination by the best in class $L_2^2$ loss,

$$\mathrm{OPT} \;=\; \min_{\|\mathbf{w}\| \le W} \mathbf{E}\Big[(u(\mathbf{w} \cdot \mathbf{x}) - y)^2\Big].$$

Our goal is to design learners that output a function $h(\mathbf{x})$ that achieve a constant-factor approximation, meaning that the output hypothesis $h$ satisfies

$$\mathbf{E}[(h(\mathbf{x}) - y)^2] \le C \cdot \mathrm{OPT} + \epsilon \quad \text{with } C = O(1)$$

in $\mathrm{poly}(d, 1/\epsilon)$ time.

### 1.3.4 Learning Generalized Linear Models under Well-Behaved Distributions

The recent success of deep learning has served as a practical motivation for the development of provable efficient learning algorithms for various natural classes of neural networks. Despite extensive investigation, our theoretical understanding of the assumptions under which neural networks are provably efficiently learnable remains somewhat limited. Here we focus on arguably the simplest possible setting of learning a Generalized Lineal Model (GLM), i.e., a real-valued function of the form $\mathbf{x} \mapsto \sigma(\mathbf{w} \cdot \mathbf{x})$, where $\mathbf{w}$ is the weight vector of parameters and $\sigma : \mathbb{R} \to \mathbb{R}$ is a fixed non-linear and monotone activation function. Concretely, the learning problem is the following: Given i.i.d. samples from a distribution $D$ on $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^d$ is the feature vector and $y \in \mathbb{R}$ is the corresponding label, our goal is to learn the underlying function in $L_2^2$-loss. That is, the learner's objective is to output a hypothesis $h : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbf{E}_{(\mathbf{x},y)\sim D}[(h(\mathbf{x}) - y)^2]$ is as small as possible, compared to the minimum possible loss $\mathrm{OPT} := \min_{\mathbf{w}\in\mathbb{R}^d} \mathbf{E}_{(\mathbf{x},y)\sim D}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2]$. Settings of particular interest for the activation $\sigma$ include the ReLU and sigmoid functions, corresponding to $\sigma(u) = \mathrm{ReLU}(u) \stackrel{\mathrm{def}}{=} \max\{0, u\}$ and $\sigma(u) \stackrel{\mathrm{def}}{=} 1/(1 + \exp(-u))$ respectively. Recall that a learning algorithm is called *proper* if the hypothesis $h$ is restricted to be of the form $h_{\widehat{\mathbf{w}}}(\mathbf{x}) = \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x})$. Throughout this work, we focus on developing efficient proper learners.

In the realizable case, i.e., when the labels $y$ are consistent with a function in the class, the above learning problem is known to be solvable in polynomial time for a range of activation functions. A line of work, see, e.g., [KS09, Sol17, YS20] and references therein, has shown that simple algorithms like gradient-descent efficiently converge to

an optimal solution under additional assumptions on the marginal distribution $D_{\mathbf{x}}$ on examples.

We focus on the *agnostic* learning model, where no realizability assumptions are made on the distribution $D$. Roughly speaking, the agnostic model corresponds to learning in the presence of adversarial label noise.

**Definition 1.3.15** (Learning Generalized Linear Models with Adversarial Noise)**.** *Fix $\epsilon, W > 0$, $\delta \in (0,1)$, and a class of distributions $\mathcal{G}$ on $\mathbb{R}^d$. Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be an activation function and $\mathcal{D}$ a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such that its $\mathbf{x}$-marginal belongs in $\mathcal{G}$. We define the population $L_2^2$-loss as $F^{D,\sigma}(\mathbf{w}) \triangleq (1/2)\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2]$. We say that $D$ is $(\epsilon, W)$-corrupted if $\inf_{\|\mathbf{w}\|_2 \leq W} F^{D,\sigma}(\mathbf{w}) \leq \epsilon$. For some $C \geq 1$, a $C$-approximate learner is given $\epsilon, \delta, W$ and i.i.d. labeled examples from $\mathcal{D}$ and outputs a function $h(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ such that, with probability at least $1 - \delta$, it holds $(1/2)\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(h(\mathbf{x}) - y)^2] \leq C\epsilon$.*

Some comments are in order. First, we note that the parameter $\epsilon$ quantifies the degree of contamination — in the sense that the closest function in the class has $L_2^2$-loss $\epsilon$. (Sometimes, this is denoted by OPT in the relevant literature.) The parameter $W$ quantifies the radius of the ball that contains an optimal vector $\mathbf{w}^*$. For uniformly bounded activations (e.g., sigmoids), one can remove the restriction on the norm of the weight vector $\mathbf{w}$, i.e., take $W = +\infty$. In this case, we call $D$ $\epsilon$-corrupted. When the distribution and activation are clear from the context, we will write $F(\mathbf{w})$ instead of $F^{D,\sigma}(\mathbf{w})$.

Ideally, one would like an efficient learner that succeeds for all marginal distributions and achieves optimal error of $\mathrm{OPT} + \epsilon$ (corresponding to $C = 1$). Unfortunately, known computational hardness results rule out this possibility. Even for the very special

case that the marginal distribution is Gaussian and the link function is known (e.g., a ReLU), there is strong evidence that any algorithm achieving error $\mathrm{OPT} + \epsilon$ requires $d^{\mathrm{poly}(1/\epsilon)}$ time [DKZ20, GGK20, DKPZ21, DKR23]. Moreover, even if we relax our goal to constant factor approximation (i.e., $C = O(1)$), distributional assumptions are required both for proper [Sím02, MR18] and improper learning [DKMR22b]. As a consequence, algorithmic research in this area has focused on constant factor approximate learners that succeed under mild distributional assumptions. On the other hand, even under Gaussian marginals, achieving error $\mathrm{OPT} + \epsilon$ (corresponding to $C = 1$ in Definition 1.3.15) requires time $d^{F(1/\epsilon)}$, for some function $F$ with $\lim_{u\to\infty} F(u) = \infty$ [GKK19, DKZ20, GGK20, DKPZ21]. These hardness results motivate the design of constant-factor approximate learners under "well-behaved" distributions.

**Related Prior Work** On the algorithmic side, [GKKT17] gave an algorithm with error $\mathrm{OPT} + \epsilon$ and runtime $\mathrm{poly}(d)2^{\mathrm{poly}(1/\epsilon)}$ that succeeds as long as the distribution on examples is supported on the unit sphere. [FCG20] study ReLU activations under structured distributions and show that gradient-descent on the $L_2$ loss converges to a weight vector with error $O(d\,\epsilon)$. The most relevant prior work is by [DGK+20] who gave a $\mathrm{poly}(d/\epsilon)$-time constant-factor approximate proper learner for ReLU activations under isotropic log-concave distributions. Their algorithm makes essential use of the ReLU activation. In fact, [DGK+20] asked whether efficient constant-factor approximations exist for other activations, including sigmoids. In this work, we answer this open question in the affirmative.

The aforementioned discussion motivates the following broad question:

**Question 1.3.16.** Is there an efficient constant-factor approximate learner for

generalized linear models (GLMs) under well-behaved distributions?

In this work, we answer this question in the affirmative for a range of activations including ReLUs and sigmoids and a variety of well-behaved distributions. In fact, we show that a simple gradient-based method on the $L_2^2$-loss suffices.

## Main Results

**Distributional Assumptions** We develop algorithms that are able to learn GLMs under a large class of structured distributions. We make mild distributional assumptions requiring only concentration, anti-concentration, and anti-anti-concentration on the $\mathbf{x}$-marginal of the examples. In particular, we consider the following class of well-behaved distributions.

**Definition 1.3.17** (Well-behaved Distributions)**.** *Let $L, R > 0$. An isotropic (i.e., zero mean and identity covariance) distribution $\mathcal{D}_{\mathbf{x}}$ on $\mathbb{R}^d$ is called $(L, R)$-well-behaved if for any projection $(\mathcal{D}_{\mathbf{x}})_V$ of $\mathcal{D}_{\mathbf{x}}$ onto a subspace $V$ of dimension at most two, the corresponding pdf $\gamma_V$ on $\mathbb{R}^2$ satisfies the following:*

- *For all $\mathbf{x} \in V$ such that $\|\mathbf{x}\|_\infty \leq R$ it holds $\gamma_V(\mathbf{x}) \geq L$ (anti-anti-concentration).*
- *For all $\mathbf{x} \in V$ it holds that $\gamma_V(\mathbf{x}) \leq (1/L)(e^{-L\|\mathbf{x}\|_2})$ (anti-concentration and concentration).*

*When the parameters $L, R$ are bounded above by universal constants (independent of the dimension), we will simply say that the distribution $\mathcal{D}_{\mathbf{x}}$ is well-behaved.*

The class of well-behaved distributions is fairly broad. Specifically, isotropic log-concave distributions are well-behaved, i.e., they are $(L, R)$-well-behaved for some $L, R = O(1)$,

see, e.g., [LV07, KLT09]. Similar assumptions were introduced in [DKTZ20a] and have been used in various classification and regression settings [DKTZ20c, DKTZ20b, DKK+20, DKK+21b, FCG20, FCG21, ZL21, ZFG21].

**Learning Sigmoidal Activations** Our first main result holds for a natural class of activations that roughly have "sigmoidal" shape.

**Definition 1.3.18** (Sigmoidal Activations)**.** *Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be a non-decreasing activation function and $\tau, \mu, \xi > 0$. We say that $\sigma$ is $(\tau, \mu, \xi)$-sigmoidal if it satisfies (a) $\sigma'(t) \geq \tau$, for all $t \in [-1, 1]$, and (b) $\sigma'(t) \leq \xi e^{-\mu|t|}$, for all $t \in \mathbb{R}$. We will simply say that an activation is "sigmoidal" when $\tau, \mu, \xi$ are universal constants.*

Arguably the most popular sigmoidal activation is the logistic activation or sigmoid, corresponding to $\sigma(t) = 1/(1+e^{-t})$. Other well-studied sigmoidal activations include the hyperbolic tangent, the Gaussian error function, and the ramp activation (see Figure 5.1). We note that all these activations satisfy the requirement of Definition 1.3.18 for some universal constants $\tau, \mu, \xi$. In what follows, we will simply refer to them as sigmoidal.

The most commonly used method to solve our learning problem in practice is to directly attempt to minimize the $L_2^2$-loss via (stochastic) gradient descent. Due to the non-convexity of the objective, this method is of a heuristic nature in general and comes with no theoretical guarantees, even in noiseless settings. In our setting, the situation is even more challenging due to the adversarial noise in the labels. Indeed, we show that the "vanilla" $L_2^2$-objective may contain bad local optima, even under Gaussian marginals. Specifically, even with an arbitrarily small amount of adversarial noise, the vanilla $L_2^2$ objective will have local-minima whose $L_2^2$-error is larger than $1/2$ (which is essentially trivial, since sigmoidal activations take values in $[-1, 1]$)( see Proposition 5.2.1 and

Figure 5.2).

Our main structural result for sigmoidal activations is that we can "correct" the optimization landscape of the $L_2^2$-loss by introducing a standard $\ell_2$-regularization term.

Standard gradient methods (such as SGD) are known to efficiently converge to stationary points of non-convex objectives under certain assumptions. By running any such method on our regularized objective, we readily obtain an efficient algorithm that outputs a weight vector with $L_2^2$-loss $O(\epsilon)$. This is already a new result in this context. Yet, black-box application of optimization results for finding stationary points of non-convex functions would resulting in a sample complexity with suboptimal dependence on $\epsilon$, up to polynomial factors.

Aiming towards an algorithmic result with near-optimal sample complexity, we perform a "white-box" analysis of gradient descent, leveraging the optimization landscape of sigmoidals. Specifically, we show that "vanilla" gradient descent, with a fixed step size, finds an approximately optimal solution when run on the empirical (regularized) $L_2^2$-loss with a near-optimal number of samples.

**Theorem 1.3.19** (Informal: Learning Sigmoidals via Gradient Descent)**.** *For sigmoidal activations and $\epsilon$-corrupted well-behaved distributions, gradient descent on the empirical regularized loss $F_\rho(\mathbf{w}) = F(\mathbf{w}) + (\rho/2)\|\mathbf{w}\|_2^2$ with $N = \widetilde{\Theta}(d/\epsilon)$ samples, converges, in* $\mathrm{poly}(1/\epsilon)$ *iterations, to a vector $\widetilde{\mathbf{w}}$ satisfying $F(\widetilde{\mathbf{w}}) \leq O(\epsilon)$ with high probability.*

Theorem 1.3.19 gives the first efficient constant-factor approximate learner for sigmoid activations in the presence of adversarial label noise, answering an open problem of [DGK+20]. As an additional bonus, our algorithm is simple and potentially practical (relying on gradient-descent), has near-optimal sample complexity (see Lemma C.1.3), and succeeds for a broad family of bounded activation functions (i.e., the ones satisfying

Definition 1.3.18).

**Learning Unbounded Activations** We next turn our attention to activation functions that are not uniformly bounded. The most popular such activation is the ReLU function $\sigma(t) = \text{Relu}(t) = \max(0, t)$. Our algorithmic results apply to the following class of unbounded activations.

**Definition 1.3.20** (Unbounded Activations)**.** *Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be a non-decreasing activation function and $\alpha, \lambda > 0$. We say that $\sigma$ is $(\alpha, \lambda)$-unbounded if it satisfies (a) $\sigma$ is $\lambda$-Lipschitz and (b) $\sigma'(t) \geq \alpha$, for all $t \in [0, +\infty)$. We will simply say that an activation is unbounded when the parameters $\alpha, \lambda$ are universal constants.*

We use the term "unbounded" for these activations, as they tend to $\infty$ as $t \to +\infty$. Most well-known unbounded activation functions such as the ReLU, Leaky-ReLU, ELU are $(\alpha, \lambda)$-unbounded for some absolute constants $\alpha, \lambda > 0$ (see Figure 5.4). For example, the ReLU activation is $(1, 1)$-unbounded.

To obtain near-optimal sample complexity, we again perform an "white-box" analysis of (approximate) gradient descent, showing that simply by initializing at $\mathbf{0}$ we can avoid bad stationary points.

**Theorem 1.3.21** (Informal: Learning Unbounded Activations via Gradient Descent)**.** *For unbounded activations and $(\epsilon, W)$-corrupted well-behaved distributions, (approximate) gradient descent on the $L_2^2$-loss $F(\cdot)$ with sample size $N = \widetilde{\Theta}(dW^2 \max(\text{polylog}(W/\epsilon), 1))$ and $\text{polylog}(1/\epsilon)$ iterations, converges to a vector $\widetilde{\mathbf{w}} \in \mathbb{R}^d$, satisfying $F(\widetilde{\mathbf{w}}) \leq O(\epsilon)$ with high probability.*

Theorem 1.3.21 is a broad generalization of the main result of [DGK+20], which gave a constant-factor approximation (using a different approach) for the special case of ReLU

activations under isotropic log-concave distributions. While the result of [DGK+20] was tailored to the ReLU activation, Theorem 5.1.7 works for a fairly broad class of unbounded activations (and a more general class of distributions). A key conceptual difference between the two results lies in the overall approach: The prior work [DGK+20] used a *convex* surrogate for optimization. In contrast, we leverage our structural result and directly optimize the natural non-convex objective.

Additionally, Theorem 1.3.21 achieves significantly better (and near-optimal) sample complexity as a function of both $d$ and $1/\epsilon$. In more detail, the algorithm of [DGK+20] had sample complexity $\Omega(d/\epsilon)$ (their results are phrased for the special case that $W = 1$), while our algorithm has sample complexity $\widetilde{O}(d)\mathrm{polylog}(1/\epsilon)$ — i.e., near-linear in $d$ and polylogarithmic in $1/\epsilon$.

### 1.3.5 Robustly Learning Single-Index Models under Well-Behaved Distributions

We now turn to the more general and realistic problem of learning a Single-Index Model (SIM) where the link function $u$ is also unknown, assumed only to be monotone and belong to the class of unbounded activations.

Classical works [KS09, KKSK11] studied the efficient learnability of SIMs for monotone and Lipschitz link functions and data distributed on the unit ball. These early algorithmic results succeed in the realizable setting (i.e., with clean labels) or in the presence of zero-mean label noise.

The focus of this part is on learning SIMs in the challenging agnostic (or adversarial label noise) model [Hau92, KSS94], where no assumptions are made on the labels of the examples and the goal is to compute a hypothesis that is competitive with the *best-fit*

function in the class. Importantly, as will be formalized below, we will not assume a priori knowledge of the link function. In more detail, let $\mathcal{D}$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ and $\mathcal{L}_2(h) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(h(\mathbf{x}) - y)^2]$ be the squared loss of the hypothesis $h : \mathbb{R}^d \to \mathbb{R}$ with respect to $\mathcal{D}$. Given i.i.d. samples from $\mathcal{D}$, the goal of the learner is to output a hypothesis $h$ with squared error competitive with OPT, where $\mathrm{OPT} = \inf_{f\in\mathcal{C}} \mathcal{L}_2(f)$ is the best attainable error by any function in the target class $\mathcal{C}$.

In the context of this part, the class $\mathcal{C}$ above is the class of SIMs, i.e., all functions of the form $f(\mathbf{x}) = u(\mathbf{w} \cdot \mathbf{x})$ where both the weight vector $\mathbf{w}$ and the link function $u$ are unknown. For this task to be even information-theoretically solvable, one requires some assumptions on the vector $\mathbf{w}$ and the link function $u$. We will assume, as is standard, that the $\ell_2$-norm of $\mathbf{w}$ is bounded by a parameter $W$. We will similarly assume that the link function lies in a family of well-behaved functions that are monotone and satisfy certain Lipschitz properties (see Definition 1.3.20).

For a weight vector $\mathbf{w}$ and link function $u$, the $L_2^2$ loss of the SIM hypothesis $u(\mathbf{w} \cdot \mathbf{x})$ (defined by $u$ and $\mathbf{w}$) is $\mathcal{L}_2(\mathbf{w}; u) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w} \cdot \mathbf{x}) - y)^2]$. Our problem of robustly learning SIMs is defined as follows.

**Problem 1.3.22** (Robustly Learning Single-Index Models)**.** *Fix a class of distributions $\mathcal{G}$ on $\mathbb{R}^d$ and a class of link functions[3] $\mathcal{F}$. Let $\mathcal{D}$ be a distribution of labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such that its $\mathbf{x}$-marginal $\mathcal{D}_\mathbf{x}$ belongs to $\mathcal{G}$. We say that an algorithm is a* $C$-approximate proper SIM learner*, for some $C \geq 1$, if given $\epsilon > 0$, $W > 0$, and i.i.d. samples from $\mathcal{D}$, the algorithm outputs a link function $\hat{u} \in \mathcal{F}$ and a vector $\widehat{\mathbf{w}} \in \mathbb{R}^d$ such that with high probability it holds $\mathcal{L}_2(\widehat{\mathbf{w}}; \hat{u}) \leq C\,\mathrm{OPT} + \epsilon$, where $\mathrm{OPT} \triangleq \min_{\|\mathbf{w}\|_2\leq W, u\in\mathcal{F}} \mathcal{L}_2(\mathbf{w}; u)$.*

[3]Throughout this thesis, we will use the terms "link function" and "activation" interchangeably.

We use $u^*(\mathbf{w}^* \cdot \mathbf{x})$ to denote a fixed (but arbitrary) optimal solution to the above learning problem, i.e., one satisfying $\mathcal{L}_2(\mathbf{w}^*; u^*) = \mathrm{OPT}$.

Recent works [DGK+20, DKTZ22a, ATV23, WZDD23] gave efficient, constant factor approximate learners, under natural distributional assumptions, for the special case of Problem 1.3.22 where the link function is known *a priori* (see also [FCG20]). For the general setting, the only prior algorithmic result was recently obtained in [GGKS23]. Specifically, [GGKS23] gave an efficient algorithm that succeeds for the class of monotone 1-Lipschitz link functions and any marginal distribution with second moment bounded by $\lambda$. Their algorithm achieves $L_2^2$ error

$$O(W\sqrt{\lambda}\sqrt{\mathrm{OPT}}) + \epsilon \tag{1.1}$$

under the assumption that the labels are bounded in $[0, 1]$. The error guarantee (1.1) is substantially weaker — both qualitatively and quantitatively — from the goal of this work. Firstly, the dependence on OPT scales with its square root, as opposed to linearly. Secondly, and arguably more importantly, the multiplicative factor inside the big-O scales (linearly) with the diameter of the space $W$.

Interestingly, [GGKS23] showed — via a hardness construction from [DKMR22b] — that, under their distributional assumptions, a multiplicative dependence on $W$ (in the error guarantee) is inherent for efficient algorithms. That is, to obtain an efficient constant factor approximation, it is necessary to restrict ourselves to distributions with *additional* structural properties. This discussion raises the following question:

**Question 1.3.23.** Can we obtain efficient *constant factor* learners for Problem 1.3.22 under mild distributional assumptions?

The natural goal here is to match the guarantees of known algorithmic results for the special case of known link function [DKTZ22a, WZDD23].

*As our main contribution, we answer this question in the affirmative.* That is, we give *the first* efficient constant-factor approximate learner that succeeds for natural and broad families of distributions (including log-concave distributions) and a broad class of link functions. We emphasize that this is the first polynomial-time constant factor approximate learner even for Gaussian marginals and for any nontrivial class of link functions. Roughly speaking, our distributional assumptions require concentration and (anti)-anti-concentration (see Definition 1.3.17).

## Main Results

Focusing on the class of Well-Behaved distributions (Definition 1.3.17) and the class of Unbounded Activations (Definition 1.3.20) we present our results. A simplified version of our main algorithmic result is as follows (see Theorem 6.3.2 for a more detailed statement):

**Theorem 1.3.24** (Main Algorithmic Result, Informal)**.** *Given Problem 1.3.22, where $\mathcal{G}$ is the class of $(L,R)$-well behaved distributions with $L, R = O(1)$ and $\mathcal{F} = \mathcal{U}_{(a,b)}$ such that $(1/a), b = O(1)$, there is an algorithm that draws $N = \mathrm{poly}(W)\tilde{O}(d/\epsilon^2)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(N,d)$ time, and outputs a hypothesis $\hat{u}(\widehat{\mathbf{w}} \cdot \mathbf{x})$ with $\hat{u} \in \mathcal{U}_{(a,b)}$, $\|\widehat{\mathbf{w}}\|_2 \leq W$ such that $\mathcal{L}_2(\widehat{\mathbf{w}}; \hat{u}) = C\,\mathrm{OPT} + \epsilon$ with high probability, where $C > 0$ is an absolute constant.*

We reiterate that the approximation factor $C$ in Theorem 1.3.24 is a universal constant, independent of the dimension and the diameter of the space. That is, our main result

provides the first efficient learning algorithm achieving a constant factor approximation, even for the most basic case of Gaussian data and any non-trivial class of link functions.

## Technical Overview

When it comes to learning SIMs in the agnostic model with target error $C\mathrm{OPT} + \epsilon$, to the best of our knowledge, all prior work that achieves such a guarantee with $C$ being an absolute constant only applies to the special case of known link function $u^*$. Such results are established by proving growth conditions (local error bounds) that relate either the $L_2^2$ loss or a surrogate loss to (squared) distance to the set of target solutions, using assumptions about the link function and the data distribution, such as concentration and (anti-)anti-concentration [DGK+20, DKTZ22a, WZDD23]. Among these, most relevant to our work is [WZDD23], which proved a "sharpness" property for the convex surrogate function defined by

$$\mathcal{L}_{\mathrm{sur}}(\mathbf{w}; u) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\int_0^{\mathbf{w}\cdot\mathbf{x}} (u(r) - y)\,\mathrm{d}r\right], \tag{1.2}$$

based on certain assumptions about the link function (that are the same as ours) and distributional assumptions (that are somewhat weaker but comparable to ours). Their sharpness result corresponds to guaranteeing that for vectors $\mathbf{w}$ that are not already $O(\mathrm{OPT}) + \epsilon$ accurate solutions, the following holds:

$$\nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}; u^*) \cdot (\mathbf{w} - \mathbf{w}^*) \gtrsim \|\mathbf{w} - \mathbf{w}^*\|_2^2, \tag{1.3}$$

where $\mathbf{w}^*$ is a vector that achieves error $O(\mathrm{OPT}) + \epsilon$ and $u^*$ is the (a priori known) link function.

One may hope that the sharpness result of [WZDD23] can be generalized to the case of unknown link function and leveraged to obtain constant factor robust learners in this more general setting. However, as we discuss below, such direct generalizations are not possible and there are several technical challenges that had to be overcome in our work. To illustrate some of the intricacies, consider first the following example.

**Example 1.3.25.** Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{w} = (1/2)\mathbf{w}^*$, where $\mathbf{w}^*$ is an arbitrary but fixed target unit vector. Let $b > 2a$. Suppose that the link function at hand is $u(z) = bz$ and the target link function is $u^*(z) = az$. Observe that both $u, u^* \in \mathcal{U}_{(a,b)}$, as required by our model. Furthermore, suppose there is no label noise, in which case $\mathrm{OPT} = 0$. Note that the $L_2^2$ error of $u(\mathbf{w} \cdot \mathbf{x})$ in this case is

$$\begin{aligned}\mathcal{L}_2(\mathbf{w}; u) &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(u(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \\ &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}[(b/2-a)^2 z^2] = (b/2-a)^2 = \Theta(1).\end{aligned}$$

However, the gradient of the surrogate loss, $\nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}; u) = \mathbf{E}[(u(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))\mathbf{x}]$, is negatively correlated with $\mathbf{w} - \mathbf{w}^*$, i.e., $\nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}; u)\cdot(\mathbf{w} - \mathbf{w}^*) < 0$, contrary to what we would hope for if a sharpness property as in [WZDD23] were to hold. Thus, although $\mathbf{w}$ and $u$ are both still far away from the target parameters $\mathbf{w}^*$ and $u^*$, the gradient of the surrogate loss cannot provide useful information about the direction in which to update $\mathbf{w}$.

What Example 1.3.25 demonstrates is that we cannot hope for the surrogate loss to satisfy a local error bound for an arbitrary parameter pair $(u, \mathbf{w})$ that would guide the convergence of an algorithm toward a target parameter pair $(u^*, \mathbf{w}^*)$. This seemingly insurmountable obstacle is surpassed by observing that we do not, in fact, need the

surrogate loss to contain a "signal" that would guide us toward target parameters for an *arbitrary* pair $(u, \mathbf{w})$. Instead, we can restrict our attention to pairs $(u, \mathbf{w})$ satisfying that $u$ is a "reasonably good" link function for the vector $\mathbf{w}$. Ideally, we would like to only consider link functions $u$ that minimize the $L_2^2$ loss — considering that $u^*$ must minimize the $L_2^2$ loss for a given, fixed $\mathbf{w}^*$ — but it is unclear how to achieve that in a statistically and computationally efficient manner. As a natural approach, we consider link functions that are the best fitting functions in an empirical distribution sense. In particular, given a sample set $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ and a parameter $\mathbf{w}$, we select a function $\hat{u}_{\mathbf{w}}$ that solves the following (convex) optimization problem:

$$\hat{u}_{\mathbf{w}} \in \operatorname*{argmin}_{u \in \mathcal{U}_{(a,b)}} \frac{1}{m} \sum_{i=1}^{m} (u(\mathbf{w} \cdot \mathbf{x}^{(i)}) - y^{(i)})^2. \tag{P}$$

For notational simplicity, we drop the parameter $\mathbf{w}^t$ from $\hat{u}_{\mathbf{w}^t}$ and use $\hat{u}^t$ instead. It is worth pointing out here that in general the problem of finding the best function that minimizes the $L_2^2$ error fails under the category of non-parametric regression, which unfortunately requires exponentially many samples (namely, $\Omega(1/\epsilon^d)$). Fortunately, in our setting, we are looking for the best function that lies in a *one-dimensional space*. Therefore, instead of looking at all possible directions, we can project all the points of the sample set $S$ to the direction $\mathbf{w}$ and find the best fitting link function efficiently. We provide the full details for efficiently solving the optimization problem (P) in Section D.2.

Having set on the "best-fit" link functions in the sense of the problem (P), the next obstacle one encounters when trying to prove a "sharpness-like" result is that *neither the $L_2^2$ loss nor its surrogate convey information about the scale* of $\mathbf{w}$ and $\mathbf{w}^*$. This is because models determined by $u, \mathbf{w}$ and $u/c, c\mathbf{w}$ for some parameter $c > 0$ can have the same value of both loss functions. Thus, it seems unlikely that a more

traditional local error bound, as in (1.3), can be established in general, for either the surrogate loss or the original $L_2^2$ loss. Instead, we prove a weaker property that establishes strong correlation between the gradient of the empirical surrogate loss $\nabla \widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t) = (1/m) \sum_{i=1}^m (\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - y^{(i)})\mathbf{x}^{(i)}$ and the direction $\mathbf{w}^t - \mathbf{w}^*$ that holds whenever $\mathbf{w}^t$ is not an $O(\text{OPT}) + \epsilon$ error solution and which is *independent of the scale* of $\mathbf{w}^t$. This constitutes our key structural result, stated as Proposition 6.2.1 and discussed in detail in Section 6.2. We further discuss how this result relates to classical and recent local error bounds in Section (6.1.1).

In addition to this weaker version of a sharpness property, we further prove in Corollary 6.2.5 that given a parameter $\mathbf{w}^t$ and a dataset of $m$ samples from $\mathcal{D}$, the activation $\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x})$ generated by optimizing the empirical risk on the dataset as in (P) satisfies $\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \lesssim b^2 \|\mathbf{w}^t - \mathbf{w}^*\|_2^2$ with high probability. As a result, we can guarantee that when $\|\mathbf{w}^t - \mathbf{w}^*\|_2$ decreases, the $L_2^2$ distance between $\hat{u}^t$ and $u^*$ diminishes as well. This is crucial, since without such a coupling we would not be able to argue about convergence over both model parameters $u, \mathbf{w}$.

Leveraging these results, we arrive at an algorithm that alternates between "gradient descent-style" updates for $\mathbf{w}$ and best-fit updates for $u$. We note in passing that similar alternating updates have been used in classical work on SIM learning in the less challenging, non-agnostic setting [KKSK11]. In more detail, our algorithm fixes the scale $\beta$ of $\|\mathbf{w}^t\|_2$ and alternates between taking a Riemannian gradient descent step on a sphere for $\mathbf{w}^t$ with respect to the empirical surrogate loss and solving (P). The unknown scale for the true parameter vector $\mathbf{w}^*$ is resolved by applying this approach using $\beta$ chosen from a sufficiently fine grid of the interval $[0, W]$ and employing a testing procedure at the end to select the best parameter vector. Although the idea is simple,

the proof of correctness is quite technical, as it requires ensuring that the entire process does not accumulate spurious errors arising from the stochastic nature of the problem, adversarial labels, and approximate minimization of the surrogate loss, and, as a result, that it converges to the target error.

### 1.3.6 Learning Generalized Linear Models with Gaussian Marginals

The previous part of this thesis established robust learning guarantees for GLMs models under the general family of well-behaved distributions (Definition 1.3.17) . Those results, however, required the activation functions to belong to specific classes, such as sigmoidal or unbounded activations (Definitions 1.3.18 and 1.3.20), which possess strong structural properties. This naturally raises the question: if we restrict the distribution from the broad well-behaved class to the specific special case of the standard normal, can we in turn handle a more general class of activations? This prompts the following question:

**Question 1.3.26.** Is there an efficient *constant-factor* approximate learner for monotone Lipschitz GLMs under Gaussian marginals?

This chapter affirmatively answers this question by showing that for the Gaussian distribution, we can develop an efficient algorithm for nearly all monotone activations, a significantly broader functional class.

**Theorem 1.3.27** (Robustly Learning Monotone & Lipschitz GLMs)**.** *There exists an algorithm with the following performance guarantee: For any known monotone and $b$-Lipschitz activation $\sigma$, given $\epsilon > 0$, $W > 0$, and $N = \tilde{\Theta}(d(bW)^2/\epsilon + d/\epsilon^2)$ samples, the algorithm runs in* $\mathrm{poly}(d, N)$ *time and returns a vector $\widehat{\mathbf{w}}$ such that with high probability, $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}) - y)^2] \leq C\mathrm{OPT} + \epsilon$, where $C$ is an absolute constant independent of $\epsilon, d, b, W$.*

We emphasize that the approximation ratio of our algorithm is a universal constant—independent of the dimension, the desired accuracy, the Lipschitz constant, and the radius of the space.

The key qualitative difference between prior work and Theorem 1.3.27 is in the assumptions on the activation. Specifically, prior constant-factor GLM learners succeed for a much smaller subclass of activations. In fact, our main algorithmic result (Theorem 7.4.1) applies more generally, encompassing all monotone activations with bounded $(2+\zeta)$ moment, for any $\zeta > 0$ (Corollary 7.4.4). This in particular implies that the case of LTFs fits in our setting. We stress here that some assumption on top of monotonicity is information-theoretically necessary, even for realizable learning (Theorem 7.5.13).

**Comparison to Prior Work** [GGKS23] gave an efficient GLM learner for monotone Lipschitz activations and marginal distribution with bounded second moment. However, the error of their algorithm scales linearly with $W$ and the Lipschitz constant. [WZDD23, ZWDD24] studied Problem 1.3.22 under 'well-behaved' distributions, where $\sigma$ is monotone and $(a, b)$ unbounded, meaning that $|\sigma'(z)| \leq b$ and $\sigma'(z) \geq a$ when $z \geq 0$. They provided an efficient algorithm with error $O(\mathrm{poly}(b/a))\mathrm{OPT} + \epsilon$. Note that when $a = 0$, this error guarantee is vacuous. More recently, [WZDD24] studied the same problem under Gaussian marginals for activations with bounded information-exponent. The approximation ratio of their method inherently scales polynomially with the radius $W$ of the space.

**Remark** In the sequel, we assume that the scale of the target vector $\mathbf{w}^*$, $\|\mathbf{w}^*\|_2$, is known and, by, rescaling the space, we optimize $\mathbf{w}$ on the unit sphere. The unknown scale of $\mathbf{w}^*$ can be resolved by a simple grid search. For our approach, this rescaling is w.l.o.g. because—unlike in prior work [WZDD23, WZDD24, ZWDD24]—the approximation

ratio of our algorithm is **independent** of any problem parameters. For a formal justification, see Remark 7.5.3 and Lemma 7.5.4.

## Technical Overview

Our work relies on three main technical ingredients: (1) data augmentation, which we use as a method to mitigate the effect of the adversarial label noise, (2) an optimization-theoretic local error bound, which in our work is a structural result that identifies the "signal" vector field that guides the algorithm toward the set of target solutions, (3) a suite of structural results for $(B, L)$-regular monotone activations (see Definition 7.3.1), leveraging their piecewise-constant approximations, smoothing through data augmentation, and representation via Hermite polynomials.

**Data Augmentation** Data augmentation encompasses a broad set of techniques for modifying or artificially generating data to enhance learning and estimation tasks. In the context of our work, data augmentation refers to the injection of Gaussian noise into the data vectors $\mathbf{x}$ while retaining the same labels. In particular, given any labeled example $(\mathbf{x}, y) \sim \mathcal{D}$ and a parameter $\rho \in (0, 1)$, the considered data augmentation process generates labeled examples $(\tilde{\mathbf{x}}, y)$, where $\tilde{\mathbf{x}} = \rho\mathbf{x} + \sqrt{1-\rho^2}\mathbf{z}$ and $\mathbf{z}$ is an independently generated sample from the standard normal distribution. While this type of data augmentation is a common empirical technique in machine learning, it is considered to be a wild card: although sometimes helpful, it can also be detrimental to learning guarantees (see, e.g., [YGLS+19, LKDM24]). Thus, on a conceptual level, one of our contributions is showing that for the considered GLM learning task, data augmentation is provably beneficial.

The effect of data augmentation on the considered GLM learning task is that it

simulates the Ornstein–Uhlenbeck semigroup $\mathrm{T}_\rho f(t) \stackrel{\text{def}}{=} \mathbf{E}_{z\sim\mathcal{N}(0,1)}[f(\rho t + \sqrt{1-\rho^2}z)]$ applied to any function $f(\mathbf{w}\cdot\mathbf{x})$. This process smoothens the function $f$ and induces other regularity properties. Unlike the common use of smoothing in the optimization literature, where the key utilized properties are continuity and smoothness of the smoothed objective function (see, e.g., [NS17, DBW12, BJL+19, DG24]), in our work, the key feature is the effect of injected noise on enhancing the signal in the data, as explained below.

Suppose we were given a GLM learning task. Since the goal of a learning algorithm is to minimize the mean squared loss $\mathcal{L}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2]$, a natural approach is to follow a gradient field associated with the error $\sigma(\mathbf{w}\cdot\mathbf{x}) - y$. Indeed, all prior work for this task proceeds by applying (stochastic) gradient-based algorithms to either the original squared loss $\mathcal{L}(\mathbf{w})$ or its surrogate $\mathcal{L}_{\text{sur}}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\int_0^{\mathbf{w}\cdot\mathbf{x}}(\sigma(t) - y)\mathrm{d}t\right]$. In either case, the associated gradient field can be represented by $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)h(\mathbf{w}\cdot\mathbf{x})\mathbf{x}]$ for some function $h$ (in particular, for the squared loss $h(\mathbf{w}\cdot\mathbf{x}) = 2\sigma'(\mathbf{w}\cdot\mathbf{x})$, while for the surrogate loss, $h \equiv 2$). Since we are considering optimizing $\mathbf{w}$ over the unit sphere (see Remark 7.5.3), the relevant information for updating $\mathbf{w}$ is in its orthogonal complement (as we are not changing its length), so it suffices to consider $\mathbf{g}(\mathbf{w}, h) \stackrel{\text{def}}{=} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)h(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp\mathbf{w}}]$.

Intuitively, if we can show that $-\mathbf{g}(\mathbf{w}, h)$ strongly correlates with $\mathbf{w}^*$, then this information can be used to update $\mathbf{w}$ to better align with $\mathbf{w}^*$, until we reach the target approximation error. Observe first that, as the Gaussian distribution is independent across orthogonal directions, we have $-\mathbf{g}(\mathbf{w}, h)\cdot\mathbf{w}^* = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[yh(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp\mathbf{w}}\cdot\mathbf{w}^*]$. Writing $y = \sigma(\mathbf{w}^*\cdot\mathbf{x}) + y - \sigma(\mathbf{w}^*\cdot\mathbf{x})$, the quantity $-\mathbf{g}(\mathbf{w}, h)\cdot\mathbf{w}^*$ can be decomposed into two parts: (i) corresponding to "clean" labels $\sigma(\mathbf{w}^*\cdot\mathbf{x})$, and (ii) corresponding

to label noise $y - \sigma(\mathbf{w}^* \cdot \mathbf{x})$. Letting $\theta$ denote the angle between $\mathbf{w}$ and $\mathbf{w}^*$, it is possible to argue (using Stein's lemma, see Fact 7.1.4) that the "clean label" portion of $-\mathbf{g}(\mathbf{w}, h) \cdot \mathbf{w}^*$ equals $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\sigma'(\mathbf{w}^* \cdot \mathbf{x})h(\mathbf{w} \cdot \mathbf{x})] \sin^2 \theta$. For the "noisy" label portion, by the Cauchy-Schwarz inequality and the definition of OPT, we can write

$$- \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))h(\mathbf{w} \cdot \mathbf{x})(\mathbf{w}^* \cdot \mathbf{x}^{\perp \mathbf{w}})] \leq \sqrt{\mathrm{OPT}}\|h\|_{L_2} \sin(\theta(\mathbf{w}, \mathbf{w}^*)),$$

where $\|h\|_{L_2} := \left(\mathbf{E}_{z\sim\mathcal{N}(0,1)}[h^2(z)]\right)^{1/2}$. Since labels are adversarial, the inequality can in fact be made to hold with equality. Thus, summarizing the above discussion, we have

$$-\mathbf{g}(\mathbf{w}, h) \cdot \mathbf{w}^* \geq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\sigma'(\mathbf{w}^* \cdot \mathbf{x})h(\mathbf{w} \cdot \mathbf{x})] \sin^2 \theta - \sqrt{\mathrm{OPT}}\|h\|_{L_2} \sin \theta. \tag{1.4}$$

We can assume w.l.o.g. that $\|h\|_{L_2} = 1$, since dividing both sides by $\|h\|_{L_2}$ would give us the same conclusion. For $-\mathbf{g}(\mathbf{w}, h)$ to contain a useful signal guiding the algorithm towards target solutions, we need that $-\mathbf{g}(\mathbf{w}, h) \cdot \mathbf{w}^* > 0$, for which we ought to argue that $G(h) := \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\sigma'(\mathbf{w}^* \cdot \mathbf{x})h(\mathbf{w} \cdot \mathbf{x})] > 0$. It is possible to argue that $G(h)$ is maximized for the "ideal" choice of $h(\mathbf{w}\cdot\mathbf{x}) \propto \sigma'(\cos\theta \mathbf{w}\cdot\mathbf{x} + \sin\theta \mathbf{z}\cdot\mathbf{x})$ with independently sampled $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This can equivalently be seen as applying the Ornstein–Uhlenbeck semi-group with parameter $\rho = \cos\theta$ to $\sigma'$, which motivates its use in our work. Of course, since $\cos\theta$ is not known to the algorithm, the smoothing parameter $\rho$ needs to be carefully chosen and adjusted between the algorithm updates.

**Alignment and Optimization** Local error bounds have long history in optimization and represent some of the most important technical tools for establishing iterate convergence to target solutions, especially in the context of gradient-based algorithms (see, e.g., [Pan97]). Broadly speaking, local error bounds are inequalities that bound

below some measure of the problem "residual" or error by a measure of distance to the target solution set. "Local" in the name refers to such inequalities being valid only in a local region around the target solution set. Within learning theory and in the context of GLM learning, they have played a crucial role in the analysis of (stochastic) gradient-based algorithms [MBM18, WZDD23, ZWDD24, WZDD24].

Our main structural result, stated in Proposition 7.2.3, is a local error bound for which the residual is $-\mathbf{g}(\mathbf{w}) \cdot \mathbf{w}^*$ for the gradient field $\mathbf{g}(\mathbf{w})$ corresponding to the data augmented squared loss function, as discussed above. This residual has the meaning of the "alignment" between $-\mathbf{g}(\mathbf{w})$ and $\mathbf{w}^*$. Specifically, we prove that in a local region around a certain set $\mathcal{S}$, the following inequality holds for any $\rho \in (0, 1)$ and $\theta$ being the angle between $\mathbf{w}, \mathbf{w}^*$ :

$$-\mathbf{g}(\mathbf{w}) \cdot \mathbf{w}^* \geq (2/3)\|\mathrm{T}_{\sqrt{\rho\,\cos\theta}}\sigma'\|_{L_2}^2 \sin^2\theta. \tag{1.5}$$

Observe that, since we are optimizing over the unit sphere, $\|\mathbf{w} - \mathbf{w}^*\|_2^2 \approx \sin^2(\theta)$. This structural result allows us to argue that, provided an initial parameter vector $\mathbf{w}_0$ for which (1.5) holds, we can update iterates $\mathbf{w}$ to contract the angle $\theta$, until the set $\mathcal{S}$ is reached. While this general idea seems relatively simple, making it work requires a rather technical argument to (i) ensure we can initialize the algorithm in the region where (1.5) holds, (ii) adjust the value of $\rho$ between the algorithm updates to ensure we remain in the region where (1.5) applies, and (iii) argue that all parameter vectors in $\mathcal{S}$ are $O(\mathrm{OPT})$ approximate solutions. Part (ii) is handled using an intricate inductive argument. Parts (i) and (iii) are addressed by proving a series of structural results for the class of $(B, L)$-regular monotone activations, discussed below.

**Approximation and Regularity of Monotone Functions** While handling arbitrary monotone functions is provably impossible, we show that fairly minimal assumptions suffice for our approach. In particular, we handle all $(B, L)$-regular monotone activations, which we show can be well-approximated by monotone piecewise-constant (staircase) functions. In more detail, instead of directly proving the desired properties of monotone $(B, L)$-regular activations, we consider the class of staircase functions, which only increase within a compact interval (and are constant outside it). For this class of staircase functions, we prove that the high-degree terms in their Hermite expansion —namely, terms with degree $> 1/\theta^2$ for $\theta$ sufficiently small—are bounded by $\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 \sin^2\theta$, and, further, this result extends to all $(B, L)$-regular functions (Proposition 7.4.6). Proving this structural result relies on auxiliary results relating Ornstein–Uhlenbeck semigroups of activations and their derivatives that may be of independent interest. Proposition 7.4.6 is then used to argue that the target set $\mathcal{S}$ to which the iterates of the algorithm converge only contains vectors with $L_2^2$ error $O(\mathrm{OPT})$, addressing the aforementioned issue (iii).

Since the result from Proposition 7.4.6 only applies for sufficiently small $\theta$, we need to argue that the algorithm can be appropriately initialized. In particular, random initialization is insufficient since we need roughly that $\theta_0 \leq O((\log(1/\epsilon))^{-1/2})$. To address this requirement, we apply a label transformation $\tilde{y} = \mathbb{1}\{y \geq t\}$ for a carefully chosen threshold $t$, where $\mathbb{1}$ denotes the indicator function. In particular, to select $t$, we leverage the staircase approximation of monotone functions discussed above. We argue that the problem reduces to learning $\mathrm{sign}(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - t)$, which is an instance of learning halfspaces with adversarial noise.[4] In particular, we argue that constant approximate

[4]Note here that $\mathrm{sign}(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - t)$ being a halfspace crucially relies on $\sigma$ being monotone.

solutions to this halfspace learning problem suffice for our initialization.

### 1.3.7 Learning Single-Index Models with Gaussian Marginals

The preceding discussion demonstrates that under the Gaussian distribution, we can robustly learn single-index models for a wide range of monotone activations, provided the activation function is known. The next step is to answer the following question:

**Question 1.3.28.** Does there exist an efficient constant-factor approximation algorithm for learning monotone & Lipschitz SIMs under Gaussian inputs?

As our main contribution, we resolve this question in the affirmative.

**Theorem 1.3.29** (Robustly Learning Monotone & Lipschitz SIMs)**.** *There exists a universal constant $C > 1$ such that the following holds. Let $\mathcal{C}$ be the class of all SIMs on $\mathbb{R}^d$ with a monotone and $L$-Lipschitz activation. There is an algorithm that, given $\epsilon > 0$ and $W > 0$, draws $N = d^2\mathrm{poly}(1/\epsilon, W, L)$ samples, runs in $\mathrm{poly}(N)$ time, and returns a predictor $(\widehat{\sigma}, \widehat{\mathbf{w}})$ such that, with high probability, $\mathcal{L}_2(\widehat{\mathbf{w}}; \widehat{\sigma}) \leq C\mathrm{OPT} + \epsilon$.*

It is worth noting that our algorithm does not require the Lipschitz assumption on the unknown activation. In fact, it applies for the broader class of monotone activations with bounded $(2+\zeta)$ moment, for any $\zeta > 0$ (Corollary 8.2.7). This in particular implies that the case of LTFs fits in our class. As pointed out in [ZWDD25], this assumption on top of monotonicity is essentially information-theoretically necessary (in particular, bounded second moment alone does not suffice).

## Technical Overview

Before getting to the details of our algorithm and its analysis, we first provide "the big picture" of the conceptual novelty of our approach. Prior work on agnostic learning of GLMs and SIMs has as one of its main components a gradient-based algorithm applied to either the square loss [DKTZ22a], its smoothing [ZWDD25], or a suitable surrogate [GGKS23, WZDD23, WZDD24, ZWDD24]. Such approaches are reasonable, considering the long history of gradient-based optimization methods and their applications in learning. However, the considered problems are not black-box, and if we think about optimization algorithms as choosing vector fields to guide the algorithm updates, then the "negative gradient" appearing in gradient-based methods is but one possible choice that could work.

On a conceptual level, our work makes the case for stepping outside the usual boundaries of gradient-based methods and instead looking to directly design a vector field that can be computed from the information given to the algorithm and that carries useful information about the location of target solutions. In the spirit of prior work such as [WZDD23, WZDD24, ZWDD24, ZWDD25], this useful information is captured by the alignment of the chosen vector field and a target parameter vector $\mathbf{w}^*$.

We point out here that this a rather nontrivial goal: even in the case of GLMs (known activation), the negative gradient of the square loss (as used in e.g., [DKTZ22a]) or a standard surrogate loss (as used in e.g., [WZDD23]) can "point in the wrong direction" on some monotone Lipschitz functions (see the discussion in [ZWDD25]). This issue was addressed in the recent work [ZWDD25] by demonstrating that the negative (Riemannian) gradient of the squared loss with Gaussian-smoothed activation, $\mathcal{L}_\rho(\mathbf{w};\sigma) = (1/(2\rho))\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \mathrm{T}_\rho\sigma(\mathbf{w}\cdot\mathbf{x}))^2]$, correlates with $\mathbf{w}^*$ for an appropriate

choice of the smoothing parameter $\rho \in (0,1)$. Here $\mathrm{T}_\rho\sigma$, $\rho \in (0,1)$, is the smoothed activation using the Ornstein–Uhlenbeck semi-group.

Briefly, [ZWDD25] shows that when $\mathbf{w}$ is still far away from the target $\mathbf{w}^*$, the Riemannian gradient $-\nabla_{\mathbf{w}}\mathcal{L}_\rho(\mathbf{w};\sigma) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \mathrm{T}_\rho\sigma(\mathbf{w}\cdot\mathbf{x}))\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp\mathbf{w}}] = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp\mathbf{w}}]$[5] possesses the 'gradient alignment' property, that is (denoting $\theta \overset{\text{def}}{=} \theta(\mathbf{w},\mathbf{w}^*)$),

$$-\nabla_{\mathbf{w}}\mathcal{L}_\rho(\mathbf{w};\sigma)\cdot\mathbf{w}^* \geq \sin^2\theta\,\mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})] - \sqrt{\mathrm{OPT}}\sin\theta\|\mathrm{T}_\rho\sigma'\|_{L_2}.$$

Then, by carefully and adaptively updating the smoothing parameter $\rho$ so that $\rho \approx \cos\theta$ (for simplicity, we take $\rho = \cos\theta$ below), we have $-\nabla_{\mathbf{w}}\mathcal{L}_{\cos\theta}(\mathbf{w};\sigma)\cdot\mathbf{w}^* \gtrsim \sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_2^2 - \sqrt{\mathrm{OPT}}\sin\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$, which implies that $-\nabla_{\mathbf{w}}\mathcal{L}_{\cos\theta}(\mathbf{w};\sigma)\cdot\mathbf{w}^* \gtrsim \sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_2^2$ when $\sin\theta \gtrsim \sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$; in other words, the gradient alignment holds when $\theta(\mathbf{w},\mathbf{w}^*)$ is large. Consequently, [ZWDD25] proved that the algorithm linearly converges to $\mathbf{w}^*$ until $\mathbf{w}$ is too close to $\mathbf{w}^*$ (and gradient alignment fails), in which case a $C\mathrm{OPT}+\epsilon$ error solution is found.

A critical issue arises, however, when trying to adapt the methods from [ZWDD25] to the SIM setting: the correlation between $\mathbf{w}^*$ and the Riemannian gradient $\nabla_{\mathbf{w}}\mathcal{L}_{\cos\theta}(\mathbf{w};u)$ becomes uncontrollable, even if we choose $u$ as the "best-fit" activation given $\mathbf{w}$. To see this, a simple calculation shows that

$$\begin{aligned}-\nabla_{\mathbf{w}}\mathcal{L}_{\cos\theta}(\mathbf{w};u)\cdot\mathbf{w}^* &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathrm{T}_{\cos\theta}u'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp\mathbf{w}}\cdot\mathbf{w}^*]\\ \gtrsim \sin^2\theta\,\mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathrm{T}_{\cos\theta}u'(\mathbf{w}\cdot\mathbf{x})] &- \sqrt{\mathrm{OPT}}\sin\theta\|\mathrm{T}_{\cos\theta}u'\|_{L_2}.\end{aligned}$$

[5] $\mathbf{x}^{\perp\mathbf{w}}$ is independent of $\mathbf{w}\cdot\mathbf{x}$ (as $\mathbf{x}$ is standard Gaussian) hence $\mathbf{E}[\mathrm{T}_\rho\sigma(\mathbf{w}\cdot\mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp\mathbf{w}}] = 0$.

Even though it is possible to control the $L_2^2$ distance between $u(z)$ and $\sigma(z)$ by $\theta(\mathbf{w}, \mathbf{w}^*)$ following similar steps as in [ZWDD24], it is unclear how to show that $\mathrm{T}_{\cos\theta} u'(z)$ is close to $\mathrm{T}_{\cos\theta}\sigma'(z)$ so that $-\nabla_{\mathbf{w}}\mathcal{L}_\rho(\mathbf{w}; u)\cdot\mathbf{w}^* \gtrsim \sin^2\theta \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$ and the arguments of [ZWDD25] can go through. Intuitively, the smoothing operator $\mathrm{T}_{\cos\theta}$ filters out the high-order components of the derivative of the activation $\sigma'$ in the Hermite basis while keeping only the low-order components. Thus, to ensure $\|\mathrm{T}_{\cos\theta}u' - \mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$ is small, it necessitates approximating the low degree coefficients of $\sigma'$ (under adversarial noise and without the knowledge of $\sigma$), imposing formidable technical challenges. In particular, it is unclear whether prior SIM learning frameworks [KKSK11, ZWDD24, HTY24]—alternating between the "best-fit" updates for activation $u$ and gradient-style updates for $\mathbf{w}$—can resolve this issue. Hence, our work represents a departure from this seemingly natural approach.

**Alignment with a New Vector Field** Our method deviates from prior works in that we no longer cling to the gradient field of a particular loss function, but rather, we identify a vector field that aligns with $\mathbf{w}^*$ *without the need to estimate the target function* $\sigma$ on the run. Revisiting the correlation between the gradient of the smoothed $L_2^2$ loss and the target vector $\mathbf{w}^*$: $-\nabla_{\mathbf{w}}\mathcal{L}_{\cos\theta}(\mathbf{w}; \sigma)\cdot\mathbf{w}^* = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp\mathbf{w}}\cdot\mathbf{w}^*]$, we observe that the right-hand side of the equation consists of three parts: the label $y$, a random variable $\mathbf{x}^{\perp\mathbf{w}}\cdot\mathbf{w}^*$ that is independent of $\mathbf{w}\cdot\mathbf{x}$, and a function $\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})$ of $\mathbf{w}\cdot\mathbf{x}$ that is not available when $\sigma$ is unknown. Critically, we replace the unknown function $\mathrm{T}_{\cos\theta}\sigma'(z)$ with any function $h(z)$ such that $\|h(z)\|_{L_2} = 1$, and we ask: which function $h^*(z)$ maximizes the correlation $K(h) \stackrel{\text{def}}{=} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}}h(\mathbf{w}\cdot\mathbf{x})]\cdot\mathbf{w}^*$? Let $h_0(z) \stackrel{\text{def}}{=} \mathrm{T}_{\cos\theta}\sigma'(z)/\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$. By maximality of $h^*$, we know that the correlation $K(h^*)$ is at least as large as $K(h_0) = -\nabla_{\mathbf{w}}\mathcal{L}_{\cos\theta}(\mathbf{w}; \sigma)\cdot\mathbf{w}^*/\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$, indicating

the vector field $\mathbf{H}^*_{\mathbf{w}} \stackrel{\text{def}}{=} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}}h^*(\mathbf{w}\cdot\mathbf{x})]$ is at least as good as the gradient $\nabla_{\mathbf{w}}\mathcal{L}_{\cos\theta}(\mathbf{w};\sigma)$ of the smoothed loss with respect to the *target activation* $\sigma$, after normalization.

In fact, letting $\mathbf{v}^*_{\mathbf{w}} = (\mathbf{w}^*)^{\perp\mathbf{w}}/\|(\mathbf{w}^*)^{\perp\mathbf{w}}\|_2$ and $\mathbf{w}^* = \cos\theta\mathbf{w} + \sin\theta\mathbf{v}^*_{\mathbf{w}}$, we can write $K(h)$ as:

$$K(h) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[yh(\mathbf{w}\cdot\mathbf{x})\mathbf{w}^*\cdot\mathbf{x}^{\perp\mathbf{w}}] = \sin\theta\,\mathbf{E}\left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{x}) \mid \mathbf{w}\cdot\mathbf{x}]h(\mathbf{w}\cdot\mathbf{x})\right]. \quad (1.6)$$

Consider the $L_2$ space of the standard Gaussian random variable $\mathbf{w}\cdot\mathbf{x}$ equipped with the inner product $\langle a,b\rangle = \mathbf{E}_{\mathbf{w}\cdot\mathbf{x}\sim\mathcal{N}}[a\cdot b]$; it is known from duality that $\langle a,b\rangle \le \|a\|_{L_2}\|b\|_{L_2}$ with equality if $a = b$ almost surely. Hence, the choice $h^*(\mathbf{w}\cdot\mathbf{x}) = \mathbf{E}[y\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{x} \mid \mathbf{w}\cdot\mathbf{x}]/\|\mathbf{E}[y\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{x} \mid \mathbf{w}\cdot\mathbf{x}]\|_{L_2}$ maximizes Equation (1.6). Having access to such $h^*$ guarantees that the corresponding update rule using $\mathbf{H}^*_{\mathbf{w}} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}}h^*(\mathbf{w}\cdot\mathbf{x})]$ performs at least as well as the gradient descent on the smoothed loss $\mathcal{L}_{\cos\theta}(\mathbf{w};\sigma)$. Note that by the definition of $K(h)$, we have $K(h^*) = \sin\theta\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{H}^*_{\mathbf{w}}$.

There are two obstacles in finding such an $h^*$: 1) the learner does not have knowledge of $\mathbf{w}^*$, and 2) even if the learner knew $\mathbf{w}^*$, it would not be possible to estimate $h^*(z) = \mathbf{E}[y\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{x} \mid \mathbf{w}\cdot\mathbf{x} = z]/\|\mathbf{E}[y\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{x} \mid \mathbf{w}\cdot\mathbf{x}]\|_{L_2}$ on any desired point $\mathbf{w}\cdot\mathbf{x} = z$ (as the probability of observing a single point twice is zero). For this reason, we need to consider a different way of finding an update rule.

**The Spectral Subroutine** For the first obstacle, our critical observation is that instead of estimating $h^*$, we can approximate the vector $\mathbf{H}^*_{\mathbf{w}}$ directly via spectral methods. Let us define $\mathbf{g}^*_{\mathbf{w}}(z) \stackrel{\text{def}}{=} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}} \mid \mathbf{w}\cdot\mathbf{x} = z]$ and consider the matrix

$\mathbf{M}^*_{\mathbf{w}} = \mathbf{E}_{z\sim\mathcal{N}}[\mathbf{g}^*_{\mathbf{w}}(z)\mathbf{g}^*_{\mathbf{w}}(z)^\intercal]$. Observe that

$$(\mathbf{v}^*_{\mathbf{w}})^\intercal \mathbf{M}^*_{\mathbf{w}} \mathbf{v}^*_{\mathbf{w}} = \mathbf{E}\left[\Big(\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}}\cdot\mathbf{v}^*_{\mathbf{w}} \mid \mathbf{w}\cdot\mathbf{x}]\Big)^2\right] = (\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{H}^*_{\mathbf{w}})\|\mathbf{E}[y\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{x} \mid \mathbf{w}\cdot\mathbf{x}]\|_{L_2}$$
$$= \frac{K(h^*)}{\sin\theta}\|\mathbf{E}[y\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{x} \mid \mathbf{w}\cdot\mathbf{x}]\|_{L_2} \geq \frac{K(\mathrm{T}_{\cos\theta}\sigma')}{\sin\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}}\|\mathbf{E}[y\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{x} \mid \mathbf{w}\cdot\mathbf{x}]\|_{L_2},$$

where in the last inequality we used the maximality of $K(h^*)$. The first equality above also implies $\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{M}^*_{\mathbf{w}}\mathbf{v}^*_{\mathbf{w}} = \|\mathbf{E}[y\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{x} \mid \mathbf{w}\cdot\mathbf{x}]\|^2_{L_2}$; therefore, we have $(\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{M}^*_{\mathbf{w}}\mathbf{v}^*_{\mathbf{w}})^{1/2} = \mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{H}^*_{\mathbf{w}}$. Since we know that $\mathbf{H}^*_{\mathbf{w}}$ is at least as aligned with $\mathbf{w}^*$ as the gradient vector $-\nabla_{\mathbf{w}}\mathcal{L}_{\cos\theta}(\mathbf{w};\sigma)$ and we know from [ZWDD25] that $K(\mathrm{T}_{\cos\theta}\sigma') = -\nabla_{\mathbf{w}}\mathcal{L}_\rho(\mathbf{w};\sigma)\cdot\mathbf{w}^* \gtrsim \sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|^2_{L_2} \gg 0$ since the gradient alignment condition holds, we get that both $\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{M}^*_{\mathbf{w}}\mathbf{v}^*_{\mathbf{w}}$ and $\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{H}^*_{\mathbf{w}}$ are far away from 0. This implies that much of the information on $\mathbf{H}^*_{\mathbf{w}}$ is contained in $\mathbf{v}^*_{\mathbf{w}}$ and, in addition, $\mathbf{v}^*_{\mathbf{w}}$ is contained in the eigenspace of the large eigenvectors of $\mathbf{M}^*_{\mathbf{w}}$. Furthermore, one can show that for any direction $\mathbf{u}$ that is orthogonal to $\mathbf{v}^*_{\mathbf{w}}$, both $\mathbf{u}\cdot\mathbf{M}^*_{\mathbf{w}}\mathbf{u} \lesssim \mathrm{OPT}$ and $\mathbf{H}^*_{\mathbf{w}}\cdot\mathbf{u} \lesssim \sqrt{\mathrm{OPT}}$ are small. In other words, $\mathbf{H}^*_{\mathbf{w}}$ is almost completely captured by $\mathbf{v}^*_{\mathbf{w}}$, and $\mathbf{v}^*_{\mathbf{w}}$ is effectively contained in the space of the highest eigenvalues. Consequently, $\mathbf{H}^*_{\mathbf{w}}$ is approximated by the top eigenvectors of $\mathbf{M}^*_{\mathbf{w}}$.

**Approximation and Regularity of Monotone Functions** The second obstacle is to estimate the function $\mathbf{g}^*_{\mathbf{w}}(z) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}} \mid \mathbf{w}\cdot\mathbf{x} = z]$ that constructs the matrix $\mathbf{M}^*_{\mathbf{w}}$. This is not a trivial task even in the noiseless setting because we are conditioning on a hyperplane that has measure zero in the space. One would hope that conditioning on small bands suffices for this purpose; namely, that if $\mathbf{w}\cdot\mathbf{x} \in (a,b)$ with $|b-a| = \mathrm{poly}(\epsilon)$, then for all $z \in (a,b)$ it would be $\mathbf{E}[y\mathbf{x}^{\perp\mathbf{w}} \mid \mathbf{w}\cdot\mathbf{x} = z] \approx \mathbf{E}[y\mathbf{x}^{\perp\mathbf{w}} \mid \mathbf{w}\cdot\mathbf{x} \in (a,b)]$. Unfortunately, since the labels $y$ are adversarially corrupted, the adversary could

corrupt $y$ for each value of $\mathbf{w} \cdot \mathbf{x} = z$ —in which case such an approximation would not yield an accurate estimate. Instead, we proceed as follows: consider restricting $h(z)$ to be a piecewise-constant function on a fixed set of small bands $\mathcal{E}_i = [a_i, a_{i+1}), i \in [I]$. We show that the argument about maximizing $K(h)$ on continuous functions $h$ can be carried out similarly to piecewise constants. Let $\widetilde{h}^*$ be the piecewise constant function that maximizes $K(h)$. We further show that $\mathrm{T}_{\cos\theta}\sigma'$ can be approximated by a fixed value on each band $\mathcal{E}_i$. Hence, using a piecewise-constant approximate function $\widetilde{\mathrm{T}}_{\cos\theta}\sigma'$, we have that $0 \ll K(\mathrm{T}_{\cos\theta}\sigma') \approx K(\widetilde{\mathrm{T}}_{\cos\theta}\sigma') \leq K(\widetilde{h}^*)$, and thus the argument that large eigenvectors of $\mathbf{M}^*_{\mathbf{w}} = \mathbf{E}_{z\sim\mathcal{N}}[\mathbf{g}^*_{\mathbf{w}}(z)\mathbf{g}^*_{\mathbf{w}}(z)^\intercal]$ contain information about $\mathbf{H}^*_{\mathbf{w}}$ can be extended to $\mathbf{M}_{\mathbf{w}} = \mathbf{E}_{z\sim\mathcal{N}}[\mathbf{g}_{\mathbf{w}}(z)\mathbf{g}_{\mathbf{w}}(z)^\intercal]$, where $\mathbf{g}_{\mathbf{w}}(z)$ is the piecewise constant version of $\mathbf{g}^*_{\mathbf{w}}(z)$, which can now be efficiently estimated.

**Optimization via Random Walk** A final obstacle in this approach is that both $\mathbf{u}$ and $-\mathbf{u}$ are eigenvectors that correspond to the maximum eigenvalue and we cannot determine whether $\mathbf{u}$ or $-\mathbf{u}$ correlates positively with $\mathbf{w}^*$. To address this issue, at each iteration, we pick the direction from $\{\mathbf{u}, -\mathbf{u}\}$ at random. Consequently, with probability 1/2, the algorithm will decrease the angle with $\mathbf{w}^*$. Consider the random variable $Z_t = \theta(\mathbf{w}^{(t)}, \mathbf{w}^*)$—the angle between $\mathbf{w}_t$ and $\mathbf{w}^*$. Let $\theta^*$ be the largest angle such that if $Z_t \leq \theta^*$ then $\mathcal{L}_2(\mathbf{w}^{(t)}; \sigma) \leq O(\mathrm{OPT}) + \epsilon$. Furthermore, let $\theta_0 \stackrel{\mathrm{def}}{=} \theta(\mathbf{w}^{(0)}, \mathbf{w}^*)$. Assume without loss of generality that the initialized vector $\mathbf{w}^{(0)}$ is not a constant approximate vector, hence $\theta_0 \geq \theta^*$. Now let $\tau_1 = \inf_t\{t \geq 1 \mid Z_t \leq \theta^*\}$, i.e., $\tau_1$ is the first iteration that has $Z_{\tau_1} \leq \theta^*$, and let $\tau_2 = \inf_t\{t \geq 1 \mid Z_t \geq \theta_0\}$. If $\mathbf{Pr}[\tau_1 < \tau_2] \geq \alpha > 0$, then repeating the process $O(1/\alpha)$ times guarantees that with high probability the event $\tau_1 < \tau_2$ happens. Note that: $\mathbf{Pr}[\tau_1 < \tau_2] \geq \mathbf{Pr}[\text{chooses the correct direction for all the } T \text{ steps until } Z_t \leq \theta^*] = 2^T$. We will show

that $T \lesssim \log(BL/\epsilon)$. Thus, we have $\mathbf{Pr}[\tau_1 < \tau_2] \geq 1/2^T = \mathrm{poly}(\epsilon, 1/B, 1/L)$. Therefore, repeating the algorithm $2^T = \mathrm{poly}(1/\epsilon, B, L)$ times suffices.

## Part III: Robustly Learning Multi-Index Models

A common (semi)-parametric modeling assumption in high- dimensional statistics is that the target function depends only on a few relevant directions. Specifically, multi-index models [FJS81, Hub85, Li91, HL93, XTLZ02, Xia08] prescribe that the target function is of the form $f(\mathbf{x}) = g(\mathbf{W}\mathbf{x})$ for a link function $g : \mathbb{R}^k \mapsto \mathcal{Y}$ and a $k \times d$ weight matrix $\mathbf{W}$. Hence, a function is called a MIM if it depends only on the projection onto a "low-dimensional" subspace. Formally, we have:

**Definition 1.3.30** (Multi-Index Model)**.** *A function $f : \mathbb{R}^d \to \mathcal{Y}$ is called a $k$-MIM if there exists a subspace $U \subseteq \mathbb{R}^d$ of dimension at most $k$ such that $f$ depends only on the projection onto $U$, i.e., $f(\mathbf{x}) = f(\mathrm{proj}_U \mathbf{x})$, where $\mathrm{proj}_U \mathbf{x}$ is the projection of $\mathbf{x}$ on $U$. Equivalently, $f$ is of the form $f(\mathbf{x}) = g(\mathbf{W}\mathbf{x})$ for a link function $g : \mathbb{R}^k \to \mathcal{Y}$ and a $k \times d$ weight matrix $\mathbf{W}$.*

In typical applications, the dimension $k$ of the hidden subspace is usually significantly smaller than the ambient dimension $d$.

Multi-index models capture a wide range of parametric models studied in the statistics and computer science literatures, including neural networks and classes of geometric Boolean functions (e.g., intersections of halfspaces). An extensive recent line of work [JSA15a, GLM18, DH18, BJW19, GKLW19, DKKZ20, CM20, DLS22, BBSS22, CDG+23, CN23, DK23b] have studied the efficient learnability of (natural classes of) MIMs from random examples under well-behaved marginal distributions — most notably under the Gaussian distribution on examples. The aforementioned works exclusively

focus on the PAC model with random samples and the underlying algorithms succeed in the realizable setting (or in the presence of additive Gaussian label noise).

Despite extensive prior investigation, there is still a vast gap in our understanding of the complexity of learning MIMs, in particular in the presence of noisy data. This discussion motivates the general direction addressed by this work:

*What is the complexity of learning general classes of MIMs in the presence of noisy data?*

### 1.3.8 Agnostically Learning Multi-Index Models with Queries

**PAC Learning with Queries** In Valiant's PAC learning model [Val84], the learner is given access to random examples labeled according to an unknown function in a known concept class. The goal of the learner is to compute a hypothesis that is close to the target function with respect to a specified loss function[6]. The standard PAC learning model is "passive" in that the learning algorithm has no control over the selection of the training set. Interestingly, while this has become known as *the* PAC model, Valiant's landmark paper [Val84] allowed queries (in addition to random samples), i.e., black-box access to the target function. We will refer to this as PAC+Query model.

A *query oracle*[7] allows the learner to obtain the value of the target function on any desired point in the domain. PAC learning with access to a query oracle can be viewed as an "active" learning model, intuitively capturing the ability to perform experiments or the availability of expert advice. A long line of research in computational learning

[6]For Boolean functions, one typically uses the 0-1 loss, while for real-valued functions a typical choice is the $L_2$ loss.

[7]In the special case of learning Boolean-valued functions, these are known as "membership" queries, as the answer to a query determines membership in the set of satisfying assignments of the target concept.

theory has explored the power of queries in the context of PAC learning. This line of investigation has spanned the distribution-free versus distribution-specific settings and the realizable (i.e., clean label) setting versus the agnostic (i.e., adversarial label noise) setting; see, e.g., [Ang87, GL89, KM93, Jac97] for some classical early works and [GKK08a, BLQT22] for some more recent results in this broad area. A conceptual message of this line of work is that, in the realizable setting, access to queries can be stronger than random samples (from a computational standpoint) for a range of natural concept classes.

In addition to being a fundamental open question in learning theory, the general problem of understanding the effect of query access in the *computational complexity* of learning has received renewed attention over the past decade in the context of deep neural networks. A recent line of inquiry from the machine learning security community has studied *model extraction attacks* — see, e.g., [TZJ+16, SSG17, PMG+17, MSDH19, JCB+20, RK20, JWZ20] and references therein — where black-box query access to publicly deployed networks may allow efficient reconstruction of the hidden model – thus exposing potential vulnerability of the deployed models. These practical applications served as a motivation for the design of the first computationally efficient learners for simple neural networks using query access to the target function [CKM21, DG22]. Importantly, the latter algorithmic results apply in the realizable PAC model under the Gaussian distribution.

**Agnostically Learning Multi-index Models with Queries** Here we study the power of queries in the *agnostic* PAC model [Hau92, KSS94] for a wide class of multi-index models. In the agnostic model, no assumptions are made on the labels of the examples and the goal is to compute a hypothesis that is competitive with the *best-fit*

function in a known class. This is a notoriously challenging model of learning with very few positive results in the distribution-free setting. For example, it is known that even *weak* (distribution-free) agnostic learning (i.e., outputting a hypothesis with non-trivial advantage over random) is computationally hard for very simple classes of single-index models with known link functions. These include linear threshold gates and single neurons with ReLU activations [Dan16, DKMR22b, Tie23][8].

In this work, we focus on the general problem of agnostically learning multi-index models under the standard Gaussian distribution using queries. At a high-level, our results also encompass the challenging setting where the link function is *unknown* and only require an average smoothness condition on the target function. Classes covered by our framework include real-valued function classes such as constant-depth neural networks with ReLU activations and Boolean concept classes such as intersections of halfspaces. In summary, we are interested in the following question:

**Question 1.3.31.** *Does* query access *affect the complexity of distribution-specific agnostic learning of multi-index models? In particular, does the availability of queries allow for* qualitatively *more efficient algorithms, compared to the vanilla random example setting?*

The main contribution of this work is a simple and general methodology that answers this question in the affirmative for a broad family of multi-index function models (including all the aforementioned examples).

A special case of Question 1.3.31 was explicitly asked — in the Boolean setting — for the class of Linear Threshold Functions by Feldman [Fel08] and by Gopalan, Kalai,

[8]We note that these computational hardness results hold even with query access, as follows from [Fel08].

and Klivans [GKK08b] As a corollary of our approach, we answer this open question. Specifically, we provide a new query algorithm for agnostically learning halfspaces implying a super-polynomial separation between the two learning models (learning with random samples versus with queries), subject to standard cryptographic assumptions. In the following subsection, we describe our contributions in detail.

## Our Results

**Problem Definition** Before we formally state our main results, we define the agnostic learning model with queries. For concreteness, Definition 1.3.32 concerns real-valued functions, where the accuracy is measured with respect to the $L_2$ loss. The definition for Boolean-valued concepts is essentially identical, where the $L_2$ loss is replaced by the 0-1 loss.

**Definition 1.3.32** (Agnostically Learning Real-valued Functions with Queries)**.** *Fix $\epsilon \in (0,1)$ and a class $\mathcal{C}$ of real-valued functions on $\mathbb{R}^d$. The adversary picks a label function $y(\mathbf{x}) \in \mathbb{R}$ for every $\mathbf{x} \in \mathbb{R}^d$. The learner is allowed to either draw $\mathbf{x} \sim \mathcal{N}$ (sample access) or select any desired point $\mathbf{x} \in \mathbb{R}^d$ (query access) and obtain the value $y(\mathbf{x})$. Let $N_s \in \mathbb{Z}_+$ be the number of samples and $N_q \in \mathbb{Z}_+$ the number of queries used by the learner. The goal of the learner is to output a hypothesis $h : \mathbb{R}^d \to \mathbb{R}$ that, with high probability, has excess $L_2^2$ error at most $\epsilon$ (with respect to $\mathcal{C}$), i.e., it satisfies $\mathcal{E}_2(h, \mathcal{C}; y) \stackrel{\text{def}}{=} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(h(\mathbf{x}) - y(\mathbf{x}))^2] - \inf_{c\in\mathcal{C}} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(c(\mathbf{x}) - y(\mathbf{x}))^2] \leq \epsilon$ .*

**Remark 1.3.33** (Boolean-valued Functions)**.** In the boolean-valued setting, we focus on learning with respect to the 0-1 loss. That is, the goal of the learner is to output a hypothesis $h : \mathbb{R}^d \mapsto \{\pm 1\}$ with excess 0-1 error at most $\epsilon$, i.e., $\mathcal{E}_{0/1}(h, \mathcal{C}; y) \stackrel{\text{def}}{=} \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[h(\mathbf{x}) \neq y(\mathbf{x})] - \inf_{c\in\mathcal{C}} \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[c(\mathbf{x}) \neq y(\mathbf{x})] \leq \epsilon$ .

### Agnostically Learning Real-valued Multi-index Models

We start by describing the family of multi-index models for which our results are applicable. Roughly speaking, our algorithmic approach can be used to agnostically learn any family of multi-index models $\mathcal{C}$ such that any function in $\mathcal{C}$ has "bounded variation", in the sense that the $L_2$-norm of its gradient is bounded with respect to the standard normal. We remark that similar "smoothness" assumptions, i.e., that $f$ belongs in a Sobolev space, are standard (and necessary) in non-parametric and semi-parametric regression [Tsy08]. Under this assumption, we show that there exists an *efficient dimension-reduction* scheme that yields a "fixed parameter tractable" agnostic learner significantly improving over the best known algorithmic results in the agnostic PAC setting with random examples.

We are now ready to formally define the semi-parametric class of MIMs that we consider in this work. In the following definition, we require that the target function is bounded in $L_4$-norm (with respect to the standard normal distribution) and also that the norm of its gradient is bounded in $L_2$-norm.

**Definition 1.3.34** (Bounded Variation Multi-index Models)**.** *Fix $L, M > 0$ and $k \in \mathbb{Z}_+$. We define the class $\mathfrak{R}(M, L, k)$ of continuous, (almost everywhere) differentiable real-valued functions such that for every $f \in \mathfrak{R}(M, L, k)$:*

1. *It holds $(\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f^4(\mathbf{x})])^{1/2} \leq M$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2] \leq L$.*

2. *There exists a subspace $U$ of $\mathbb{R}^d$ of dimension at most $k$ such that $f$ depends only on $U$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$ it holds that $f(\mathbf{x}) = f(\mathrm{proj}_U \mathbf{x})$, where $\mathrm{proj}_U \mathbf{x}$ is the projection of $\mathbf{x}$ on $U$.*

We will subsequently see that this is a very broad class of functions subsuming

commonly studied classes such as multi-layer neural networks with ReLUs and other activations.

Our main result is an efficient algorithm that exploits the power of queries to significantly reduce the runtime of agnostically learning the semi-parametric class of Definition 1.3.34.

**Theorem 1.3.35** (Agnostic Query Learner for Real-valued Multi-index Models)**.** *Fix the function class* $\mathfrak{R}(M, L, k)$ *given in Definition 1.3.34. There exists an algorithm that makes* $N_q = \mathrm{poly}(dML/\epsilon)$ *queries, draws* $N_s = \mathrm{poly}(dML/\epsilon) + k^{\mathrm{poly}(L,M,1/\epsilon)}$ *random labeled examples, runs in time* $\mathrm{poly}(N_s, N_q, d)$*, and outputs a polynomial* $h : \mathbb{R}^d \mapsto \mathbb{R}$ *such that with high probability* $h$ *has* $L_2^2$*-excess error* $\mathcal{E}_2(h, \mathfrak{R}(L, M, k); y) \leq \epsilon$*.*

**Comparison with Sample-Based Algorithms** As a corollary of Theorem 1.3.35, we establish a strong separation between the agnostic PAC+Query model and the agnostic PAC model (with random samples only). We first compare with the best-known algorithm for agnostically PAC learning real-valued functions, which is the $L_2$-polynomial regression algorithm. To agnostically learn the class of Definition 1.3.34 to excess error $\epsilon$, one needs polynomials of degree $\mathrm{poly}(L, M, 1/\epsilon)$, and thus $d^{\mathrm{poly}(L,M,1/\epsilon)}$ samples and time are necessary. Theorem 1.3.35 leverages the power of queries to efficiently reduce the dimensionality of the problem, and thus qualitatively improve the computational complexity of agnostic learning to $\mathrm{poly}(d)\, k^{\mathrm{poly}(L,M,1/\epsilon)}$.

Given the assumption of Definition 1.3.34 that the target function depends only on an unknown $k$-dimensional subspace, it is natural to attempt some kind of dimension-reduction technique in order to reduce the sample and computational complexity of learning. Such reductions are indeed often possible *in the realizable setting* by using

some form of PCA and then working in the obtained low-dimensional subspace; see, e.g., [Vem10b].

On the other hand, in the agnostic setting considered here, there is strong evidence that such dimension-reduction schemes, or any other runtime improvements whatsoever, are impossible using only sample access to the target function. Specifically, a recent line of work (see, e.g., [DKPZ21, DKR23]) has shown that for agnostically learning real-valued MIMS (even very special cases thereof), the standard $L_2$-regression algorithm is qualitatively optimal computationally (e.g., under standard cryptographic assumptions) in the standard agnostic PAC model. This in particular implies that the best possible runtime without query access is $d^{\mathrm{poly}(1/\epsilon)}$. In fact, even for learning a single ReLU activation, which satisfies Definition 1.3.34 with $L, M = O(1)$ and $k = 1$, $d^{\mathrm{poly}(1/\epsilon)}$ samples and time are required [DKPZ21, DKR23]. In contrast, Theorem 1.3.35 decouples the dimension dependence from the dependence on $1/\epsilon$ and yields an algorithm with runtime $\mathrm{poly}(d)\, 2^{\mathrm{poly}(1/\epsilon)}$.

**Concrete Applications** Theorem 1.3.35 applies to a fairly general non-parametric class of functions. Here we provide specific applications to well-studied classes of neural networks.

***Single Non-Linear Gates.*** The simplest case is that of agnostically learning a ReLU, i.e., a function of the form $f(\mathbf{x}) = \mathrm{ReLU}(\mathbf{w} \cdot \mathbf{x})$, where $\mathbf{w} \in \mathbb{R}^d$ and $\mathrm{ReLU}(t) = \max\{0, t\}$. In the vanilla agnostic PAC setting, the complexity of this problem is $d^{\mathrm{poly}(1/\epsilon)}$(both upper and lower bounds). On the positive side, the $L_2$-polynomial regression algorithm has sample and computational complexity $d^{\Theta(\mathrm{poly}(1/\epsilon))}$. On the negative side, there is strong evidence that this complexity upper bound is qualitatively best possible, both for SQ algorithms [GGK20, DKZ20, DKPZ21] and

under plausible cryptographic assumptions [DKR23]. Our agnostic query learner has complexity $\mathrm{poly}(d)\, 2^{\mathrm{poly}(1/\epsilon)}$, implying a super-polynomial separation between the two learning models.

**Corollary 1.3.36** (Agnostic Query Learning for ReLUs)**.** *There exists an agnostic query learner for the class of ReLUs on* $\mathbb{R}^d$ *with running time* $\mathrm{poly}(d)\, 2^{\mathrm{poly}(1/\epsilon)}$.

Corollary 1.3.36 follows from Theorem 1.3.35 by observing that ReLUs satisfy Definition 1.3.34 for $k = 1$ and $L, M = O(1)$ (assuming that the norm of the weight vector is bounded, i.e., $\|\mathbf{w}\|_2 = O(1)$).

Note that selecting the excess error to be $\epsilon = 1/\log^c(d)$, where $c > 0$ is a small constant, the query algorithm of Corollary 1.3.36 has $\mathrm{poly}(d)$ runtime. On the other hand, the complexity of agnostic learning problem with random samples is super-polynomial in $d$ for *any* $\epsilon = o_d(1)$.

Finally, we note that Corollary 1.3.36 holds for other link functions satisfying smoothness assumptions, e.g., sigmoidal activations of the form $t \mapsto 1/(1 + \exp(-t))$.

***Single-index Models.*** Our first application above assumed that the link function is known a priori.We next consider learning Single-index models (SIMs) with an *unknown* Lipschitz link function $g : \mathbb{R} \mapsto \mathbb{R}$, i.e., $f(\mathbf{x}) = g(\mathbf{w}\cdot\mathbf{x})$. Classical results [KS09, KKSK11] gave efficient algorithms for this setting in the realizable PAC setting (or with unbiased additive noise) under the additional assumption that $g$ is non-decreasing. The agnostic setting was recently considered in [GGKS23] who gave an efficient algorithm achieving error $O(\sqrt{\mathrm{OPT}}) + \epsilon$ for distributions with bounded second moments (similarly assuming weight vectors of bounded $\ell_2$-norm). Using Theorem 1.3.35, we can leverage query access to provide optimal agnostic guarantees with essentially the same complexity as for the case of known link function.

**Corollary 1.3.37** (Agnostic Query Learning for Lipschitz SIMs)**.** *There exists an agnostic query learner for the class of L-Lipschitz SIMs on* $\mathbb{R}^d$*, for* $L = O(1)$*, with running time* $\mathrm{poly}(d)\, 2^{\mathrm{poly}(1/\epsilon)}$.

***One-Hidden Layer ReLU Networks.*** Our approach naturally extends to non-negative linear combinations (aka sums) of ReLUs, i.e., functions of the form $f(\mathbf{x}) = \sum_{i=1}^{k} \alpha^{(i)} \mathrm{ReLU}(\mathbf{w}^{(i)} \cdot \mathbf{x})$ for $k$ non-negative weights $\alpha^{(i)} \geq 0$ and weight vectors $\mathbf{w}^{(i)} \in \mathbb{R}^d$. Prior work [JSA15a, GLM18, DKKZ20, DK20b] has studied this problem in the noiseless setting with random samples under the Gaussian distribution — with the best-known runtime being $\mathrm{poly}(d/\epsilon)\,(k/\epsilon)^{O(\log^2 k)}$ [DK20b]. Using Theorem 9.3.2, we obtain an agnostic query learner with complexity $\mathrm{poly}(d)O(k)^{\mathrm{poly}(1/\epsilon)}$. To see this, we note that as long as $\mathbf{E}[f^2(\mathbf{x})] = O(1)$ we also obtain that $\mathbf{E}[\|\nabla f(\mathbf{x})\|_2^2] = O(1)$ which implies only an $O(k)^{\mathrm{poly}(1/\epsilon)}$ runtime overhead.

Our approach can also be applied to the more general class of (unconstrained) linear combinations of $k$ ReLUs, i.e., functions of the form $f(\mathbf{x}) = \sum_{i=1}^{k} \alpha^{(i)} \mathrm{ReLU}(\mathbf{w}^{(i)} \cdot \mathbf{x})$. This is known [DKKZ20, CDG+23, CN23, DK23b] to be a more challenging class of functions to learn. In the noiseless setting, the best known runtime for general linear combinations is $(dk/\epsilon)^{O(k)}$ [DK23b]. Using Theorem 1.3.35, we obtain an agnostic query learner with complexity $\mathrm{poly}(d)\, 2^{\mathrm{poly}(k/\epsilon)}$.

**Corollary 1.3.38** (Agnostic Query Learning for 1-Hidden Layer ReLU Networks)**.** *There exists an agnostic query learner for sums of k ReLUs on* $\mathbb{R}^d$ *with running time* $\mathrm{poly}(d)\, O(k)^{\mathrm{poly}(1/\epsilon)}$*. For general linear combinations of ReLUs, the runtime is* $\mathrm{poly}(d)\, 2^{\mathrm{poly}(k/\epsilon)}$.

***Bounded Depth Neural Networks.*** Our non-parametric function class of Definition 1.3.34 includes deep ReLU networks with $\ell$ layers of width at most $S$. More

Table 1.1: Learning Real-Valued Functions using Queries: Running time comparisons of the best known PAC algorithms with our PAC+Queries technique (Influence PCA).

| Function Class | PAC (without queries) $L_2$ Regression | PAC+Queries **Influence PCA** |
|---|---|---|
| Single ReLU | $d^{\mathrm{poly}(1/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(1/\epsilon)}$ |
| Sum of $k$ ReLUs | $d^{\mathrm{poly}(1/\epsilon)}$ | $\mathrm{poly}(d)\, O(k)^{\mathrm{poly}(1/\epsilon)}$ |
| Linear Combinations of $k$ ReLUs | $d^{\mathrm{poly}(k/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(k/\epsilon)}$ |
| Deep Networks with $\ell$-Layers, $S$-width | $d^{\mathrm{poly}(\ell S/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(\ell S/\epsilon)}$ |
| Bounded Variation | $d^{\mathrm{poly}(k,L,M,1/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(k,L,M,1/\epsilon)}$ |

precisely, we assume that $f(\mathbf{x}) = \mathbf{W}_L \mathrm{ReLU}(\mathbf{W}_{L-1} \cdots \mathrm{ReLU}(\mathbf{W}_1 \mathbf{x}))$, for matrices $\mathbf{W}_1 \in \mathbb{R}^{k_1 \times d}, \ldots, \mathbf{W}_L \in \mathbb{R}^{k_L \times 1}$, with $\|\mathbf{W}_i\|_{op} \leq O(1)$ and $k_i \leq S$; see Definition 9.3.24 for more details. The running time of our algorithm for this class is $\mathrm{poly}(d) 2^{\mathrm{poly}(\ell S/\epsilon)}$; see Theorem 9.3.25. We remark that a similar fixed-parameter tractability result for deep ReLU networks was recently shown in [CKM22] for the realizable PAC setting (with access to random examples only). Our result exploits the power of queries to provide a learner with qualitatively similar running time in the much more challenging agnostic setting. We remark that the following result can be readily extended to other continuous activation functions, including sigmoids, LeakyReLUs, and combinations thereof.

**Corollary 1.3.39** (Agnostic Query Learning for Bounded-Depth Networks)**.** *There exists an agnostic query learner for $\ell$-depth, $S$-width, ReLU networks on $\mathbb{R}^d$ with running time* $\mathrm{poly}(d) 2^{\mathrm{poly}(\ell S/\epsilon)}$.

For a summary of our results for the above classes, we refer to Table 1.1 (where for the $L_2$-regression algorithm we only assume random sample access).

**Proper versus Improper Learning** The hypothesis computed by algorithm of Theorem 1.3.35 is not necessarily in the target concept class. That is, the agnostic learner is *improper*. With some additional effort, our approach can be used to obtain *proper* learners. As a concrete example, for the class of ReLUs, we show the following:

**Theorem 1.3.40** (Proper Agnostic Query Learner of ReLUs)**.** *There exists an algorithm that makes* $\mathrm{poly}(d/\epsilon)$ *queries, runs in time* $\mathrm{poly}(d)\, 2^{\mathrm{poly}(1/\epsilon)}$*, and properly agnostically learns the class of ReLUs on* $\mathbb{R}^d$*, i.e., it outputs a ReLU hypothesis* $h(\mathbf{x}) = \mathrm{ReLU}(\widehat{\mathbf{w}} \cdot \mathbf{x})$ *with excess* $L_2^2$ *error at most* $\epsilon$ *with high probability.*

We note that in addition to computing a ReLU hypothesis, the learner of Theorem 1.3.40 uses $\mathrm{poly}(d/\epsilon)$ labeled examples (queries plus random examples), removing the extraneous $2^{\mathrm{poly}(1/\epsilon)}$ term in our generic result.

It is natural to ask whether the $2^{\mathrm{poly}(1/\epsilon)}$ runtime dependence in Theorem 1.3.40 is inherent. We provide evidence that such a dependence may be necessary for *proper* learners. Specifically, we show (Theorem 9.5.4) that if there exists a $\mathrm{poly}(d/\epsilon)$ agnostic proper learning for our problem, there exists a polynomial-time algorithm for the small-set expansion (SSE) problem [RS10] (refuting the SSE hypothesis). This hardness result also extends to the Boolean class of halfspaces. Obtaining a computational lower bound for improper learners is left as an interesting open problem.

### Agnostically Learning Boolean Multi-index Models

We start by describing the family of Boolean functions for which our results are applicable. Roughly speaking, our algorithmic approach can be used to agnostically learn any Boolean concept class $\mathcal{C}$ satisfying the following conditions: (i) $\mathcal{C}$ has bounded Gaussian surface area, (ii) it depends on an unknown low-dimensional subspace, and

(iii) it is closed under translations. Under these assumptions, we similarly obtain a "fixed parameter tractable" agnostic learner qualitatively improving over the agnostic PAC setting with random examples only.

The Gaussian surface area of a Boolean function is the surface area of its decision boundary weighted by the Gaussian density (Definition 1.3.41). The Gaussian surface area of a concept class has played a significant role as a useful complexity measure in learning theory and related fields; see, e.g., [KOS08, Kan11, Nee14, KTZ19, DMN21]. A formal definition follows:

**Definition 1.3.41** (Gaussian Surface Area)**.** *For a Borel set $A \subseteq \mathbb{R}^d$, its Gaussian surface area is defined by $\Gamma(A) \stackrel{\text{def}}{=} \liminf_{\delta \to 0} \frac{\mathcal{N}(A_\delta \setminus A)}{\delta}$, where $A_\delta = \{x : \mathrm{dist}(x, A) \leq \delta\}$. For a Boolean function $f : \mathbb{R}^d \mapsto \{\pm 1\}$, we overload notation and define its Gaussian surface area to be the surface area of its positive region $K = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = +1\}$, i.e., $\Gamma(f) = \Gamma(K)$. For a class of Boolean concepts $\mathcal{C}$, we define $\Gamma(\mathcal{C}) \stackrel{\text{def}}{=} \sup_{f \in \mathcal{C}} \Gamma(f)$.*

We are ready to define the class of Boolean multi-index models for which our approach applies.

**Definition 1.3.42** (Bounded Surface Area, Low-Dimensional Boolean Concepts)**.** *Fix $\Gamma > 0$ and $k \in \mathbb{Z}_+$. We define the class $\mathfrak{B}(\Gamma, k)$ of Boolean concepts with the following properties:*

1. *For every $f \in \mathfrak{B}(\Gamma, k)$, it holds $\Gamma(f_{\mathbf{r}}) \leq \Gamma$ for all $\mathbf{r} \in \mathbb{R}^d$, where $f_{\mathbf{r}}(\mathbf{x}) = f(\mathbf{x} + \mathbf{r})$.*
2. *For every $f \in \mathfrak{B}(\Gamma, k)$, there exists a subspace $U$ of $\mathbb{R}^d$ of dimension at most $k$ such that $f$ depends only on $U$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$ it holds $f(\mathbf{x}) = f(\mathrm{proj}_U \mathbf{x})$.*

We remark that $\mathfrak{B}(\Gamma, k)$ is a general *non-parametric class* that contains a range of

natural and well-studied Boolean function classes. For example, $\mathfrak{B}(\Omega(k), k)$ contains arbitrary functions of $k$ halfspaces.

Our main positive result in this context is a query algorithm that agnostically learns the class $\mathfrak{B}(\Gamma, k)$ with running time $\mathrm{poly}(d)k^{\mathrm{poly}(\Gamma/\epsilon)}$. In more detail, we establish the following theorem:

**Theorem 1.3.43** (Agnostic Learner for Boolean Multi-index Models)**.** *Fix the concept class $\mathfrak{B}(\Gamma, k)$ given in Definition 1.3.42. There exists an algorithm that makes $N_q = \mathrm{poly}(d/\epsilon)$ queries, draws $N_s = \mathrm{poly}(d/\epsilon) + O(k)^{\mathrm{poly}(\Gamma/\epsilon)}$ random labeled examples, runs in sample-polynomial time, and outputs a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ with excess 0-1 error $\mathcal{E}_{0/1}(h, \mathfrak{B}(\Gamma, k); y) \leq \epsilon$.*

**Discussion** Some remarks are in order. We start by noting that, in the setting of Theorem 1.3.43, an exponential dependence on the parameter $\Gamma$ is *information-theoretically* necessary — even with access to queries. Specifically, as shown in [KOS08], there exists a Boolean concept class with Gaussian surface area $\Gamma$ (consisting of intersections of halfspaces) such that the total number of samples and queries required to obtain constant accuracy is $2^{\Omega(\Gamma)}$.

It is worth comparing Theorem 1.3.43 with the best known algorithmic results in the standard agnostic PAC model (with random samples only). Klivans, O'Donnell and Servedio [KOS08] showed that the $L_1$-polynomial regression algorithm of [KKMS08] agnostically learns any concept class on $\mathbb{R}^d$ whose Gaussian surface area is at most $\Gamma > 0$ with (sample and computational) complexity $d^{\mathrm{poly}(\Gamma/\epsilon)}$. Under the additional assumption that the concepts in the target class depend on an unknown $k$-dimensional subspace, for some parameter $k \ll d$, Theorem 1.3.43 gives a significantly improved

agnostic query algorithm with computational complexity $\mathrm{poly}(d)\, k^{\mathrm{poly}(\Gamma/\epsilon)}$.

For a concrete example, if the target class is the concept class consisting of any intersection of $\ell$ halfspaces, then we have that $k = \ell$ and $\Gamma = O(\sqrt{\log(\ell)})$ [KOS08]. So, as long as $\ell = O(1)$ or even $\ell = \mathrm{polylog}(d)$, query access allows us to obtain a super-polynomial complexity improvement.

**Concrete Applications** Theorem 1.3.43 applies to a fairly general non-parametric class of functions. Here we provide specific applications to well-studied classes of Boolean functions.

***Halfspaces.*** Arguably the simplest application is for the class of halfspaces. A halfspace (or Linear Threshold Function) is any Boolean-valued function $f : \mathbb{R}^d \to \{\pm 1\}$ of the form $f(\mathbf{x}) = \mathrm{sign}\,(\mathbf{w} \cdot \mathbf{x} - \theta)$, where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector and $\theta \in \mathbb{R}$ is the threshold. (The function $\mathrm{sign} : \mathbb{R} \to \{\pm 1\}$ is defined as $\mathrm{sign}(t) = 1$ if $t \geq 0$ and $\mathrm{sign}(t) = -1$ otherwise.) The problem of PAC learning halfspaces is a textbook problem in machine learning, whose history goes back to Rosenblatt's Perceptron algorithm [Ros58]. As a corollary of Theorem 1.3.43, we obtain the following:

**Corollary 1.3.44** (Agnostic Query Learning of Halfspaces)**.** *There exists an agnostic query learner for the class of halfspaces on* $\mathbb{R}^d$ *with running time* $\mathrm{poly}(d)\, 2^{\mathrm{poly}(1/\epsilon)}$.

Corollary 1.3.44 follows from Theorem 1.3.43 by observing that halfspaces satisfy Definition 1.3.42 for $k = 1$ and $\Gamma \leq 1/\sqrt{2\pi}$.

As mentioned in the introduction, Corollary 1.3.44 answers an open question independently posed by Feldman [Fel08] and by Gopalan, Kalai, and Klivans [GKK08b]. Specifically, as we explain below, it implies a *super-polynomial* computational separation between agnostic query learning and agnostic learning with random samples for the

class of halfspaces.

In the vanilla agnostic PAC setting, the complexity of this problem is $d^{\mathrm{poly}(1/\epsilon)}$; the upper bound follows via the $L_1$-polynomial regression algorithm [KKMS08] which has complexity $d^{\Theta(1/\epsilon^2)}$ [DKN10] in this setting. The matching lower bound follows from a recent line of work, both in the SQ model [GGK20, DKZ20, DKPZ21] and under plausible cryptographic assumptions [DKR23, Tie23].

***Functions of Halfspaces.*** A more general concept class where our general approach is applicable is that consisting of all intersections (or arbitrary functions) of a bounded number of halfspaces. For the special case of intersections, we show:

**Corollary 1.3.45** (Agnostic Query Learning for Intersections of Halfspaces)**.** *There exists an agnostic query learner for intersections of $\ell$ halfspaces on $\mathbb{R}^d$ with running time* $\mathrm{poly}(d)\, O(\ell)^{\mathrm{poly}(\log(\ell)/\epsilon)}$.

Corollary 1.3.45 follows from Theorem 1.3.43 by observing that intersections of $\ell$ halfspaces satisfy Definition 1.3.42 for $k = \ell$ and that their Gaussian surface area is bounded above by $\Gamma = O(\sqrt{\log(\ell)})$, as shown by Nazarov (see, e.g., [KOS08, CCK17]).

Analogously to the case of a single halfspace, the complexity of the agnostic learning problem with random samples is significantly worse (as long as $\ell \ll d$), namely $d^{\mathrm{poly}(\log(\ell)/\epsilon)}$; the upper bound follows from [KOS08] and a qualitatively matching SQ lower bound was given in [DKPZ21, HSSV22].

Finally, for arbitrary functions of $\ell$ halfspaces, the Gaussian surface area is bounded by $\Gamma = O(\ell)$, leading to the following corollary:

**Corollary 1.3.46** (Agnostic Query Learning for Functions of Halfspaces)**.** *There exists an agnostic query learner for arbitrary functions of $\ell$ halfspaces on $\mathbb{R}^d$ with running time* $\mathrm{poly}(d)\, O(\ell)^{\mathrm{poly}(\ell/\epsilon)}$.

Similarly, the best known complexity upper bound with random samples is $d^{\mathrm{poly}(\ell/\epsilon)}$.

***Low-degree Polynomial Threshold Functions (PTFs).*** Another notable application is for the class of low-degree PTFs that depend on a low-dimensional subspace. A degree-$\ell$ PTF is any Boolean function $f : \mathbb{R}^d \to \{\pm 1\}$ of the form $h(\mathbf{x}) = \mathrm{sign}\,(p(\mathbf{x}))$, where $p : \mathbb{R}^d \to \mathbb{R}$ is a degree at most $\ell$ polynomial. Low-degree PTFs have been extensively studied in theoretical machine learning and specifically in the context of agnostic learning [DHK+10, DSTW10, DRST14, Kan11].

Here we consider a natural subclass of low-degree PTFs where the underlying polynomial is a subspace junta. Specifically, we consider the class of Boolean functions of the form $f(\mathbf{x}) = \mathrm{sign}\,(p(\mathrm{proj}_U \mathbf{x}))$, where $U$ is an unknown $k$-dimensional subspace and $p$ is a degree-$\ell$ polynomial in $k$ variables. Since the Gaussian surface area of this class of functions is bounded above by $\Gamma = O(\ell)$ [Kan11], we obtain the following corollary:

**Corollary 1.3.47** (Agnostic Query Learning for Low-Dimensional PTFs)**.** *There exists an agnostic query learner for degree-$\ell$ PTFs on $\mathbb{R}^d$ that depend on an unknown $k$-dimensional subspace with running time* $\mathrm{poly}(d)\, O(k)^{\mathrm{poly}(\ell/\epsilon)}$.

The above running time bound should be compared with the best known complexity bound of $d^{\mathrm{poly}(\ell/\epsilon)}$ for agnostic learning with samples [Kan11].

Table 1.2 summarizes our contributions for Boolean concept classes in comparison to prior work on agnostic PAC learning (with random samples only).

## Technical Overview

We leverage query access to develop a unified dimension-reduction framework for agnostically learning both real-valued and Boolean-valued multi-index models. As

Table 1.2: Learning Boolean Concepts using Queries: Running time comparisons of the best known agnostic learners (using random samples) with our Influence PCA technique (using queries).

| Concept Class | PAC (without queries) $L_1$ Regression [KOS08] | PAC+Queries **Influence PCA (Ours)** |
|---|---|---|
| Single Halfspace | $d^{\mathrm{poly}(1/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(1/\epsilon)}$ |
| Intersections of $k$ Halfspaces | $d^{\mathrm{poly}(\log(k)/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(\log(k)/\epsilon)}$ |
| Functions of $k$ Halfspaces | $d^{\mathrm{poly}(k/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(k/\epsilon)}$ |
| Degree-$\ell$, $k$-Dim. PTFs | $d^{\mathrm{poly}(\ell/\epsilon)}$ | $\mathrm{poly}(d)\, O(k)^{\mathrm{poly}(\ell/\epsilon)}$ |
| Low-Dim. Geometric Concepts | $d^{\mathrm{poly}(\Gamma/\epsilon)}$ | $\mathrm{poly}(d)\, O(k)^{\mathrm{poly}(\Gamma/\epsilon)}$ |

already explained after the statement of Theorem 1.3.35, natural dimension-reduction approaches that work in the realizable (noiseless) setting inherently cannot be extended to the agnostic setting.

At a high-level, our framework reduces the problem of agnostically learning MIMS in $d$ dimensions to agnostically learning the same class in $\mathrm{poly}(k/\epsilon)$ dimensions. It consists of three main steps:

- First we use queries to the label function to simulate gradient queries to a "smoothed" version $\widetilde{y}(\mathbf{x})$ of the adversarial label $y(\mathbf{x})$. We show that, as long as the concept class of interest has bounded variation (real-valued MIMs of Definition 1.3.34) or bounded Gaussian surface area (Boolean MIMs of Definition 1.3.42), a hypothesis that has low excess-error with respect to the smoothed label $\widetilde{y}$ will also have low excess error with respect to the original label $y(\mathbf{x})$; see Proposition 9.1.5.

- The second step uses gradient queries to the function $\widetilde{y}$ in order to compute an accurate estimate of the influence matrix of the "smoothed" label, namely $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\nabla\widetilde{y}(\mathbf{x})(\nabla\widetilde{y}(\mathbf{x}))^{\intercal}]$. We perform PCA on $\mathbf{M}$ and find the top eigenvectors (i.e., the eigen-directions whose corresponding eigenvalues are larger than some

threshold). This method is known as outer gradient product [XTLZ02]; in the context of learning/testing Boolean concepts, it has been used in [DKK+21a, DMN21]. (See Section (9.1.5) for a detailed summary of related work.) We show that those "high-influence" directions form a low-dimensional (i.e., of dimension $\text{poly}(k/\epsilon)$) subspace such that there exists a hypothesis that (i) depends only on the low-dimensional subspace, (ii) has bounded surface area/variation, and (iii) is close to our target function. That is, we effectively reduce the dimension of our original learning task from $d$ down to $\text{poly}(k/\epsilon)$.

- The third step is to solve an agnostic learning task of a bounded variation/surface area function in the low-dimensional subspace spanned by the top eigenvectors of $\mathbf{M}$. For this step, for learning real-valued MIMs, we rely on a generic $L_2$-regression algorithm; for learning Boolean concepts, we use the $L_1$-polynomial regression agnostic learner of [KKMS08, KOS08]. Those methods yield non-proper learning algorithms – to obtain proper-learners, we essentially perform a brute-force search over a net of the *low-dimensional* parameter space found in the previous step.

## Hardness of Proper Agnostic Query Learning for ReLUs and Halfspaces

Here we sketch our hardness reduction, establishing that the exponential dependence in $1/\epsilon$ is inherent for proper agnostic learners, even with query access to the function (see Theorem 9.5.3 and Theorem 9.5.4). In particular, we show that assuming there are no polynomial-time algorithms for the Small-Set Expansion (SSE) problem [RS10], then there are no polynomial time *proper* agnostic learning algorithms for ReLUs and homogeneous halfspaces with respect to the Gaussian distribution.

The basic idea of our argument is to reduce to the problem of (approximately) optimizing a homogeneous degree-4 polynomial over the unit sphere (for the case of halfspaces we reduce to optimizing a degree-5 polynomial). As there are already known reductions from SSE to the problem of finding approximate maxima of degree-4 polynomials (and for halfspaces we can do a simple reduction from degree-4 to degree-5) this will suffice.

For this, we note that if $f(\mathbf{x})$ is a polynomial and $g(\mathbf{x}) = \mathrm{ReLU}(\mathbf{v} \cdot \mathbf{x})$ for $\mathbf{v}$ a unit vector, then $\mathbf{E}[f(\mathbf{x})g(\mathbf{x})]$ is a low-degree polynomial in $\mathbf{v}$. In fact, by specifying $f$, we can make this into any homogeneous degree-5 polynomial we desire. This gives us SSE hardness of approximating $\mathbf{E}[f(\mathbf{x})g(\mathbf{x})]$.

If $f$ were a Boolean function we would be done. However, as this is not the case, we need two additional steps. Firstly, we scale down $f$ and truncate it so that its values stay within $[-1, 1]$ (note that this introduces only a small error if the average size of $f$ is small). Second, we replace $f$ by a random Boolean function $\tilde{f}$ so that $\mathbf{E}[\tilde{f}(\mathbf{x})] = f(\mathbf{x})$. Doing this, it is not hard to see that with high probability over the randomness of defining $\tilde{f}$ that $\mathbf{E}[\tilde{f}(\mathbf{x})g(\mathbf{x})]$ is arbitrarily close to $\mathbf{E}[f(\mathbf{x})g(\mathbf{x})]$ for all functions $g$.

Now even if the algorithm was given an explicit description of our function $\tilde{f}$, finding a ReLU function $g$ that approximately maximizes $\mathbf{E}[\tilde{f}(\mathbf{x})g(\mathbf{x})]$ is essentially equivalent to approximately optimizing a homogeneous degree-5 polynomial of the sphere, which is SSE-hard.

### 1.3.9 Robustly PAC Learning Multi-Index Models under the Gaussian Distribution

In this part, we develop a new algorithm for learning a broad family of multi-index models (MIMs) under the Gaussian distribution in the presence of adversarial label noise. Before we describe our general setting and algorithm, we start with a specific basic model that was the original motivation for this work.

**Motivation: Multiclass Linear Classification** A multiclass linear classifier is a function $f : \mathbb{R}^d \to [K]$ of the form $f(\mathbf{x}) = \mathrm{argmax}_{i\in[K]}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i)$ , where $\mathbf{w}^{(i)} \in \mathbb{R}^d$ and $t_i \in \mathbb{R}$ (ties are resolved by selecting the smallest index). Multiclass linear classification (MLC) is a textbook machine learning problem [SSBD14], that has also been extensively studied in the online learning literature; see, e.g., [KSST08] and references therein. In the realizable PAC setting (i.e., with clean labels), this task is efficiently solvable to any desired error $\epsilon$ with sample size $O(dK/\epsilon)$ via linear programming (or via the multiclass Perceptron algorithm for well-behaved marginal distributions, including the Gaussian). On the other hand, despite extensive effort, see, e.g., [NDRT13, WLT18, LWS18, SKP+20, GWBL20], our understanding of the computational complexity of learning is poor in the noisy setting—even for the simplest case of *random* label noise.

The special case of *binary* classification of Linear Threshold Functions (LTFs) or halfspaces—corresponding to $K = 2$—is fairly well-understood. Specifically, sample and computationally efficient learning algorithms are known, even in the presence of partially corrupted labels [ABL17, DKS18, DGT19, DKTZ20a, DKTZ20c, DKK+21b, DKK+22]. Unfortunately, these prior approaches are specific to the binary case. In particular, it is unclear whether they can be generalized to give any non-trivial guarantees for the noisy

multiclass setting.

Perhaps the simplest model of random label noise is that of Random Classification Noise (RCN) [AL88]. In the binary classification setting, the RCN model posits that the label of each example is independently flipped with probability exactly $\eta$, for some noise parameter $0 < \eta < 1/2$. A slight generalization of the noise model, that is essentially equivalent, allows for different probabilities $\eta_+, \eta_- \in (0, 1/2)$ of flipping the label of positive and negative examples. A classical result [Kea98] shows that, in the binary classification setting, any efficient Statistical Query (SQ) algorithm can be transformed into an efficient RCN-tolerant PAC learner (with a polynomial dependence in all relevant parameters). Unfortunately, no such result is known for the multiclass setting[9].

The standard generalization of the RCN model to the multiclass setting is the following [PRKM+17, ZLA21, VRW17]. Let $D$ be the distribution on noisy samples $(\mathbf{x}, y) \in \mathbb{R}^d \times [K]$, where in our setting the marginal distribution $D_{\mathbf{x}}$ is the standard Gaussian, and $f : \mathbb{R}^d \to [K]$ be the ground truth classifier. Then there exists a row-stochastic matrix $\mathbf{H} \in \mathbb{R}_+^{K \times K}$ such that $\mathbf{H}_{i,j} = \mathbf{Pr}_{(\mathbf{x},y)\sim D}[y = i \mid f(\mathbf{x}) = j]$. Moreover, it is additionally assumed that there exists $\gamma > 0$ such that $\mathbf{H}_{i,i} \geq \mathbf{H}_{i,j} + \gamma$, for all $i \neq j \in [K]$[10]. Under these assumptions, MLC is information-theoretically solvable using $O(dK/(\gamma\epsilon))$ samples. On the other hand, no $\mathrm{poly}(d, K)$ time learning algorithm is known. Prior algorithmic work using gradient-based methods to achieve sample and time complexity that scales polynomially with $1/\sigma$, where $\sigma$ is the minimum singular value of $\mathbf{H}$ (see [LWS18]). This quantity could be arbitrarily small or even zero, hence such approaches do not in general lead to a computationally efficient algorithm. This

[9] In fact, Theorem 1.3.57 implies that such a generalization is impossible without strong assumptions on the noise matrix.

[10] Roughly speaking, the parameter $\gamma > 0$ plays the role of $1 - 2\eta$ in the binary case.

discussion leads to the following question:

**Question 1.3.48.** Is there a $\mathrm{poly}(d, K)$ time algorithm for multiclass linear classification in the presence of Random Classification Noise?

Perhaps surprisingly, we provide strong evidence that the answer to the above question is negative. Specifically, we establish a Statistical Query (SQ) lower bound showing that the complexity of any SQ algorithm for this task is $d^{\Omega(K)}$ (see Theorem 1.3.57).

An even more challenging goal is to understand the algorithmic possibilities and limitations of learning MLCs in the presence of *adversarial* label noise. In the adversarial label noise model, an adversary is allowed to arbitrarily corrupt an unknown OPT fraction of the labels, for some parameter $\mathrm{OPT} < 1/2$, and the goal of the learner is to compute a hypothesis whose 0-1 error is competitive with OPT (see Definition 1.3.52). The information-theoretically optimal error guarantee in this model is $\mathrm{OPT} + \epsilon$. Unfortunately, even for $K = 2$, known computational lower bounds rule out the existence of such algorithms with complexity better than $d^{\mathrm{poly}(1/\epsilon)}$ [DKPZ21, DKR23].

Consequently, to achieve a fixed polynomial runtime in $d$, one needs to relax the error guarantee. For the special case $K = 2$, such algorithms are known. In particular, [DKS18] gave a $\mathrm{poly}(d/\epsilon)$ time algorithm with error guarantee $C \cdot \mathrm{OPT} + \epsilon$, where $C$ is a universal constant.

A natural question is whether qualitatively similar algorithms exist for the multiclass setting. Specifically, a first goal is to establish a fixed-parameter tractability guarantee:

**Question 1.3.49.** Is there a *constant-factor* approximate learner for MLC with complexity $\mathrm{poly}(d) f(\epsilon, K)$?

As a corollary of our approach, we answer this question in the affirmative (see Theorem 1.3.56).

## Main Results

In this section, we present our main results. For a concise overview of our algorithmic contributions, see Table 1.3.

| **Function Class** | **Prior Work** | | **Our Work** | |
|---|---|---|---|---|
| | **Runtime** | **Error** | **Runtime** | **Error** |
| Agnostic $K$-MLC | $d^{\mathrm{poly}(K/\epsilon)}$ | $\mathrm{OPT}+\epsilon$ | (Theorem 1.3.56) $\mathrm{poly}(d)\,2^{\mathrm{poly}(K/\epsilon)}$ | $O(\mathrm{OPT})+\epsilon$ |
| Agnostic Intersections of $K$ halfspaces | $d^{O(\log(K)/\epsilon^4)}$ $\frac{\mathrm{poly}(d)}{\epsilon^{\mathrm{poly}(K)}}$ | $\mathrm{OPT}+\epsilon$ $\mathrm{poly}(K)\tilde{O}(\mathrm{OPT}^{\frac{1}{11}})$ | (Theorem 1.3.58) $\mathrm{poly}(d)\,2^{\mathrm{poly}(K/\epsilon)}$ | $K\tilde{O}(\mathrm{OPT})+\epsilon$ |
| Well-Behaved $K$-MIMs (Definition 1.3.50) | $d^{\mathrm{poly}(\Gamma/\epsilon)}$ | $\mathrm{OPT}+\epsilon$ | (Theorem 1.3.53) $d^{O(m)}2^{\mathrm{poly}(mK\Gamma/\epsilon)}$ | $\tau+\mathrm{OPT}+\epsilon$ |

Table 1.3: Comparison of runtime and error guarantees between state-of-the-art efficient algorithms (see [KOS08, DKS18] for intersections of $K$ halfspaces) and our subspace approximation procedure.

### Learning Well-behaved Multi-Index Models with Noise

Our general algorithmic result (Theorem 1.3.53) applies to MIMs with finite range that satisfy a collection of regularity properties (Definition 1.3.50).

Throughout this section, we will use $d$ for the dimension of the feature space and $K$ for the dimension of the hidden low-dimensional subspace. A standard assumption in the literature on MIMs (see, e.g., [KOS08, DKK+24]), satisfied by a wide range of natural concept classes, is that the underlying function has bounded Gaussian surface area (GSA)—a complexity measure of the decision boundary (see Definition 10.3.1 for the definition in the multiclass case). We denote by $\Gamma$ the upper bound on the GSA of the underlying concept class.

The structural property that we will assume for every MIM function $f$ in our class is, roughly speaking, the following: it is not possible to add noise to $f$ to conceal crucial

directions so that they remain invisible to moment-based detection. Specifically, for *any* subspace $V$, one of the following must hold:

(a) $V$ already captures all essential information of $f$, in which case there is a function depending solely on the projection onto $V$ that approximates $f$ well.

(b) Otherwise, there is at least one point in $V$ such that when we condition on $\mathbf{x}^V$ near that point, the conditional distribution of a "noisy" version of $f$ exhibits a distinguishing moment of degree at most $m$. That is, there exists a degree-$m$ polynomial whose conditional expectation, conditioned on $\mathbf{x}^V$ (the projection of $\mathbf{x}$ on $V$), cannot remain hidden; and hence reveals a dependence on directions outside the subspace $V$.

The parameter $m$ above turns out to essentially characterize the complexity of the learning problem. We show, qualitatively speaking, that there exists an algorithm with complexity $d^m$; and that this bound is essentially best possible in the SQ model.

We now present an informal definition of this property and refer the reader to Definition 10.3.2 for a more precise formulation.

**Definition 1.3.50** (Well-Behaved Multi-Index Model)**.** *Fix $m \in \mathbb{Z}_+, \zeta, \tau \in (0,1)$, $\Gamma > 0$, and $\mathcal{Y} \subseteq \mathbb{Z}$ a set of finite cardinality. We say that a $K$-MIM $f : \mathbb{R}^d \to \mathcal{Y}$ is $(m, \zeta, \tau, \Gamma)$-well-behaved if $f$ has Gaussian surface area at most $\Gamma$, and for any random variable $y(\mathbf{x})$ supported on $\mathcal{Y}$ with $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) \neq y(\mathbf{x})] \leq \zeta$ and any subspace $V \subseteq \mathbb{R}^d$, one of the following conditions holds:*

*a) there exists a function $g : V \to \mathcal{Y}$ such that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}^V) \neq f(\mathbf{x})] \leq \tau$, or*

*b) with non-negligible probability*[11] *over* $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$*, there exists a mean-zero, variance-one polynomial* $p : U \to \mathbb{R}$ *of degree at most* $m$ *and* $z \in \mathcal{Y}$ *such that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[p(\mathbf{x}^U)\mathbb{1}(y(\mathbf{x}) = z) \mid \mathbf{x}^V = \mathbf{x}_0^V] \geq \mathrm{poly}(\zeta/m) \ ,$$

*where* $U = (V + W) \cap V^\perp$ *is the projection of* $W$ *onto* $V^\perp$ *and* $W$ *is the subspace defining* $f$*.*

*We denote by* $\mathcal{F}$ *the class of* $(m, \zeta, \tau, \Gamma)$*-well-behaved* $K$*-MIMs over* $\mathbb{R}^d$*.*

The aforementioned property enforces that no direction crucial to $f$ can be fully hidden when we look at a suitable low-degree moment. Intuitively, the parameter $\zeta$ in Definition 1.3.50 will determine the noise-tolerance of our learning algorithm, while the parameter $\tau$ will determine the final error guarantee of the algorithm. The parameter $m$ is the minimum degree of a non-zero moment on the subspace $U$. It is worth pointing out that the sample and computational complexity of our algorithm will scale with $d^m$.

**Remark 1.3.51.** *Definition 1.3.50 implicitly introduces a complexity measure for learning a family of discrete MIMs. Let* $\mathcal{C}$ *be a family of MIMs with Gaussian Surface Area bounded by* $\Gamma$*, and let* $\zeta, \tau > 0$ *be parameters. Define* $m^*$ *to be the* smallest integer *for which every* $f \in \mathcal{C}$ *is* $(m^*, \zeta, \tau, \Gamma)$*-well-behaved. It turns out that the quantity* $m^*$ *is (essentially) a generalization of the generative exponent [DPVLB24], a measure characterizing the SQ complexity of learning Single-Index Models (SIMs). For further details, see Section (*10.5.1*).*

As we will show in the next subsection, fundamental and well- studied concept classes—specifically, multiclass linear classifiers and intersections of halfspaces—satisfy

[11] By this we mean that the probability is not-too-small, specifically it is independent of the dimension $d$.

Definition 1.3.50 for reasonable values of the underlying parameters (in particular, with $m \leq 2$).

Before we state our main algorithmic result, we recall the definition of PAC learning with adversarial label noise (aka agnostic learning) under the Gaussian distribution.

**Definition 1.3.52** (Agnostic Multiclass Classification)**.** *Let $\mathcal{Y}$ be finite and $\mathcal{C}$ a class of functions from $\mathbb{R}^d$ to $\mathcal{Y}$. Given i.i.d. samples $(\mathbf{x}, y)$ from a distribution $D$ on $\mathbb{R}^d \times \mathcal{Y}$ such that the marginal $D_{\mathbf{x}}$ is the standard normal and no assumptions are made on the labels, the goal is to output a hypothesis $h : \mathbb{R}^d \to \mathcal{Y}$ that, with high probability, has small 0-1 error $\mathrm{err}^D_{0-1}(h) \stackrel{\mathrm{def}}{=} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y]$, as compared to the optimal 0-1 error of the class $\mathcal{C}$, i.e., $\mathrm{OPT} \stackrel{\mathrm{def}}{=} \inf_{c\in\mathcal{C}} \mathrm{err}^D_{0-1}(c)$.*

Our main result is a general algorithm for learning well-behaved MIMs in the presence of adversarial label noise (see Theorem 10.3.3).

**Theorem 1.3.53** (Learning Well-Behaved MIMs with Adversarial Label Noise)**.** *Let $\mathcal{F}$ be the class of $(m, \zeta, \tau, \Gamma)$-well-behaved $K$-MIMs on $\mathbb{R}^d$ of Definition 1.3.50. If $\zeta \geq \mathrm{OPT} + \epsilon$, there exists an agnostic algorithm that draws $N = d^m \, 2^{\mathrm{poly}(mK\Gamma/\epsilon)}$ i.i.d. samples, it runs in $\mathrm{poly}(d, N)$ time, and computes a hypothesis $h$ such that with high probability $\mathrm{err}^D_{0-1}(h) \leq \tau + \mathrm{OPT} + \epsilon$.*

At a high-level, the main advantage of the above algorithm is that—as long as $m = O(1)$—its complexity is bounded by a *constant-degree polynomial in $d$*. In other words, we give a fixed-parameter tractability result for the task of learning well-behaved MIMs. While the complexity guarantee we obtain is new even in the realizable setting, our algorithm is additionally robust to adversarial label noise—sometimes achieving near-optimal noise-tolerance (within constant factors). In addition to its own merits, the significance

of our general algorithm is illustrated by its downstream applications (given in the following subsection).

An interesting feature of Theorem 1.3.53 is that the complexity upper bound it establishes is qualitatively tight in the Statistical Query (SQ) model, as a function of the dimension $d$. Roughly speaking, if a function class does not satisfy the assumptions of Definition 1.3.50 for a specific value of $m$, then any SQ algorithm that achieves error competitive with Theorem 1.3.53 must have SQ complexity $d^{\Omega(m)}$. In more detail, we show (see Theorem 12.1.1 for a more detailed formal statement).

**Theorem 1.3.54** (SQ Lower Bound for Learning Not-Well-behaved MIMs, Informal)**.** *Suppose that a function class $\mathcal{G}$ is rotationally invariant and contains a bounded surface area MIM not contained in the function class $\mathcal{F}$ of Definition 1.3.50 for parameters $m, \tau, \zeta$. Then, any SQ algorithm that learns $\mathcal{G}$ in the presence of $\zeta$ label noise to error smaller than $\tau - O(\zeta)$ has complexity at least $d^{\Omega(m)}$.*

**Remark 1.3.55** (Characterization)**.** *Theorems 1.3.53 and 1.3.54 together essentially characterize the complexity of learning a MIM family. In particular, let $\mathcal{C}$ be a rotationally invariant family of MIMs with Gaussian Surface Area bounded by $\Gamma$, and let $\zeta, \tau > 0$ be parameters. Define $m^*$ to be the* smallest integer *for which every $f \in \mathcal{C}$ is $(m^*, \zeta, \tau, \Gamma)$-well-behaved.*

*Then there is an SQ algorithm that learns a (slightly less than) $\zeta$-noisy function from $\mathcal{C}$ to error $\tau + O(\zeta)$ with resources $d^{m^*}$ (times some function of the other parameters), but not an SQ algorithm that learns a $\zeta$-noisy function from $\mathcal{C}$ to error $\tau - O(\zeta)$ with resources $d^{o(m^*)}$.*

### Concrete Learning Applications to Well-studied Concept Classes

In this subsection, we present algorithmic applications of our general MIM learner to two fundamental and well-studied concept classes.

**Multiclass Linear Classifiers** We denote by $\mathcal{L}_{d,K}$ the family of $K$-class linear classifiers in $\mathbb{R}^d$. For the adversarial noise setting, we obtain the following result (see Theorem 10.2.1):

**Theorem 1.3.56** (Learning $\mathcal{L}_{d,K}$ with Adversarial Label Noise)**.** *There exists an agnostic learning algorithm that draws $N = d\,2^{\mathrm{poly}(K/\epsilon)}$ i.i.d. labeled samples, runs in $\mathrm{poly}(d, N)$ time, and outputs a hypothesis $h$ such that with high probability $\mathrm{err}^D_{0-1}(h) \leq O(\mathrm{OPT}) + \epsilon$, where OPT is defined with respect to $\mathcal{L}_{d,K}$.*

Theorem 1.3.56 can be derived from Theorem 1.3.53 via a new structural result that we establish, roughly showing that the class $\mathcal{L}_{d,K}$ is $(1, \zeta, O(\zeta), O(K))$-well behaved for any $\zeta \in (0, 1)$ (see Proposition 10.2.4). For the sake of exposition, in Section 10.2 we give a direct algorithm establishing Theorem 1.3.56.

Theorem 1.3.56 gives the first *constant-factor* approximate learner for multiclass linear classification with adversarially corrupted labels, whose complexity is a *fixed-degree polynomial in $d$*. We provide some remarks to interpret this statement. First, even for $K = 2$, known hardness results imply that achieving error $\mathrm{OPT} + \epsilon$ requires $d^{\mathrm{poly}(1/\epsilon)}$ time [DKPZ21, DKR23]. Hence, to achieve a fixed-degree $\mathrm{poly}(d)$ complexity guarantee, it is necessary to relax the final accuracy. The best thing to hope for here is a constant-factor approximation, i.e., a final error of $O(\mathrm{OPT}) + \epsilon$ (like the one we achieve). For the $K = 2$ case, prior work achieves such an error bound [DKS18] (earlier work [ABL17] achieves such a constant-factor approximation for the special case of homogeneous

halfspaces); and in fact with complexity $\mathrm{poly}(d/\epsilon)$. On the other hand, *no non-trivial approximation algorithm was previously known for the multiclass case.* Importantly, no polynomial-time learner with error $o(d)\,\mathrm{OPT} + \epsilon$ is known even for $K = 3$. The crucial difference is that the approximation ratio of our algorithm is an *absolute constant*, independent of the dimension and the number of classes.

We also investigate the complexity of learning multiclass linear classifiers in the presence of RCN. We start by establishing an SQ lower bound of $d^{\Omega(K)}$ for achieving $\mathrm{OPT} + \epsilon$ error. Specifically, we show (see Theorem 12.1.4):

**Theorem 1.3.57** (SQ Lower Bound for Learning $\mathcal{L}_{d,K}$ with RCN)**.** *For any $\epsilon \leq 1/K^C$, where $C$ a sufficiently large universal constant, any SQ algorithm that learns $\mathcal{L}_{d,K}$ under the Gaussian distribution in the presence of RCN to error $\mathrm{OPT} + \epsilon$, has complexity at least $d^{\Omega(K)}$.*

**Intersections of Halfspaces** An intersection of $K$ halfspaces is any function $f : \mathbb{R}^d \to \{\pm 1\}$ for which there exist $K$ halfspaces $h_i(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i)$ such that $f(\mathbf{x}) = 1$ if and only if $h_i(\mathbf{x}) = 1$ for all $i \in [K]$. We denote by $\mathcal{F}_{K,d}$ the class of intersections of $K$ halfspaces on $\mathbb{R}^d$.

The family of intersections of halfspaces is a fundamental concept class that has been extensively studied in many contexts for at least the past five decades. We show that our general MIM algorithm can be applied to $\mathcal{F}_{K,d}$, leading to a novel accuracy-complexity tradeoff in the agnostic setting. Specifically, we obtain the following result (see Theorem 10.4.1 for a more detailed statement):

**Theorem 1.3.58** (Learning $\mathcal{F}_{K,d}$ with Adversarial Label Noise)**.** *There exists an agnostic learning algorithm that draws $N = O(d^2)\, 2^{\mathrm{poly}(K/\epsilon)}$ i.i.d. labeled samples, runs in time*

$\mathrm{poly}(d, N)$, *and computes a hypothesis* $h$ *such that with high probability* $\mathrm{err}^D_{0-1}(h) \leq K\widetilde{O}(\mathrm{OPT}) + \epsilon$, *where* OPT *is defined with respect to* $\mathcal{F}_{K,d}$.

Theorem 1.3.58 is obtained from Theorem 1.3.53 via a new structural result, roughly showing that the class $\mathcal{F}_{d,K}$ is $(2, \zeta, K\tilde{O}(\zeta), O(\sqrt{\log(K)}))$-well behaved for any $\zeta \in (0, 1)$ (see Proposition 10.4.2).

To highlight the significance of Theorem 1.3.58, some remarks are in order. Note that the information-theoretically optimal error in our setting is $\mathrm{OPT} + \epsilon$. The only known algorithm achieving this guarantee is the $L_1$ polynomial regression [KKMS08, KOS08], which has complexity $d^{O(\log K/\epsilon^4)}$ for the class $\mathcal{F}_{d,K}$. (This complexity bound was recently shown to be near-optimal in the SQ model [DKPZ21, HSSV22].)

A fundamental open question in the literature has been whether one can achieve *near-optimal error* guarantees with qualitatively improved sample and computational complexity. Theorem 1.3.58 answers this question in the affirmative. Specifically, for $K = O(1)$, it achieves near-optimal error guarantee (up to a $O(\log(1/\mathrm{OPT}))$ factor), and its complexity has *fixed-polynomial dependence in the dimension* (namely, $d^2$ as opposed to $d^{f(k,1/\epsilon)}$).

In comparison, [DKS18] developed a $\mathrm{poly}(d)(1/\epsilon)^{\mathrm{poly}(K)}$ time algorithm, achieving an error guarantee of $\mathrm{poly}(K)\tilde{O}(\mathrm{OPT}^{1/11})$. The key distinction is that the error guarantee achieved by the [DKS18] algorithm is highly suboptimal as a function of OPT (by a large polynomial factor). Additionally, the dependence on $K$ in our Theorem 1.3.58 is significantly better by a polynomial factor.

## Overview of Techniques

**Prior Techniques and Obstacles for Learning Multi-Index Models** Recall that our goal is, given access to noisy samples from a well-behaved MIM with finite label set, to learn the underlying function $f$ to high accuracy. We will use $W$ to denote the hidden $K$-dimensional subspace to whose projection $f$ depends. The standard approach for solving such learning problems (see, e.g., [Vem10b]) is to first learn some suitable approximation $V$ to $W$, and then use exhaustive search to learn a function of the projection onto $V$. Computing such an approximation to $W$ is usually achieved by estimating the low-order moments of the level sets of $f$. In particular, for each output value $i$ and each low-degree polynomial $p$, we want to approximate the expectation of $p(\mathbf{x})\mathbb{1}(f(\mathbf{x}) = i)$. In the absence of label noise, these moments should depend solely on how $p$ interacts with the directions aligned with $W$. Using this property, one can extract an approximation of a subspace of $W$, by choosing $V$ to be the space spanned by all directions with reasonably large moments. This approach yields an algorithm with runtime approximately $d^{O(m)}$, where $m$ is the number of moments being computed, plus some (often exponential) function of $\dim(V)$—which is typically regarded as "constant", since $d$ is much larger than $\dim(V)$.

Unfortunately, there may exist directions within $W$ where all low-degree moments vanish. An innocuous such example occurs when $f$ depends only on the projection of $\mathbf{x}$ onto a subspace $W' \subseteq W$. In this case, all moments would vanish for directions in $W \cap W'^{\perp}$. However, if such moments vanish for different reasons—for example, due to label noise—the aforementioned approach could be problematic. Therefore, in order to analyze this class of algorithms, one needs to prove that every direction $\mathbf{v} \in W$ for which the moments of $f$ vanish is a direction where $f$ approximately does not depend

on—i.e., $f$ is close to a function of the projection onto $\mathbf{v}^\perp$. Unfortunately, showing that the moments of $f$ are non-vanishing for all relevant directions can be challenging, which limits the applicability of this approach.

**New Dimensionality Reduction Technique** The main algorithmic contribution of this work is the design of a new, more general, approach that leads to significant improvements on the aforementioned dimensionality reduction technique. At a high-level, we design an iterative dimensionality reduction procedure (Algorithms 26 and 27) working as follows. At each step/iteration, after learning an approximating subspace $V$, we partition $V$ into small rectangular regions $S$, each of probability mass at least $\alpha$, where $\alpha$ is a carefully selected parameter. We then compute the first $m$ moments of $y(\mathbf{x})$, where $y(\mathbf{x})$ is the noisy version of $f(\mathbf{x})$, conditioned on the projection of $\mathbf{x}$ to $V$ lying in each region $S$; this step requires roughly $d^m(1/\alpha)$ samples and runtime (see Algorithm 27). In the noiseless case (i.e., with clean labels), non-vanishing moments should still align with $W$, providing us with a better approximation. We then find all directions with non-vanishing moments across all regions $S$, add them to $V$, and repeat the process. We terminate this procedure once $y$ is sufficiently close to a function $g$ of the projection of $\mathbf{x}$ onto $V$. Finally, using exhaustive search methods, we can recover an $\epsilon$-approximation to $g$, using roughly $(1/\epsilon)^{\dim(V)}$ samples and runtime.

The main advantage of our new method is that it succeeds under much weaker conditions, as compared to previous approaches. As mentioned earlier, prior techniques require that every direction in $W$ to be either unimportant or have a non-trivial low-degree moment. In contrast, our approach only requires that for each subspace $V$, either $f$ is approximately a function of the projection onto $V$, or that there exists some region $S$ with a non-trivial low-degree moment in a direction in $V^\perp$ correlated with $W$ (see

Definition 10.3.2). If we can establish this structural property, it follows that unless $f$ is already close to a function of the projection onto $V$ (in which case we can apply a brute-force method), one iteration of the algorithm will find a direction in $V^{\perp}$ that is at least slightly correlated with $W$. As a result, in the next iteration, the distance between $V$ and $W$ decreases; hence, our algorithm will eventually terminate.

This generalization of the previously used one-shot dimension reduction technique has two main advantages. Practically speaking, the more relaxed condition is often easier to prove. This allows, for example, our qualitative improvements in the error for robustly learning intersections of halfspaces. As far as we know, the one-shot dimension reduction technique may suffice to obtain error $\tilde{O}(\mathrm{OPT})$ in this case. However, the difficulty of proving such results means that the best analysis we know of gives substantially worse bounds (and only after a quite involved analysis). On the other hand, the condition needed for our iterative dimension reduction is somewhat easier to obtain, giving us a substantially better $\tilde{O}(K\mathrm{OPT})$ error with a simpler analysis.

Furthermore, our new technique allows us to efficiently learn a broader family of MIMs compared to what was previously possible with one-shot dimension reduction techniques, even in principle. In particular, as we show in Section 10.6, there exist Multi-Index Models that cannot be learned to better than constant error with any constant number of moments using the one-shot technique; and on the other hand, can be efficiently learned to arbitrary accuracy using only second moments via our iterative technique.

One disadvantage of our method as described above is that, while the runtime is typically polynomial in the dimension $d$, the dependence on other parameters can be problematic. Specifically, the number of regions $S$ is usually exponential in $\dim(V)$;

hence, at each iteration, $\dim(V)$ can increase in the worst-case by an exponential amount. This can lead to power-tower dependencies on some parameters. We circumvent this issue by assuming that our function has a small fraction of regions $S$ with non-vanishing moments. With this assumption, we can only add directions with large moments in a reasonable fraction of regions using a filtering step (see Line 13), keeping the increase in $\dim(V)$ polynomial. For an illustration of our iterative procedure for the special case of multiclass linear classification, see Figure 10.1 in Section 10.2.

**SQ Lower Bounds** An appealing feature of our algorithmic technique is that it is provably qualitatively tight in the Statistical Query (SQ) model (see Chapter 12). In particular, suppose that a function $f$ has a subspace $V$ so that on the preimage of the projection onto any point $\mathbf{x}$ in $V$, the low-degree moments of $f$ are nearly vanishing. Then we prove an SQ lower bound showing that it is SQ-hard to distinguish between (i) the distribution $(\mathbf{x}, h(\mathbf{x}))$, where $h$ is a random rotation of the function $f$ about $V$, and (ii) a product distribution $(\mathbf{x}, y)$, where $\mathbf{x}$ and $y$ are conditionally independent on $\mathbf{x}^V$ (see Theorem 12.1.1). This means that unless $f$ is close to some function of the projection onto $V$, it is hard to learn $f$ in the SQ model. In more detail, unless $f$ has the property that for each $V$, either $f$ is $\tau$-close to a function of the projection onto $V$, or there is some preimage with a non-trivial moment of degree at most $m$, then any SQ algorithm that learns $f$ to accuracy better than roughly $\tau$ either uses exponentially many queries or at least one query of accuracy smaller than $d^{-\Omega(m)}$. In summary, our algorithm learns $f$ if and only if there is an SQ algorithm that does so with comparable parameters as a function of $d$.

Finally, we give an overview of our SQ lower bound for learning multiclass linear classifiers with RCN (see Theorem 12.1.4 in Section (12.1.2)). This result leverages

the fact that statistical queries correspond to evaluations of the confusion matrix (see Definition 12.1.3).

If this confusion matrix is close to being non-invertible, then there is missing information between the noisy and noiseless queries. To prove this, we find a low-dimensional concept, $f$, and a corresponding confusion matrix, $\mathbf{H}$, for which the first $\Omega(K)$ moments of the noisy label sets match the moments of the standard Gaussian distribution. As a result, we establish SQ-hardness in distinguishing between the distribution of random rotations of $f$ in the $d$-dimensional space and a product distribution.

**Applications to Well-studied Concept Classes** We show that our iterative dimension-reduction technique yields significantly improved learners for multiclass linear classifiers and intersections of halfspaces in the presence of noise. In the regime of agnostically learning multiclass classifiers with $K$ classes (see Section 10.2), we show that our algorithm achieves an error of $O(\mathrm{OPT}) + \epsilon$, using roughly $d2^{\mathrm{poly}(K/\epsilon)}$ samples and runtime (see Theorem 10.2.1). We show that if the noisy labels $y$ are at least $\Theta(\mathrm{OPT}) + \epsilon$ far from any function depending on the projection onto a subspace $V$ (otherwise, an exhaustive search in $V$ would yield a function $g$ with the desired error bound), then at least an $\epsilon$ fraction of our partition $S$ of $V$, will have the optimal $f$ more than $\Theta(\mathbf{Pr}[f(\mathbf{x}) \neq y \mid S])$ far from any constant function. In this case, we establish a new structural result (see Proposition 10.2.8) showing that the first moments of the sub-level sets of $y$ align with $W$ on a non-trivial fraction of the partition sets $S$. By using this structural result, we conclude that the class of multiclass linear classifiers is $(1, \zeta, O(\zeta), O(K))$-well behaved, as in Definition 1.3.50, for any $\zeta \in (0, 1)$. Therefore, by applying our general algorithm we can obtain a learner with the aforementioned efficiency (Theorem 1.3.53)

Furthermore, we obtain an algorithm with improved time-accuracy tradeoff for agnostically learning intersections of $K$ halfspaces under the Gaussian distribution (see Section 10.4). Specifically, we show that an error of $K\widetilde{O}(\mathrm{OPT})$ can be achieved with complexity $d^2 2^{\mathrm{poly}(K/\epsilon)}$ (see Theorem 10.4.1). An analogous structural result to the one we showed for multiclass linear classifiers can be established to prove that unless the optimal classifier is $K\widetilde{O}(\mathrm{OPT})$-close to a constant function, there exists a low-order moment (of degree at most two) that correlates non-trivially with the unknown subspace $W$ (see Proposition 10.4.4). Using this result, we deduce that the class of intersections of $K$ halfspaces is $(2, \zeta, K\tilde{O}(\zeta), O(\sqrt{\log(K)}))$-well behaved, as in Definition 1.3.50, for any $\zeta \in (0, 1)$. Hence, our learner again follows by applying our general iterative dimension-reduction technique.

## Part IV: Statistial Computational Limits Limits Via SQ Complexity

While the preceding results establish algorithmic progress across a range of structured learning problems, a natural question arises: to what extent are these results optimal? To rigorously understand the computational limits of learning in these regimes, this thesis also develops nearly-matching lower bounds under the Statistical Query (SQ) model.

### 1.3.10 SQ Lower Bounds Using Linear Programming

While agnostic learning is known to be computationally intractable in a distribution-independent setting, even for simple concept classes, the problem becomes more tractable under specific distributional assumptions. A significant body of research has been

dedicated to developing efficient agnostic learning algorithms for natural concept classes, with a particular emphasis on the Gaussian distribution. For Boolean-valued concept classes, the problem of agnostic learning under Gaussian marginals is formally defined as follows:

**Definition 1.3.59** (Agnostic Learning Boolean-valued Functions with Gaussian Marginals)**.** *Let $\mathcal{C}$ be a class of Boolean-valued concepts on $\mathbb{R}^n$. Given i.i.d. samples $(\mathbf{x}, y)$ from a distribution $D$ on $\mathbb{R}^n \times \{\pm 1\}$, where the marginal $D_{\mathbf{x}}$ on $\mathbb{R}^n$ is the standard Gaussian $\mathcal{N}_n$ and no assumptions are made on the labels $y$, the goal is to output a hypothesis $h : \mathbb{R}^n \to \{\pm 1\}$ such that with high probability we have $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, where $\mathrm{OPT} = \inf_{f\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y]$.*

The only known algorithmic technique for agnostic learning in the setting of Definition 1.3.59 is the $L^1$-polynomial regression algorithm [KKMS08]. This algorithm uses linear programming to compute a low-degree polynomial that minimizes the $L^1$-distance to the target function. Its performance hinges on how well the underlying concept class $\mathcal{C}$ can be approximated, in $L^1$-norm, by low-degree polynomials. In more detail, if $d$ is the (minimum) degree such that any $f \in \mathcal{C}$ can be $\epsilon$-approximated (in $L^1$-norm) by a degree-$d$ polynomial, the algorithm has sample complexity and running time $n^{O(d)}/\mathrm{poly}(\epsilon)$ and outputs a hypothesis with misclassification error $\mathrm{OPT} + \epsilon$.

For many important geometric concept classes (including Linear Threshold Functions [KKMS08, DGJ$^+$10, DKN10], Polynomial Threshold Functions [DHK$^+$10, Kan11, DRST14, HKM14], and Intersections of Halfspaces) [KKMS08, KOS08, Kan14], this algorithm acts as a polynomial-time approximation scheme (PTAS) because the required degree $d$ depends only on the error parameter $\epsilon$ and not on the dimension $n$. Specifically, for the class of LTFs under the Gaussian distribution, the $L^1$-regression algorithm is

known to have sample and computational complexity of $n^{O(1/\epsilon^2)}$.

Since $L^1$-polynomial regression is the fastest known agnostic learner for these fundamental classes, it is natural to ask:

**Question 1.3.60.** Can we beat $L^1$-polynomial regression for agnostic learning under Gaussian marginals?

As our first main contribution, we answer the above question in the negative for all concept classes satisfying some mild properties (including all the geometric concept classes mentioned above). Our lower bound applies for the class of Statistical Query (SQ) algorithms. Statistical Query (SQ) algorithms are a class of algorithms that are allowed to query expectations of bounded functions of the underlying distribution rather than directly access samples. Formally, an SQ algorithm has access to the following oracle.

**Definition 1.3.61** (STAT Oracle)**.** *Let $D$ be a distribution on labeled examples supported on $X \times [-1,1]$, for some domain $X$. A statistical query is a function $q : X \times [-1,1] \to [-1,1]$. We define $STAT(\tau)$ to be the oracle that given any such query $q(\cdot,\cdot)$ outputs a value $v$ such that $|v - \mathbf{E}_{(\mathbf{x},y)\sim D}[q(\mathbf{x},y)]| \leq \tau$, where $\tau > 0$ is the tolerance parameter of the query.*

The SQ model was introduced by [Kea98] as a natural restriction of the PAC model [Val84] and has been extensively studied in learning theory; see, e.g., [FGR+17, FPV15, FGV17, Fel17] for some recent references. The reader is referred to [Fel16] for a survey. The class of SQ algorithms is fairly broad: a wide range of known algorithmic techniques in machine learning are known to be implementable using SQs (see, e.g., [CKL+06, FGR+17, FGV17]).

Returning to our agnostic learning setting, roughly speaking, *we show that a lower bound of $d$ on the degree of any $L^1$ approximating polynomial can be translated to an SQ lower bound of $n^{\Omega(d)}$ for the agnostic learning problem.* This lower bound is tight, since the $L^1$-regression algorithm can be implemented in the SQ model with complexity $n^{O(d)}$.

We note that a similar characterization had been previously shown, under somewhat different assumptions, for agnostic learning under the uniform distribution on the hypercube [DFT+15]. We explain the technical differences and similarities with our results in Section (11.1.1). It is worth pointing out that learning under the Gaussian distribution is generally believed to be computationally easier than learning under the uniform distribution on the hypercube in a number of settings. For example, prior work [ABL17, DKS18, DKTZ20c] has given "constant factor" agnostic learners for LTFs on $\mathbb{R}^n$ under the Gaussian distribution — i.e., algorithms with error $O(\mathrm{OPT}) + \epsilon$ — that run in $\mathrm{poly}(n/\epsilon)$ time. No polynomial time algorithm with such an error guarantee is known for any discrete distribution. At a high-level, known algorithms for these problems make essential use of the *anti-concentration* of the Gaussian distribution, which fails in the discrete setting. Similar algorithmic gaps exist for robustly learning low-degree PTFs and intersections of halfspaces [DKS18].

Our generic lower bound result for the Boolean case (Theorem 1.3.63) reduces the problem of proving explicit SQ lower bounds for agnostic learning to the structural question of proving lower bounds on the $L^1$ polynomial approximation degree (under the Gaussian measure). As our second contribution, we provide a toolkit to prove explicit degree lower bounds. As a corollary, we prove optimal or near-optimal SQ lower bounds for various natural classes, including LTFs, PTFs, and intersections of halfspaces.

Moving away from the Boolean-valued setting, an interesting direction is to understand the complexity of agnostic learning for real-valued function classes. In recent years, this broad question has been intensely investigated in learning theory, in part due to its connections to deep learning. Here we focus on agnostic learning under the $L^2$-loss.

**Definition 1.3.62** (Agnostic Learning Real-valued Functions with Gaussian Marginals)**.** *Let $\mathcal{C}$ be a class of real-valued concepts on $\mathbb{R}^n$. Given i.i.d. samples $(\mathbf{x}, y)$ from a distribution $D$ on $\mathbb{R}^n \times \mathbb{R}$, where the marginal $D_{\mathbf{x}}$ on $\mathbb{R}^n$ is the standard Gaussian $\mathcal{N}_n$ and no assumptions are made on the labels $y$, the goal is to output a hypothesis $h : \mathbb{R}^n \to \mathbb{R}$ such that with high probability we have $\mathbf{E}_{(\mathbf{x},y)\sim D}[(h(\mathbf{x}) - y)^2]^{1/2} \leq \mathrm{OPT} + \epsilon$, where $\mathrm{OPT} = \inf_{f\in\mathcal{C}} \mathbf{E}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - y)^2]^{1/2}$.*

A prototypical concept class of significant recent interest are Rectified Linear Units (ReLUs). A ReLU is any real-valued function $f : \mathbb{R}^n \to \mathbb{R}_+$ of the form $f(\mathbf{x}) = \mathrm{ReLU}(\langle \mathbf{w}, \mathbf{x} \rangle + \theta)$, $\mathbf{w} \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$, where $\mathrm{ReLU} : \mathbb{R} \to \mathbb{R}_+$ is defined as $\mathrm{ReLU}(u) = \max\{0, u\}$. ReLUs are the most commonly used activation functions in modern deep neural networks. The corresponding agnostic learning problem is a fundamental primitive in the theory of neural networks that has been extensively studied in recent years [GKKT17, MR18, GKK19, DGK+20, GGK20, DKZ20].

Our techniques extend to real-valued concepts leading to improved and nearly tight SQ lower bounds for natural concept classes. We describe our contributions in the following subsection.

**Contributions for Boolean-valued Concepts** Our main general result for Boolean-valued concepts is the following:

**Theorem 1.3.63** (Generic SQ Lower Bound, Boolean Case)**.** *Let $n, m \in \mathbb{Z}_+$ with $m \leq n^a$ for any constant $0 < a < 1/2$ and $\epsilon \geq n^{-c}$ for some sufficiently small constant $c > 0$. Fix a function $f : \mathbb{R}^m \to \{\pm 1\}$. Let $d$ be the smallest integer such that there exists a degree at most $d$ polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}_m}[|p(\mathbf{x}) - f(\mathbf{x})|] < 2\epsilon$. Let $\mathcal{C}$ be a class of Boolean-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any $\mathbf{P} \in \mathbb{R}^{m \times n}$ such that $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Any SQ algorithm that agnostically learns $\mathcal{C}$ under $\mathcal{N}_n$ to error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

The $L^1$-polynomial regression algorithm and Theorem 1.3.63 characterize the complexity of agnostic learning under the Gaussian distribution – within the class of SQ algorithms – for a range of concept classes. If $d$ is the (minimum) degree for which any function in $\mathcal{C}$ can be $\epsilon$-approximated by a degree-$d$ polynomial in $L^1$-norm, the complexity of agnostically learning $\mathcal{C}$ is, roughly, $n^{\Theta(d)}$.

***Applications of Theorem 1.3.63.*** Note that the above result does not tell us what the optimal degree $d$ is for any given concept class $\mathcal{C}$. Using analytic techniques, we establish explicit lower bounds on the $L^1$ polynomial approximation degree for three fundamental concept classes: Linear Threshold Functions (LTFs), Polynomial Threshold Functions, and Intersections of Halfspaces. As a corollary, we obtain explicit SQ lower bounds for these classes. Our applications are summarized in Table 1.4.

For the class of LTFs, using a known degree lower bound for the sign function [Gan02], we immediately obtain an SQ lower bound of $n^{\Omega(1/\epsilon^2)}$. This bound is optimal (within polynomial factors), improving on the previous SQ lower bound of $n^{\Omega(1/\epsilon)}$ [GGK20, DKZ20]. Our approach is simpler and more general compared to these prior works, immediately extending to other families. For the broader class of degree-$k$ PTFs, we

| **Concept Class** | Lower Bound | Upper Bound |
|---|---|---|
| LTFs | $\Omega\left(1/\epsilon^2\right)$ [Gan02] | $O\left(1/\epsilon^2\right)$ [Gan02, DKN10] |
| Degree-$k$ PTFs | $\Omega\left(k^2/\epsilon^2\right)$ (Thm 11.3.1) | $O\left(k^2/\epsilon^4\right)$ [Kan11] |
| Intersections of $k$ Halfspaces | $\tilde{\Omega}\left(\sqrt{\log k}/\epsilon\right)$ (Thm 11.3.5) | $O\left(\log k/\epsilon^4\right)$ [KOS08] |

Table 1.4: Bounds on the degree $d$ of $\epsilon$-approximating polynomials in $L^1$-error under the Gaussian measure. For each concept class, we obtain an SQ lower bound of $n^{\Omega(d)}$.

establish a degree lower bound of $\Omega(k^2/\epsilon^2)$ (Proposition 11.3.2), which yields an SQ lower bound of $n^{\Omega(k^2/\epsilon^2)}$ for the agnostic learning problem.

Our third explicit degree lower bound is for intersections of $k$ halfspaces. For this concept class, we prove a degree lower bound of $d = \tilde{\Omega}(\sqrt{\log k}/\epsilon)$ which implies a corresponding SQ lower bound of $n^{\tilde{\Omega}(\sqrt{\log k}/\epsilon)}$. In the process, we establish a new structural result translating lower bounds on the Gaussian Noise Sensitivity (GNS) of *any* Boolean function to the $L^1$-polynomial approximation degree of the same function.

Recall that the Gaussian Noise Sensitivity (GNS) of a function $f : \mathbb{R}^n \to \{\pm 1\}$ is defined as $\mathrm{GNS}_\rho(f) \stackrel{\text{def}}{=} \mathbf{Pr}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^\rho}[f(\mathbf{x}) \neq f(\mathbf{y})]$, where $\mathcal{N}_n^\rho$ is the distribution of a $(1-\rho)$-correlated Gaussian pair (i.e., $\mathbf{x}$ and $\mathbf{y}$ are standard Gaussians with correlation $(1-\rho)$). We show the following:

**Theorem 1.3.64** (Structural Result)**.** *Let $f : \mathbb{R}^n \to \{\pm 1\}$ and $p : \mathbb{R}^n \to \mathbb{R}$ be a degree at most $d$ polynomial. Then, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x})-p(\mathbf{x})|] \geq \Omega(1/\log(d))\mathrm{GNS}_{(\log(d)/d)^2}(f)$. Furthermore, for any $\epsilon > 0$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - p(\mathbf{x})|] \geq \mathrm{GNS}_\epsilon(f)/4 - O(d\sqrt{\epsilon})$.*

**Contributions for Real-valued Concepts** For agnostically learning real-valued concepts, we provide two generic lower bound results, analogous to Theorem 1.3.63, for Correlational SQ (CSQ) algorithms and general SQ algorithms respectively. A conceptual message of our results is that $L^2$ regression is essentially optimal against

CSQ algorithms, but not necessarily optimal against general SQ algorithms.

Recall that Correlational SQ (CSQ) algorithms are a subclass of SQ algorithms, where the algorithm is allowed to choose any bounded query function on the examples and obtain estimates of its correlation with the labels. (See Section (1.4.5) for a detailed description.) This class of algorithms is fairly broad, capturing many learning algorithms used in practice (including gradient-descent). For CSQ algorithms, we prove.

**Theorem 1.3.65** (Generic CSQ Lower Bound, Real-valued Case)**.** *Let $n, m \in \mathbb{Z}_+$ with $m \leq n^a$ for any constant $0 < a < 1/2$ and $\epsilon \geq n^{-c}$ for some sufficiently small constant $c > 0$. Let $f : \mathbb{R}^m \to \mathbb{R}$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f^2(\mathbf{x})] = 1$ and $d$ be the smallest integer such that there exists a degree at most $d$ polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\|f - p\|_2 < \epsilon$. Let $\mathcal{C}$ be a class of real-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any matrix $\mathbf{P} \in \mathbb{R}^{m\times n}$ satisfying $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Then, any CSQ algorithm that agnostically learns $\mathcal{C}$ over $\mathcal{N}_n$ to $L^2$-error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

Our lower bound for the general SQ model is presented below. The difference between the two is that the latter uses the $L^1$-norm to measure the approximation of $f$ by polynomials.

**Theorem 1.3.66** (Generic SQ Lower Bound, Real-valued Case)**.** *Let $n, m \in \mathbb{Z}_+$ with $m \leq n^a$ for any constant $0 < a < 1/2$ and $\epsilon \geq n^{-c}$ for some sufficiently small constant $c > 0$. Let $f : \mathbb{R}^m \to \mathbb{R}$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f^2(\mathbf{x})] = 1$ and $d$ be the smallest integer such that there exists a degree at most $d$ polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\|f - p\|_1 < \epsilon$. Let $\mathcal{C}$ be a class of real-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any matrix $\mathbf{P} \in \mathbb{R}^{m\times n}$ satisfying $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Then, any SQ algorithm*

*that agnostically learns $\mathcal{C}$ over $\mathcal{N}_n$ to $L^2$-error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

***Applications of Theorems 1.3.65 and 1.3.66.*** As in the Boolean-valued setting, obtaining explicit (C)SQ lower bounds for agnostically learning real-valued concepts requires analytic tools to establish lower bounds on the degree of polynomial approximations. In this part, we give such lower bounds for two fundamental concept classes: ReLUs and sigmoids. Establishing degree lower bounds for other non-linear activations is left as a question for future work.

| | $p = 1$ | |
|---|---|---|
| **Concept Class** | Lower Bound | Upper Bound |
| ReLUs | $\Omega(1/\epsilon)$ (Cor. 11.5.3) | $O(1/\epsilon)$ |
| Sigmoids | $\Omega(\log(1/\epsilon))$ (Thm. 11.5.8) | $O(\log^2(1/\epsilon))$ |

| | $p = 2$ | |
|---|---|---|
| **Concept Class** | Lower Bound | Upper Bound |
| ReLUs | $\Omega\left(1/\epsilon^{4/3}\right)$ (Cor. 11.5.3) | $O\left(1/\epsilon^{4/3}\right)$ |
| Sigmoids | $\Omega\left(\log^2(1/\epsilon)\right)$ (Cor. 11.5.5) | $O\left(\log^2(1/\epsilon)\right)$ |

Our degree lower bounds applications for both $L^1$ and $L^2$ polynomial approximations are summarized in Section (1.3.10). Combining these degree lower bounds Theorems 1.3.65 and 1.3.66 implies explicit SQ lower bounds for ReLUs and sigmoids.

Concretely, for agnostically learning ReLUs, we establish a CSQ lower bound of $n^{\Omega(1/\epsilon^{4/3})}$ (matching the $n^{O(1/\epsilon^{4/3})}$ upper bound obtained via $L^2$-regression); and an SQ lower bound of $n^{\Omega(1/\epsilon)}$, improving on the previous best bound of $n^{\Omega((1/\epsilon)^{1/36})}$ [GGK20, DKZ20].

**Application: Proving SQ Lower Bounds for the Massart Noise Model** Finally, we note that the power of our techniques is not limited to the agnostic setting. Moving away from the worst-case noise assumption, our framework can be adapted to analyze learning under more structured noise models, such as the semi-random models, i.e., Massart and General Massart (Section (1.3.1)). Our approach can be used to provide the lower bounds in Theorem 1.3.5.

## Overview of Techniques

**SQ Lower Bounds for Boolean-valued Functions** The starting point for our lower bounds is the work of [DKS17], which shows that if $D$ is a univariate distribution whose low-degree moments match those of a standard Gaussian (and which satisfies some other mild niceness conditions), then it is SQ-hard to distinguish between a standard multivariate Gaussian and a distribution that is a copy of $D$ in a random direction and a standard Gaussian in the orthogonal directions. (This is shown in [DKS17] for $D$ a 1-dimensional distribution, but it is not hard to generalize to higher dimensional distributions.)

Note that the above setting is unsupervised. To go from distributions to functions, we will try to produce a Boolean function $f$ of a few variables such that the distributions of $X$ conditioned on $f(X) = 1$ and on $f(X) = -1$ match moments with a Gaussian. We generalize the techniques of [DKKZ20] to show that such a function $f$ embedded in a hidden *low-dimensional subspace* is SQ hard to distinguish from a random function. Our goal then is to find such a function $f$ that is $(1/2 - \epsilon)$-close to a function in our family. Given this construction, learning the function to error $\mathrm{OPT} + \epsilon/2$ requires being able to distinguish $f$ from a random function.

The aforementioned approach was recently used by [DKZ20]. However, while that work constructs the function $f$ somewhat directly, here we take a more general approach. In more detail, it is not hard to phrase the conditions that (1) $f$ is bounded in $[-1, 1]$, (2) it matches moments with low-degree polynomials, and (3) is not too far from the function we are trying to learn, as an infinite-dimensional linear program (LP). We can then non-constructively attempt to find the optimal value of such an LP by duality. We note that "LP duality" in this setting is non-trivial – we require some (basic) functional analysis tools to show that duality applies for the LPs we are considering on function spaces. Given this, we find that the dual program is equivalent to finding a low-degree polynomial that approximates the function we are trying to learn in $L^1$-norm. The degree of such a polynomial conveniently matches the parameter that determines the runtime of the $L^1$- polynomial regression algorithm. We can thus show that, for reasonable function families, the $L^1$-regression algorithm is in fact optimal, among SQ algorithms, up to polynomial factors.

The above characterization allows us to determine the complexity of agnostically learning LTFs, by leverage tight degree lower bounds for the sign function. For the cases of degree-$k$ PTFs and intersections of $k$ halfspaces, we do not know what the correct answer is, but we are able to prove non-trivial, and qualitatively close to optimal, lower bounds.

We note that the $L^1$ approximation theory for these functions is more challenging than the $L^2$ approximation theory (which is entirely determined by the Fourier decay). To that end, we develop new techniques relating $L^1$ approximability to the Gaussian Noise sensitivity (Theorem 1.3.64), which allows us to prove the first non-trivial lower bounds. The proof of Theorem 1.3.64 works via a symmetrization technique. In

particular, let $\theta = \arccos(1-\epsilon)$ and let $X$ and $Y$ be standard Gaussians. Let $F_{X,Y}(\phi) := f(\sin(\phi)X + \cos(\phi)Y)$. Then we can write $\mathrm{GNS}_\epsilon(f) = \mathbf{Pr}[F_{X,Y}(\phi) \neq F_{X,Y}(\phi+\theta)]$. On the other hand, $\|f-p\|_1 = \mathbf{E}[|F_{X,Y}(\phi) - p(\sin(\phi)X + \cos(\phi)Y)|]$. Thus, it suffices to show that if $F$ is *any* Boolean function on the circle that the $L^1$ approximation error of $F$ by low degree polynomials can be bounded below by $\mathbf{Pr}[F(\phi) \neq F(\phi+\theta)]$. To show this, we use basic Fourier analysis to show that any low-degree polynomial with small $L^1$ norm cannot have any large higher derivatives. This implies that if $F$ transitions from being 0 to being 1 over some small interval, that any low-degree polynomial will not be able to match it very well in this interval.

**(C)SQ Lower Bounds for Real-valued Functions** We now move to real-valued functions and sketch our CSQ and SQ lower bounds. For CSQ lower bounds, we obtain a similar characterization. The difference is that, in the real-valued setting, we need to find a real-valued function $f$ whose low-degree moments vanish, and which is close to the function we are trying to learn *in* $L^2$ *norm*. This can be phrased as a similar LP and, applying duality, we find that the complexity is determined by the degree needed to approximate the function we are trying to learn in $L^2$ norm. For this particular setting, the LP can actually be solved explicitly and the best possible approximation function is obtained by taking the high-degree Hermite component of $f$. This lower bound matches (up to polynomial factors in the final error) the upper bound coming from the $L^2$ polynomial regression algorithm. This means that we can qualitatively characterize the complexity of agnostic learning using CSQ algorithms. In particular, we use this characterization to obtain new CSQ lower bounds on agnostically learning ReLUs and sigmoids.

Our SQ lower bounds against learning real-valued functions are somewhat more

challenging, since the approximating function $f$ must have more than just vanishing moments. It must have all its level-sets match low-degree moments with a standard Gaussian (which is equivalent only for Boolean-valued functions). Because of this additional requirement, we restrict our "imitating functions" to Boolean-valued functions. We can still find an LP defining $f$, however the dual gives us the relevant parameter of the degree needed to approximate the function we are trying to learn in $L^1$-norm (rather than $L^2$-norm) for which a matching upper bound is not known. So, in this case, while we can still obtain significantly improved SQ lower bounds for agnostically learning a number of concept classes, we do not obtain optimal results.

**SQ Lower Bounds: Theorems 2.1.3 and 2.1.6** The connection between low-degree moments and the correlation between $\mathbf{x}$ and $y$ is even more evident when trying to establish SQ lower bounds. Using the framework introduced in [DKS17], we can show the following: if there exists a Massart halfspace whose degree-$k$ moments have $\mathbf{x}$ independent of $y$, then distinguishing a random rotation of this distribution from one in which $\mathbf{x}$ is *truly* independent of $y$ is roughly $d^{\Omega(k)}$-hard in the SQ model. Thus, proving SQ lower bounds amounts to finding Massart halfspace distributions that fool low-order moments in this way.

We note that the moment-matching and noise conditions amount to a system of linear inequalities in terms of the noise function $\eta(\mathbf{x})$. To construct our hard examples, we use linear programming (LP) duality to show the existence of solutions to this system. It should be noted that LP duality has been previously used to provide SQ lower bounds for agnostic learning of halfspaces, see, e.g., [DFT+15, DKPZ21]. In the current setting, however, we have to construct instances that satisfy the much more restrictive (constant-bounded) Massart noise assumption.

To achieve this, we show that it is possible to add constant-bounded Massart noise to a classifier $f(\mathbf{x})$ in order to cause it to fool polynomials of degree at most $k$ if and only if there is no polynomial $p$ of degree at most $k$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[p(\mathbf{x})] = 0$, such that the sign of $p(\mathbf{x})$ agrees everywhere with the optimal halfspace $f(\mathbf{x})$. In order to prove that there exists no such low-degree, zero-mean, sign-matching polynomial, we establish a much more general result: the sign of any low-degree, zero-mean polynomial $p(\mathbf{x})$ cannot be too biased, and therefore $p(\mathbf{x})$ cannot match the sign of any heavily-biased function; see Lemma 11.7.10. In particular, if $f$ is a halfspace with bias $(1-\gamma)$, this is true for all zero-mean polynomials of degree $k$ up to (roughly) $\log(1/\gamma)$, implying the SQ lower bound of Theorem 2.1.3. We remark that our main structural lemma (Lemma 11.7.10) can readily be used to establish SQ lower bounds for Massart learning other geometric concept classes such as intersections of two homogeneous halfspaces.

It remains to provide a proof overview of Theorem 2.1.6 for learning homogeneous halfspaces with *general* Massart noise ($\eta = 1/2$). Interestingly, instead of adapting the methodology described in the preceding paragraphs, we prove Theorem 2.1.6 via a "reduction" to our SQ lower bound of Theorem 2.1.3, i.e., for learning (general) halfspaces with *constant-bounded* Massart noise. At a high-level, this is achieved by using the noise to "wash out" any information except when $\mathbf{x}$ lies in a thin strip not passing through the origin. Restricted to this strip, $f(\mathbf{x})$ is now a non-homogeneous halfspace, and our lower bound for general halfspaces applies.

## 1.4 Preliminaries

### 1.4.1 Miscellaneous Notation

**Basic Notation** For $n \in \mathbb{Z}_+$, let $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$. We use $\mathbb{1}_A$ to denote the characteristic function of the set $A$, i.e., $\mathbb{1}_A(\mathbf{x}) = 1$ if $\mathbf{x} \in A$ and $\mathbb{1}_A(\mathbf{x}) = 0$ if $\mathbf{x} \notin A$.

**Linear Algebra and Geometry Notation** We use small boldface characters for vectors and capital bold characters for matrices. Let $\mathbf{e}_i$ be the $i$-th standard basis vector in $\mathbb{R}^d$. For $\mathbf{x} \in \mathbb{R}^d$ and $i \in [d]$, $\mathbf{x}_i$ denotes the $i$-th coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2 \stackrel{\text{def}}{=} (\sum_{i=1}^d \mathbf{x}_i^2)^{1/2}$ denotes the $\ell_2$-norm of $\mathbf{x}$. We use $\mathbf{x} \cdot \mathbf{y}$ for the inner product of $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\theta(\mathbf{x}, \mathbf{y})$ for the angle between $\mathbf{x}, \mathbf{y}$. We slightly abuse notation and denote $\mathbf{e}_i$ the $i$-th standard basis vector in $\mathbb{R}^d$. For $\mathbf{x} \in \mathbb{R}^d$ and $V \subseteq \mathbb{R}^d$, $\mathbf{x}_V$ denotes the projection of $\mathbf{x}$ onto the subspace $V$. Note that in the special case where $V$ is spanned from one unit vector $\mathbf{v}$, then we simply write $\mathbf{x}_\mathbf{v}$ to denote $\mathbf{x} \cdot \mathbf{v}$, i.e., the projection of $\mathbf{x}$ onto $\mathbf{v}$. For a subspace $U \subset \mathbb{R}^d$, let $U^\perp$ be the orthogonal complement of $U$. For a vector $\mathbf{w} \in \mathbb{R}^d$, we use $\mathbf{w}^\perp$ to denote the subspace spanned by vectors orthogonal to $\mathbf{w}$, i.e., $\mathbf{w}^\perp = \{\mathbf{u} \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{u} = 0\}$. For unit vectors $\mathbf{u}, \mathbf{v}$, we use $\mathbf{u}^{\perp\mathbf{v}}$ to denote the component of $\mathbf{u}$ that is orthogonal to $\mathbf{v}$ i.e., $\mathbf{u}^{\perp\mathbf{v}} = (\mathbf{I} - \mathbf{v}\mathbf{v}^\intercal)\mathbf{u}$. We denote $\mathbf{v}^{\|\mathbf{u}}$ the projection of $\mathbf{v}$ into the direction $\mathbf{u}$, i.e., $\mathbf{v}^{\|\mathbf{u}} \stackrel{\text{def}}{=} ((\mathbf{v} \cdot \mathbf{u})\mathbf{u})/\|\mathbf{u}\|_2^2$. We use $\mathbb{B}(r)$ to denote the $\ell_2$ ball in $\mathbb{R}^d$ of radius $r$, centered at the origin. Finally, we use $\mathbb{S}^{d-1}$ to denote the unit sphere in $\mathbb{R}^d$ and $\mathbb{B}$ to denote the unit ball.

For a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\mathrm{tr}(\mathbf{A})$ denotes the trace of the matrix $\mathbf{A}$. For a square symmetric matrix $\mathbf{M}$, we say that $\mathbf{M}$ is positive semi-definite if only if all the eigenvalues of $\mathbf{M}$ are non-negative. For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the Frobenius norm of $\mathbf{A}$ is denoted by $\|\mathbf{A}\|_F = (\sum_{i=1}^m \sum_{j=1}^n A_{i,j}^2)^{1/2}$.

**Asymptotic notation** We use the standard $O(\cdot), \Theta(\cdot), \Omega(\cdot)$ asymptotic notation and $\widetilde{O}(\cdot)$ to omit polylogarithmic factors in the argument. We write $E \gtrsim F$ for two expressions $E$ and $F$ to denote that $E \geq c\,F$, where $c > 0$ is a sufficiently large universal constant (independent of the variables or parameters on which $E$ and $F$ depend). Similarly, we write $E \lesssim F$ to denote that $E \leq c\,F$, where $c > 0$ is a sufficiently small universal constant.

### 1.4.2 Probability Basics

**Probability Notation** We use $\mathbf{E}_{x\sim\mathcal{D}}[x]$ for the expectation of the random variable $x$ according to the distribution $\mathcal{D}$ and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$. For simplicity of notation, we may omit the distribution when it is clear from the context. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the $d$-dimensional Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, we denote $\phi_d(\cdot)$ the pdf of the $d$-dimensional Gaussian and we use the $\phi(\cdot)$ for the pdf of the standard normal. In this work we usually consider the standard normal, i.e., $\mu = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$, and therefore we denote it simply $\mathcal{N}$; the dimension will be always clear from the context. Moreover, we denote as $\mathcal{N}(\mu_1, \mu_2)$ the 2-dimensional Gaussian distribution with mean $(\mu_1, \mu_2)$ and $\mathcal{N}(\mu)$ the 1-dimensional Gaussian distribution with mean $\mu$. For $(\mathbf{x}, y)$ distributed according to $\mathcal{D}$, we denote $\mathcal{D}_{\mathbf{x}}$ to be the distribution of $\mathbf{x}$ and $\mathcal{D}_y$ to be the distribution of $y$. For unit vector $\mathbf{v} \in \mathbb{R}^d$, we denote $\mathcal{D}_{\mathbf{v}}$ the distribution of $\mathbf{x}$ on the direction $\mathbf{v}$, i.e., the distribution of $\mathbf{x}_{\mathbf{v}}$. For a set $B$ and a distribution $\mathcal{D}$, we denote $\mathcal{D}_B$ to be the distribution $D$ conditional on $B$. We define the standard $L_p$ norms with respect to the Gaussian measure, i.e., $\|g(\mathbf{x})\|_{L_p} = (\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|g(\mathbf{x})|^p)^{1/p}$. Sometimes, we use the notation $\|g(\mathbf{x})\|_p$ instead of $\|g(\mathbf{x})\|_{L_p}$, to distinguish them from vector norms we shall write $\|g(\mathbf{x})\|_p$ instead of $\|g\|_p$ when it is not clear from the context. We use

$\widehat{D}_N$ to denote the empirical distribution obtained by drawing $N$ i.i.d. samples from $D$.

**Error Metrics with respect to Distributions** We use $\mathrm{err}_D(\mathbf{w})$ to denote the 0-1 error of the halfspace defined by the weight vector $\mathbf{w}$ with respect to the distribution $D$, i.e., $\mathrm{err}_D(\mathbf{w}) \stackrel{\mathrm{def}}{=} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y]$. We will use $\mathrm{err}(\mathbf{w},\mathbf{x})$ for the 0-1 error of $\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})$ conditioned on $\mathbf{x}$, i.e., $\mathrm{err}(\mathbf{w},\mathbf{x}) := \mathbf{Pr}_{y\sim D_y(\mathbf{x})}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y]$. Note that $\mathrm{err}_D(\mathbf{w}) = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathrm{err}(\mathbf{w},\mathbf{x})]$. If $D$ satisfies the $\eta$-Massart noise condition with respect to the halfspace $\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})$, then $\mathrm{err}(\mathbf{w},\mathbf{x}) = \eta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) = \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\} + (1-\eta(\mathbf{x}))\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\}$ .

### 1.4.3 Gaussian Space and Hermite Polynomials

**Gaussian Space** We denote by $L_2(\mathcal{N})$ the vector space of all functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\|f\|_{L_2} < \infty$. $\mathrm{He}_i(z)$ denotes the *normalized* probabilist's Hermite polynomial of degree $i$. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the $d$-dimensional Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. In this work we usually consider the standard normal distribution, i.e., $\mu = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$, and thus denote it simply by $\mathcal{N}$. The usual inner product for this Gaussian space is $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})g(\mathbf{x})]$. We write $f(z) \doteq g(z)$ to mean that $\mathbf{E}_{z\sim\mathcal{N}(0,1)}[(f(z)-g(z))^2] = 0$. We use *normalized* probabilists' Hermite polynomial of degree $i$, defined via $\mathrm{He}_i(x) = \mathrm{he}_i(x)/\sqrt{i!}$, where by $\mathrm{he}_i(x)$ we denote the probabilist's Hermite polynomial of degree $i$:

$$\mathrm{he}_k(z) = (-1)^k \exp(z^2/2)\frac{\mathrm{d}^k}{\mathrm{d}z^k}\exp(-z^2/2).$$

These normalized Hermite polynomials form a complete orthonormal basis for the single-dimensional version of the inner product space defined above. Given a function $f : \mathbb{R} \to$

$\mathbb{R}$, $f \in L_2(\mathcal{N})$, we compute its Hermite coefficients as $\hat{f}(i) = \mathbf{E}_{z\sim\mathcal{N}(0,1)}[f(z)\mathrm{He}_i(z)]$, and express the function uniquely as $f(z) \doteq \sum_{i\geq 0} \hat{f}(i)\mathrm{He}_i(z)$. In some sections, we use the notation $H_i(z)$ instead of $\mathrm{He}_i(z)$. To get an orthonormal basis for $L_2(\mathcal{N})$ in the $d$-dimensional, we use a multi-index $\alpha \in \mathbb{N}^d$ to define the $d$-variate normalized Hermite polynomial as $\mathrm{He}_\alpha(\mathbf{x}) = \prod_{i=1}^d \mathrm{He}_{\alpha_i}(\mathbf{x}_i)$. The total degree of $\mathrm{He}_\alpha$ is $|\alpha| := \sum_i \alpha_i$. Given a function $f \in L_2(\mathcal{N})$, we compute its Hermite coefficients as $\hat{f}(\alpha) = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x})\mathrm{He}_\alpha(\mathbf{x})]$ and express it uniquely as $\sum_{\alpha\in\mathbb{N}^d} \hat{f}(\alpha)\mathrm{He}_\alpha(\mathbf{x})$.

We denote by $f^{[k]}(x)$ the degree $k$ part of the Hermite expansion of $f$, $f^{[k]}(\mathbf{x}) = \sum_{|J|=k} \hat{f}(J)\cdot \mathrm{He}_J(\mathbf{x})$. We say that a polynomial $q$ is harmonic of degree $k$ if it is a linear combination of degree $k$ Hermite polynomials, that is $q$ can be written as

$$q(\mathbf{x}) = q^{[k]}(\mathbf{x}) = \sum_{J:|J|=k} c_J \mathrm{He}_J(\mathbf{x}) \ .$$

For a single dimensional Hermite polynomial it holds $\mathrm{He}'_m(x) = \sqrt{m}\mathrm{He}'_{m-1}(x)$. Using this we obtain that for a multivariate Hermite polynomial $\mathrm{He}_M(\mathbf{x})$, where $M = (m_1, \ldots, m_n)$ it holds

$$\nabla \mathrm{He}_M(\mathbf{x}) = \sqrt{m_i} H_{M-E_i}(\mathbf{x}) \in \mathbb{R}^n, \tag{1.7}$$

where $E_i = \mathbf{e}_i$ is the multi-index that has 1 position $i$ and 0 elsewhere. From this fact and the orthogonality of Hermite polynomials we obtain

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[\nabla \mathrm{He}_M(\mathbf{x}) \cdot \nabla \mathrm{He}_L(\mathbf{x})] = |M|\delta_{M,L}. \tag{1.8}$$

The following fact gives us a formula for the inner product of

**Fact 1.4.1.** *Let $p, q : \mathbb{R}^n \to \mathbb{R}$ be harmonic polynomials of degree $k$. Then*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}\left[\nabla^\ell p(\mathbf{x}) \cdot \nabla^\ell q(\mathbf{x})\right] = k(k-1)\dots(k-\ell+1) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[p(\mathbf{x})q(\mathbf{x})] \,.$$

*In particular,*

$$\nabla^k p(\mathbf{x}) \cdot \nabla^k q(\mathbf{x}) = k! \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[p(\mathbf{x})q(\mathbf{x})] \,.$$

*Proof.* Write $p(\mathbf{x}) = \sum_{M:|M|=k} b_M \mathrm{He}_M(\mathbf{x})$ and $q(\mathbf{x}) = \sum_{M:|M|=k} c_M \mathrm{He}_M(\mathbf{x})$. Since the Hermite polynomials are orthonormal we obtain $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[p(\mathbf{x})q(\mathbf{x})] = \sum_{M:|M|=k} c_M b_M$. Now, using Equation (1.7) iteratively we obtain

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\nabla^\ell \mathrm{He}_M(\mathbf{x}) \cdot \nabla^\ell \mathrm{He}_L(\mathbf{x})\right] = k(k-1)\dots(k-\ell+1)\delta_{M,L}.$$

Using this equality we obtain

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\nabla^\ell p(\mathbf{x}) \cdot \nabla^\ell q(\mathbf{x})\right] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\sum_M b_M \nabla^\ell \mathrm{He}_M(\mathbf{x}) \cdot \sum_L c_L \nabla^\ell H_L(\mathbf{x})\right] \\
&= \sum_{M,L} b_M c_L \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\nabla^\ell \mathrm{He}_M(\mathbf{x}) \cdot \nabla^\ell \mathrm{He}_L(\mathbf{x})\right] \\
&= \sum_{M,L} b_M c_L k(k-1)\dots(k-\ell+1)\delta_{M,L} \\
&= k(k-1)\dots(k-\ell+1) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[p(\mathbf{x})q(\mathbf{x})].
\end{aligned}$$

□

Observe that for every harmonic polynomial $p(\mathbf{x})$ of degree $k$ we have that $\nabla^k p(\mathbf{x})$ is a symmetric tensor of order $k$. Since the degree of the polynomial is $k$ and we differentiate $k$ times this tensor no longer depends on $\mathbf{x}$. Using Fact 1.4.1, we observe

that this operation (modulo a division by $\sqrt{k!}$) preserves the $L^2$-norm of the harmonic polynomial $p$, that is $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[p^2(\mathbf{x})] = \left\|\nabla^k p(\mathbf{x})\right\|_2^2/k!$. For more details on the Gaussian space and Hermite analysis, we refer the reader to [O'D14]. Most of the facts about Hermite polynomials that we use in this work are well-known properties and can be found, for example, in [Sze67]. We are going to use the following simple fact about the gradient of Hermite polynomials; for a proof see, for example, Lemma 6 of [KTZ19].

**Fact 1.4.2.** *Let $f \in L_2(\mathcal{N})$. It holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(\nabla f(\mathbf{x})\cdot\mathbf{e}_i)^2] = \sum_{\alpha\in\mathbb{N}^d}\alpha_i\hat{f}(\alpha)^2$.*

### 1.4.4 Preliminaries: Multilinear Algebra

Here we introduce some multilinear algebra notation. An order $k$ tensor $\mathbf{A}$ is an element of the $k$-fold tensor product of subspaces $\mathbf{A} \in \mathcal{V}_1 \otimes \ldots \otimes \mathcal{V}_k$. We will be exclusively working with subspaces of $\mathbb{R}^d$ so a tensor $A$ can be represented by a sequence of coordinates, that is $A_{i_1,\ldots,i_k}$. The tensor product of a order $k$ tensor $\mathbf{A}$ and an order $m$ tensor $\mathbf{B}$ is an order $k+m$ tensor defined as $(\mathbf{A}\otimes\mathbf{B})_{i_1,\ldots,i_k,j_1,\ldots,j_m} = \mathbf{A}_{i_1,\ldots,i_k}\mathbf{B}_{j_1,\ldots,j_m}$. We are also going to use capital letters for multi-indices, that is tuples of indices $I = (i_1,\ldots,i_k)$. We denote by $E_i$ the multi-index that has 1 on its $i$-th co-ordinate and 0 elsewhere. For example the previous tensor product can be denoted as $\mathbf{A}_I\mathbf{B}_J$. To simplify notation we are also going to use Einstein's summation where we assume that we sum over repeated indices in a product of tensors. For example if $\mathbf{A}\in\mathbb{R}^d\otimes\mathbb{R}^d$, $\mathbf{v}\in\mathbb{R}^d$, $\mathbf{u}\in\mathbb{R}^d$ we have $\sum_{i,j=1}^d \mathbf{v}_i\mathbf{u}_j\mathbf{A}_{ij} = \mathbf{v}_i\mathbf{u}_j\mathbf{A}_{ij}$. We define the dot product of two tensors (of the same order) to be $\langle\mathbf{A},\mathbf{B}\rangle = \mathbf{A}_{i_1,\ldots,i_k}\mathbf{B}_{i_1,\ldots,i_k} = \mathbf{A}_I\mathbf{B}_I$. We also denote the $\ell_2$-norm of a tensor by $\|\mathbf{A}\|_2 = \sqrt{\mathbf{A}\cdot\mathbf{A}}$. We denote by $\mathbf{A}(\mathbf{X})$ a function that maps the tensor $\mathbf{X}$ to a tensor $\mathbf{A}(\mathbf{X})$. Let $\mathcal{V}$ be a vector space and let $\mathbf{A}(\mathbf{x}):\mathbb{R}^d\to\mathcal{V}^{\otimes k}$ be a tensor valued function. We denote by $\partial_i\mathbf{A}(\mathbf{x})$ the tensor of partial derivatives of $A(\mathbf{x})$,

$\partial_i \mathbf{A}(\mathbf{x}) = \partial_i \mathbf{A}_J(\mathbf{x})$ is a tensor of order $k+1$ in $\mathcal{V}^{\otimes k} \otimes \mathbb{R}^d$. We also denote this tensor $\nabla \mathbf{A}(\mathbf{x}) = \partial_i \mathbf{A}_J(\mathbf{x})$. Similarly we define higher-order derivatives, and we denote

$$\nabla^m \mathbf{A}(\mathbf{x}) = \partial_{i_1} \dots \partial_{i_m} \mathbf{A}_J(\mathbf{x}) \in \mathcal{V}^{\otimes k} \otimes (\mathbb{R}^d)^{\otimes m} .$$

Let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)$ be a $d$-dimensional multi-index vector, where for all $i \in [d]$, $\alpha_i$ is non-negative integer. We denote $|\alpha| = \sum_{i=1}^d \alpha_i$ and for a $d$-dimensional vector $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d)$, we denote $\mathbf{w}^\alpha = \prod_{i=1}^d \mathbf{w}_i^{\alpha_i}$. We also denote the Frobenius norm of a tensor by $\|\mathbf{A}\|_F = \sqrt{\mathbf{A} \cdot \mathbf{A}}$. We will be using the (appropriately) normalized Hermite tensors (Definition 2.2.6). For these tensors we have the following fact.

**Fact 1.4.3.** *Let $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d$ and let $h_m : \mathbb{R} \mapsto \mathbb{R}$ be the univariate normalized (with respect to the standard normal) degree-$m$ Hermite polynomial and $\mathbf{H}^m$ be the Hermite moment tensor of Definition 2.2.6. It holds $h_m(\mathbf{x} \cdot \mathbf{v}) = \mathbf{H}^m \cdot \mathbf{v}^{\otimes m}$.*

### 1.4.5 Statistical Query Framework

To formally reason about the computational limits of learning algorithms, we require a framework that captures a broad class of algorithmic techniques while being amenable to rigorous analysis. The Statistical Query (SQ) model provides such a framework. At a high level, SQ algorithms are restricted from inspecting individual examples and can only access statistical properties of the underlying data distribution. By proving that no efficient SQ algorithm can solve a given learning problem, we provide strong evidence that the problem is computationally hard for a wide range of practical algorithms, including all known variants of gradient descent.

In this section, we formally define the SQ model and its more restricted variant, the

Correlational SQ (CSQ) model. We then introduce the key combinatorial notion of SQ dimension, which allows us to translate the problem of proving computational lower bounds into one of constructing sets of "hard-to-distinguish" distributions.

### Statistical Query Model

We begin by formalizing the notion of a statistical query oracle, following the standard formulation in prior work.

**Definition 1.4.4** (STAT Oracle)**.** *Let $D$ be a distribution on labeled examples supported on $X \times [-1, 1]$, for some domain $X$. A statistical query is a function $q : X \times [-1, 1] \to [-1, 1]$. We define $STAT(\tau)$ to be the oracle that given any such query $q(\cdot, \cdot)$ outputs a value $v$ such that $|v - \mathbf{E}_{(\mathbf{x},y)\sim D}[q(\mathbf{x}, y)]| \leq \tau$, where $\tau > 0$ is the tolerance parameter of the query.*

**Statistical Query Dimension** To study lower bounds in this model, we adopt the general approach of reducing learning tasks to decision problems over distributions. We begin with the following formal definitions, adapted from [FGR+17].

**Definition 1.4.5** (Decision Problem over Distributions)**.** *Let $D$ be a fixed distribution and $\mathcal{D}$ be a distribution family. We denote by $\mathcal{B}(\mathcal{D}, D)$ the decision (or hypothesis testing) problem in which the input distribution $D'$ is promised to satisfy either (a) $D' = D$ or (b) $D' \in \mathcal{D}$, and the goal is to distinguish between the two cases.*

**Definition 1.4.6** (Pairwise Correlation)**.** *The pairwise correlation of two distributions with probability density functions $D_1, D_2 : \mathbb{R}^n \to \mathbb{R}_+$ with respect to a distribution with density $D : \mathbb{R}^n \to \mathbb{R}_+$, where the support of $D$ contains the supports of $D_1$ and $D_2$, is defined as $\chi_D(D_1, D_2) \stackrel{\text{def}}{=} \int_{\mathbb{R}^n} D_1(\mathbf{x}) D_2(\mathbf{x}) / D(\mathbf{x})\, d\mathbf{x} - 1$.*

A low pairwise correlation implies that the distributions are statistically close. We can now define the SQ dimension based on finding a large set of distributions that are nearly uncorrelated.

**Definition 1.4.7.** *We say that a set of $s$ distributions $\mathcal{D} = \{D_1, \ldots, D_s\}$ over $\mathbb{R}^n$ is $(\gamma, \beta)$-correlated relative to a distribution $D$ if $|\chi_D(D_i, D_j)| \leq \gamma$ for all $i \neq j$, and $|\chi_D(D_i, D_j)| \leq \beta$ for $i = j$.*

**Definition 1.4.8** (Statistical Query Dimension)**.** *For $\beta, \gamma > 0$, a decision problem $\mathcal{B}(\mathcal{D}, D)$, where $D$ is a fixed distribution and $\mathcal{D}$ is a family of distributions, let $s$ be the maximum integer such that there exists a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ such that $\mathcal{D}_D$ is $(\gamma, \beta)$-correlated relative to $D$ and $|\mathcal{D}_D| \geq s$. The* Statistical Query dimension *with pairwise correlations $(\gamma, \beta)$ of $\mathcal{B}$ is defined to be $s$, and denoted by $\mathrm{SD}(\mathcal{B}, \gamma, \beta)$.*

The following lemma establishes the formal connection between the SQ dimension of a problem and the query complexity required to solve it. A large SQ dimension implies that any successful SQ algorithm must either make many queries or use a very small tolerance.

**Lemma 1.4.9** (Corollary 3.12 of [FGR+17])**.** *Let $\mathcal{B}(\mathcal{D}, D)$ be a decision problem, where $D$ is the reference distribution and $\mathcal{D}$ is a class of distributions. For $\gamma, \beta > 0$, let $s = \mathrm{SD}(\mathcal{B}, \gamma, \beta)$. For any $\gamma' > 0$, any SQ algorithm for $\mathcal{B}$ requires queries of tolerance at most $\sqrt{\gamma + \gamma'}$ or makes at least $s\gamma'/(\beta - \gamma)$ queries.*

The following fact shows that given a one-dimensional marginal $A$ that matches $m$ moments with the standard normal, the correlation between two hidden direction distributions with directions $\mathbf{v}$ and $\mathbf{u}$ is bounded above roughly by the $(m+1)$-th power of the correlation of their corresponding directions. Therefore, using Fact 11.2.6, we

obtain that there exists an exponentially large (in the dimension $d$) set of distributions with pairwise correlation roughly $d^{-m}$.

**Fact 1.4.10** (Lemma 3.4 from [DKS17]). *Let $m \in \mathbb{Z}_+$. If the univariate distribution $A$ over $\mathbb{R}$ agrees with the first $m$ moments of $\mathcal{N}(0,1)$, then for all $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$, we have that*

$$|\chi_{\mathcal{N}(\mathbf{0},\mathbf{I})}(\mathbf{P}^A_{\mathbf{v}}, \mathbf{P}^A_{\mathbf{u}})| \leq |\mathbf{v} \cdot \mathbf{u}|^{m+1} \chi^2(A, \mathcal{N}(0,1)) \; .$$

### Correlational Statistical Query (CSQ) Model

For some of our lower bounds in the real-valued setting, we consider *correlational* or inner product queries. The CSQ model is a restriction of the SQ model, where the algorithm is allowed to choose any bounded query function, and obtain estimates for its correlation with the labels.

Specifically, for $f, h : X \to \mathbb{R}$ and a distribution $D_{\mathbf{x}}$ over the domain $X$, we denote by $\langle f, h \rangle_{D_{\mathbf{x}}}$ the quantity $\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[f(\mathbf{x})h(\mathbf{x})]$ and refer to it as the correlation of $f$ and $h$ under $D_{\mathbf{x}}$. While it is commonly assumed that the query function $h$ is pointwise bounded, it is in fact sufficient to assume that it has bounded $L^2$-norm. If $D$ is the joint distribution on points and labels, a correlational query takes $h$ and a parameter $t > 0$, and outputs a value $v \in [\mathbf{E}_{(\mathbf{x},y)\sim D}[h(\mathbf{x})y] - \tau, \mathbf{E}_{(\mathbf{x},y)\sim D}[h(\mathbf{x})y] + \tau]$.
Similarly to the general SQ model, we consider the following notions of statistical dimension.

**Definition 1.4.11** (Correlational Statistical Query Dimension). *For $\beta, \gamma > 0$, a probability distribution $D_{\mathbf{x}}$ over domain $X$ and a family $\mathcal{C}$ of functions $f : X \to \mathbb{R}$, let $s$ be the maximum integer for which there exists a finite set of functions $\{f_1, \ldots, f_s\} \subseteq \mathcal{C}$ such that $|\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[f_i^2(\mathbf{x})]| \leq \beta$ for all $i \in [s]$, and $|\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[f_i(\mathbf{x}) f_j(\mathbf{x})]| \leq \gamma$ for all $i, j \in [s]$*

*with $i \neq j$. We define the* Correlational Statistical Query Dimension *with pairwise correlations $(\gamma, \beta)$ of $\mathcal{C}$ to be $s$ and denote it by* $\mathrm{CSD}_{D_{\mathbf{x}}}(\mathcal{C}, \gamma, \beta)$.

**Definition 1.4.12** (Average Correlational Statistical Query Dimension)**.** *Let $\rho > 0$, let $D_{\mathbf{x}}$ be a probability distribution over some domain $X$, and let $\mathcal{C}$ be a family of functions $f : X \to \mathbb{R}$. We define the average pairwise correlation of functions in $\mathcal{C}$ to be $\rho(\mathcal{C}) = \frac{1}{|\mathcal{C}|^2} \sum_{g,r \in \mathcal{C}} |\mathbf{E}_{\mathbf{x} \sim D_{\mathbf{x}}}[g(\mathbf{x})r(\mathbf{x})]|$. The* Average Correlational Statistical Query Dimension *of $\mathcal{C}$ relative to $D_{\mathbf{x}}$ with parameter $\gamma$, denoted by* $\mathrm{CSDA}_{D_{\mathbf{x}}}(\mathcal{C}, \gamma)$*, is defined to be the largest integer $s$ such that every subset $\mathcal{C}' \subseteq \mathcal{C}$ of size $|\mathcal{C}'| \geq |\mathcal{C}|/s$, satisfies $\rho(\mathcal{C}') \geq \rho$.*

In most of the cases, it suffices to bound the correlational statistical query dimension, since by simple calculations this implies a bound on the average statistical query dimension.

**Lemma 1.4.13.** *Let $\mathcal{C}$ be a class of functions and $D_{\mathbf{x}}$ be a distribution and suppose that $\mathrm{CSD}_{D_{\mathbf{x}}}(\mathcal{C}, \gamma, \beta) = d$, for some $\gamma, \beta > 0$. Then, for all $\gamma' > 0$, we have that $\mathrm{CSD}_{D_{\mathbf{x}}}(\mathcal{C}, \gamma + \gamma') \geq d\gamma'/(\beta - \gamma)$.*

The following result [Szö09, GGJ+20] relates the Average Correlational SQ dimension of a concept class with the complexity of any CSQ algorithm for the class.

**Lemma 1.4.14** (Theorem B.1 in [GGJ+20])**.** *Let $D_{\mathbf{x}}$ be a distribution over a domain $X$ and let $\mathcal{C}$ be a real-valued concept class over $X$ such that $0 \in \mathcal{C}$, and $\|f\|_2 \geq \eta$ for all $f \in \mathcal{C}, f \not\equiv 0$. Suppose that for some $\gamma > 0$ we have $s = \mathrm{CSDA}_{D_{\mathbf{x}}}(\mathcal{C}, \gamma)$. Any CSQ algorithm that outputs a hypothesis $h$ such that $\|h - f\|_2 < \eta$ needs at least $s/2$ queries or queries of tolerance $\sqrt{\gamma}$.*

# Part I

# Learning Halfspaces with Noise

# Chapter 2

# Learning Halfspaces with Semi-Random Noise

## 2.1 Introduction

This chapter characterizes the computational complexity of learning general (non-homogeneous) halfspaces under (General) Massart noise when the underlying data distribution is Gaussian. We present the first efficient learning algorithms and nearly-matching computational lower bounds for this problem, revealing a surprising quasi-polynomial dependence on the halfspace's bias and the excess error, answering Questions 1.3.2 and 1.3.3.

Our investigation reveals a dichotomy in the problem's complexity. For constant-bounded Massart noise (i.e., $\eta \leq 1/2 - c$), we establish that the complexity is dictated by the bias, $\gamma$, of the target halfspace, providing an algorithm with runtime $d^{O(\log(1/\gamma))}$ (Theorem 2.1.2) and a matching SQ lower bound (Theorem 2.1.3). For the more challenging general Massart noise case (i.e., $\eta = 1/2$), we find the complexity is governed

by the excess error $\epsilon$, for which we provide an algorithm with runtime $d^{O(\log(1/\epsilon))}$ (Theorem 2.1.5) and a corresponding lower bound (Theorem 2.1.6).

**Learning with Constant-Bounded Massart Noise** Here we address the regime of learning general halfspaces with $\eta$-Massart noise, where $\eta \leq 1/2 - c$ for some constant $c > 0$. It turns out that the complexity of the problem in this setting depends on the "bias" of the target halfspace.

**Definition 2.1.1** ($(1-\gamma)$-biased Halfspace)**.** *For $\gamma \in [0, 1/2]$, we say that the halfspace $f$ is at most $(1-\gamma)$-biased (with respect to $\mathcal{D}$) if* $\min(\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) = +1], \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) = -1]) \geq \gamma$.

Note that homogeneous halfspaces correspond to the special case of $\gamma = 1/2$. Recall that prior literature gave $\mathrm{poly}(d/\epsilon)$-time learners for homogeneous Massart halfspaces with $\eta < 1/2$. Our first main result is a learning algorithm for general halfspaces whose complexity scales quasi-polynomially with $1/\gamma$. More specifically, we establish the following theorem (see also Theorem 2.3.1):

**Theorem 2.1.2** (Learning General Massart Halfspaces with Constant-Bounded Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal that satisfies the constant-bounded Massart noise condition with respect to an at most $(1-\gamma)$-biased halfspace. There exists an algorithm that draws $N = d^{O(\log(\min(1/\gamma, 1/\epsilon)))}\mathrm{poly}(1/\epsilon)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and computes a halfspace $h$ such that with high probability it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$.*

Qualitatively, Theorem 2.1.2 yields a $\mathrm{poly}(d/\epsilon)$-time algorithm for halfspaces with bias bounded away from zero, specifically $\gamma \geq c$ where $c > 0$ is an absolute constant.

Furthermore, for general halfspaces it yields a PTAS with quasi-polynomial dependence on $1/\epsilon$, i.e., with runtime $d^{O(\log(1/\epsilon))}$.

Perhaps surprisingly, we provide strong evidence that the upper bound of Theorem 2.1.2 is best possible for any value of $\gamma$. Specifically, we prove a matching lower bound in the Statistical Query (SQ) model of [Kea98]. We recall that SQ algorithms are a broad class of algorithms that are only allowed to query expectations of bounded functions of the distribution rather than directly access samples; see Section (1.4.5) for the definition and additional discussion. Formally, we prove (see also Theorem 11.7.1):

**Theorem 2.1.3** (SQ Lower Bound for Massart Halfspaces with Constant-Bounded Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal that satisfies the Massart noise condition with parameter $\eta = 1/2 - \Omega(1) < 1/2$ with respect to an at most $(1-\gamma)$-biased halfspace. For any $\gamma > \epsilon$, any SQ algorithm that for any such distribution $\mathcal{D}$ learns a hypothesis $h : \mathbb{R}^d \mapsto \{\pm 1\}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, either requires queries with tolerance at most $d^{-\Omega(\log(1/\gamma))}$ or makes at least $2^{d^{\Omega(1)}}$ statistical queries.*

Informally, Theorem 2.1.3 shows that no SQ algorithm can learn the subclass of $(1-\gamma)$-biased halfspaces in the constant-bounded Massart noise model with sub-exponential in $d^{\Omega(1)}$ many queries, unless it uses queries of very small tolerance – that would require at least $d^{\Omega(\log(1/\gamma))}$ samples to simulate (as long as $\gamma > \epsilon$[1]). This "fine-grained" lower bound that can be viewed as an information-computation tradeoff for the problem (within the class of SQ algorithms) and matches the upper bound of Theorem 2.1.2. As a corollary, we obtain that learning general halfspaces with constant-bounded Massart noise has SQ complexity $d^{\Theta(\log(1/\epsilon))}$.

[1] Of course, if $\gamma \leq \epsilon$, one of the two constant functions suffices.

Theorems 2.1.2 and 2.1.3 together qualitatively characterize the complexity of learning general halfspaces in the constant-bounded Massart setting. We view this inherent quasi-polynomial dependence as rather surprising. Even though the class of general halfspaces has one additional parameter compared to the homogeneous case (the unknown threshold), the learning problem becomes harder and exhibits a quasi-polynomial dependence on the inverse of the bias parameter $\gamma$.

**Remark 2.1.4.** While the algorithm of Theorem 2.1.2 is essentially optimal in the constant-bounded Massart regime, its running time has quasi-polynomial dependence on $1/\beta$, where $\beta := 1 - 2\eta$. As a result, the algorithm is not optimal (as far as we know) when the parameter $\eta$ is *very* close to $1/2$, e.g., $\eta = 1/2 - 1/d$. Characterizing the complexity of the learning problem when $\eta$ is very close to (but not equal to) $1/2$ is left as an open problem for future work.

**Learning with *General* Massart Noise** Our second main result essentially characterizes the complexity of learning halfspaces with general Massart noise, i.e., the $\eta = 1/2$ case. On the positive end, we develop an algorithm for this general setting with the following guarantees.

**Theorem 2.1.5** (Learning General Halfspaces with General Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal and such that $\mathcal{D}$ satisfies the Massart noise condition for $\eta = 1/2$ with respect to a general target halfspace. There is an algorithm that draws $N = d^{O(\log(1/\epsilon))}$ samples from $\mathcal{D}$, runs in time $2^{\mathrm{poly}(1/\epsilon)}\mathrm{poly}(N, d)$, and computes a halfspace $h$ such that with high probability $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$.*

We reiterate that no nontrivial algorithm was known in the general Massart model,

even for homogeneous halfspaces (our algorithm works for general halfspaces). Qualitatively, Theorem 2.1.5 gives a PTAS with runtime $O_\epsilon(1)d^{O(\log(1/\epsilon))}$ for the general problem. Note that if $\epsilon \geq 1/\log^{\Omega(1)}(d)$, the runtime of our algorithm is $d^{O(\log(1/\epsilon))}$.

It is worth comparing Theorem 2.1.5 with the algorithm of Theorem 2.1.2, which handles the case where $\eta < c$ for some constant $c < 1/2$. The latter result yields a $\mathrm{poly}(d/\epsilon)$ time algorithm when the bias of the target halfspace is a positive constant. On the other hand, the algorithm of Theorem 2.1.5 has a quasi-polynomial dependence in $1/\epsilon$, even for the subclass of homogeneous halfspaces.

Perhaps surprisingly, we establish an SQ lower bound suggesting that this quasi-polynomial dependence is necessary in the general Massart model, even for homogeneous halfspaces.

**Theorem 2.1.6** (SQ Lower Bound for Homogeneous Halfspaces with General Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal and such that $\mathcal{D}$ satisfies the Massart noise condition for $\eta = 1/2$ with respect to a homogeneous halfspace. Then any SQ algorithm that for any such distribution $\mathcal{D}$ outputs a hypothesis $h$ with $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, either requires queries with tolerance at most $d^{-\Omega(\log(1/\epsilon))}$ or makes at least $2^{d^{\Omega(1)}}$ statistical queries.*

(For a more detailed statement, see Theorem 11.7.13.) Informally, Theorem 2.1.6 shows that no SQ algorithm can learn homogeneous halfspaces in the general Massart model with sub-exponential in $d^{\Omega(1)}$ many queries, unless it uses queries of very small tolerance – that would require at least $d^{\Omega(\log(1/\epsilon))}$ samples to simulate. Note that the sample complexity of the algorithm establishing Theorem 2.1.5 (which can be implemented in the SQ model) matches our SQ lower bound. Furthermore, this implies that the runtime of our algorithm is essentially optimal (within SQ algorithms) for all

$\epsilon \geq 1/\log^{\Omega(1)}(d)$.

**Remark 2.1.7.** It is worth noting that Theorems 2.1.5 and 2.1.6 provably separate the SQ complexity of learning halfspaces with general Massart noise from the SQ complexity of the corresponding agnostic learning problem. For agnostically learning halfspaces under Gaussian marginals, the $L_1$-regression algorithm [KKMS08] has complexity $d^{O(1/\epsilon^2)}$ (and is known to be implementable in SQ). Moreover, a matching SQ lower bound of $d^{\Omega(1/\epsilon^2)}$ is known [DKPZ21]. That is, while agnostic learning requires runtime $d^{\mathrm{poly}(1/\epsilon)}$, Massart learning can be achieved in time $O_\epsilon(1)\, d^{O(\log(1/\epsilon))}$.

**Remark 2.1.8.** Interestingly, the Massart model remains meaningful even for the "very large" noise setting, where $\eta > 1/2$. For this extreme regime, it is not hard to show (see Section A.1) that this model becomes *equivalent* to the agnostic model.

**Organization** The structure of this work is as follows: In Section 2.2, we provide a detailed overview of our techniques. In Section 2.3, we give our algorithm for learning general halfspaces with constant-bounded Massart noise, establishing Theorem 2.1.2. In Section 2.4, we give our algorithm for learning halfspaces under general Massart noise, establishing Theorem 2.1.5. Finally, in Section 11.7, we prove our SQ lower bounds, Theorem 2.1.3 and Theorem 2.1.6.

## 2.2 Detailed Technical Overview

### 2.2.1 Learning General Halfspaces with Constant-Bounded Massart Noise: Theorem 2.1.2

Our learning algorithm for this regime leverages the certificate framework developed in [DKTZ20b, DKK+20, DKK+21b]. This framework makes essential use of the Massart noise condition, i.e., the fact that $\eta(\mathbf{x}) \leq 1/2$ for every $\mathbf{x} \in \mathbb{R}^d$. Using this fact, we have the following characterization of the affine function $\ell^*(\mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x} - t^*$ defining the target halfspace $f(\mathbf{x}) = \mathrm{sign}(\ell^*(\mathbf{x}))$. For every non-negative function $T(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}^+$, it holds that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell^*(\mathbf{x}) y\, T(\mathbf{x})] = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell^*(\mathbf{x})\mathrm{sign}(\ell^*(\mathbf{x}))(1-2\eta(\mathbf{x}))\, T(\mathbf{x})] \geq 0\,.$$

On the other hand, when $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] \geq \mathrm{OPT} + \epsilon$, there exists a non-negative function $T(\mathbf{x})$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x}) y\, T(\mathbf{x})] < 0$. We say that such a $T$ is a "certifying" function (or simply a certificate) for the guess $\ell(\mathbf{x})$, because it proves that $\ell(\mathbf{x})$ is not optimal.

**Definition 2.2.1** (Certificate Oracle)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal satisfying the Massart noise condition with respect to some halfspace. Fix $\epsilon, \delta \in (0,1]$. Given any affine function $\ell(\mathbf{x})$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] \geq \mathrm{OPT} + \epsilon$, a $\rho$-certificate oracle returns a non-negative function $T : \mathbb{R}^d \mapsto \mathbb{R}_+$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x}) y\, T(\mathbf{x})] \leq -\rho\|\ell(\mathbf{x})\|_2$.*

In [DKTZ20b, DKK+20, DKK+21b], it was shown that under Massart label noise, the problem of efficiently learning an optimal halfspace can be reduced to the problem

of efficiently computing certifying functions. At a high-level, the certifying functions can be viewed either as separation oracles for a convex program (that can be solved via the ellipsoid method) or as loss functions in an online convex optimization problem (that can be solved via online gradient descent). In our setting, we need to adapt the proofs of [DKTZ20b] slightly to work for general halfspaces. For completeness, we provide the details of this reduction in Subsection 2.3.3.

**Proposition 2.2.2.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal. Assume that $\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to some halfspace. Fix $\epsilon, \delta \in (0,1)$. Given a $\rho$-certificate oracle $\mathcal{O}$ with runtime $T_{\mathcal{O}}(\rho)$, there exists an algorithm that makes $M = \mathrm{poly}(\frac{d}{\epsilon\rho})$ calls to $\mathcal{O}$, draws $N = \mathrm{poly}(\frac{d}{\epsilon\rho})\log(1/\delta)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(d, N, M)T_{\mathcal{O}}(\rho)$ time and computes a hypothesis $h$, such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, with probability $1-\delta$.*

Given Proposition 2.2.2, it remains to construct an efficient certificate oracle. The first step is to restrict our search over some *parametric class* of non-negative functions.

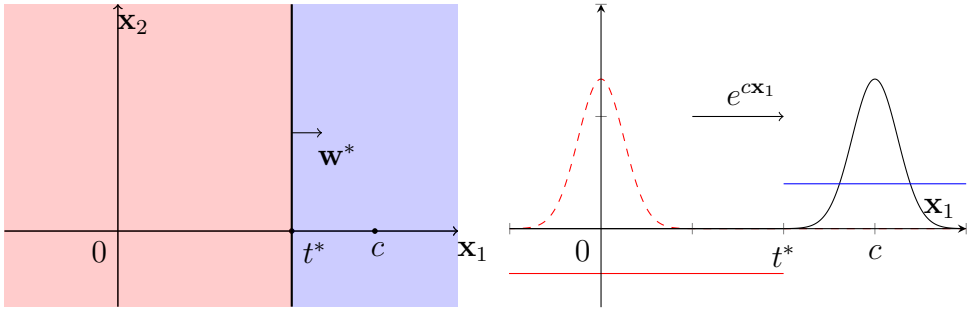
**Non-Continuous Certificates** When we do not restrict our search to continuous functions, we can use certificates of the form $T(\mathbf{x}) = \mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq \mathrm{sign}(\mathbf{v}\cdot\mathbf{x} - b)\}$, for some $\mathbf{v} \in \mathbb{R}^d, b \in \mathbb{R}$. It is not hard to see (Section A.3) that, for any $\eta \in [0, 1/2]$, taking $T(\mathbf{x}) = \mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})\}$, i.e., the indicator of the disagreement region of $\mathrm{sign}(\ell(\mathbf{x}))$ and $f(\mathbf{x})$, we obtain that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y\,T(\mathbf{x})] \leq -\Omega(\epsilon^2)\|\ell(\mathbf{x})\|_2$. Observe that, when we consider non-continuous certificates, there exist $\mathrm{poly}(\epsilon)$-certificates independent of both the noise level $\eta$ and the bias of the halfspace $\gamma$. However, finding such certificates is computationally hard in general. In [DKK+20], the authors provided a polynomial-time oracle for non-continuous certificates, i.e., subsets of the disagreement region, under

two crucial assumptions. Specifically, they assumed that (1) the target halfspace is homogeneous and (2) the distribution $\mathcal{D}$ satisfies the Tsybakov noise condition, which is significantly weaker than the 1/2-Massart noise assumption. In particular, the technique of [DKK+20] to isolate a subset of the disagreement region does not work when we relax either of the above assumptions. In what follows, we describe an efficient certificate oracle for general halfspaces under the assumption that $\eta = 1/2 - \Omega(1)$, i.e., in the constant-bounded Massart noise regime.

**The Exponential Shift Certificate** In order to get a handle on the optimization problem of finding a certificate, the first step is to consider smooth function classes. We will show how to construct a certificate function against the constant guess $-1$, when the true halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} - t^*)$ is $(1-\gamma)$-biased: the probability of the negative region is very large $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x}) = -1] = 1 - \gamma$; see Figure 2.1a. Finding certifying functions against the constant guess $-1$ captures many of the challenges of the general case. The idea is to find a continuous function $T(\mathbf{x})$ that puts more weight on the disagreement region of $f(\mathbf{x})$ (colored blue in Figure 2.1a) than the agreement region (colored red in Figure 2.1a). In particular, in order to find a certificate against the constant guess $-1$, we want to find some function $T(\mathbf{x})$ so that the following ratio is a sufficiently large constant (greater than $1/\beta$):

$$\frac{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[T(\mathbf{x})\mathbb{1}\{f(\mathbf{x}) = +1\}]}{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[T(\mathbf{x})\mathbb{1}\{f(\mathbf{x}) = -1\}]} \tag{2.1}$$

Our key idea is to use an exponential-shift certificate of the form $T(\mathbf{x}) = e^{\mathbf{v}\cdot\mathbf{x}}$. By multiplying the Gaussian density with the exponential function $T(\mathbf{x})$ we essentially shift the mean of the Gaussian from $\mathbf{0}$ to $\mathbf{v}$. Setting $\mathbf{v}$ to be a large multiple of $\mathbf{w}^*$,




(a) (b)

Figure 2.1: **(a)** An instance where $f(\mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x} - t^*$ is very biased and the constant hypothesis $\ell(\mathbf{x}) = -1$ agrees with $f(\mathbf{x})$ "almost everywhere", i.e., in the red region. A non-negative certifying function against the hypothesis $-1$ must put significantly more weight to the disagreement region (colored in blue). Notice that by Gaussian concentration we have that $t^* = O(\sqrt{\log(1/\gamma)})$.
**(b)** The certifying function $T(\mathbf{x}) = e^{c\ (\mathbf{w}^* \cdot \mathbf{x})}$ essentially moves the mean of the Gaussian from $\mathbf{0}$ to $c\mathbf{w}^*$. Choosing $c = \Theta(\sqrt{\log(1/\gamma)})$ implies that the mass of the blue region with respect to the shifted normal $\mathcal{N}(c\mathbf{w}^*, \mathbf{I})$ will be much larger than the mass of the red region, making Equation (2.1) true.

we can re-center the Gaussian of the ratio of Equation (2.1) to lie well within the disagreement (blue) region; see Figure 2.1b. In order to make the ratio of Equation (2.1) sufficiently large, it suffices to set $\mathbf{v} = c\ \mathbf{w}^*$ for $c = \Theta(\sqrt{\log(1/\gamma)})$; see also the proof of Lemma 2.3.5.

**The Polynomial-Shift Certificate** Even though we have shown that there exists an exponential-shift certificate, it is still not easy to compute: minimizing $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})ye^{\mathbf{v}\cdot\mathbf{x}}]$ over $\mathbf{v} \in \mathbb{R}^d$ is a non-convex objective that, in general, is hard to optimize. In order to circumvent this obstacle, we consider polynomial certificates of the form $q^2(\mathbf{x})$ for some low-degree polynomial $q$. Such certificates were also used in [DKTZ20b] in order to learn homogeneous halfspaces with Tsybakov noise. The advantage of having $T(\mathbf{x}) = q^2(\mathbf{x})$, for some low-degree polynomial $q$, is that we can efficiently compute certificates via

semi-definite programming (SDP); see Subsection 2.3.2.

As with the exponential-shift certificate, we can provide a high-level overview of some of the ideas involved by trying to construct a certificate against the constant hypothesis $-1$, when the optimal halfspace is almost everywhere equal to $-1$, i.e., $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x}) = -1] = 1 - \gamma$. Our goal in this case is to find a polynomial that assigns much more weight to the positive region (colored blue in Figure 2.1a) than the negative region (colored red in Figure 2.1a). The main structural result that enables our algorithm is the following lemma, where we show that we can achieve the same effect of the exponential-shift certificate discussed above using a low-degree polynomial (see Lemma 2.3.5 for a more detailed statement and proof).

**Lemma 2.2.3** (Low-Degree Polynomial Shift)**.** *Let $f(\mathbf{x})$ be an at most $(1-\gamma)$-biased halfspace. There exists an absolute constant $C \geq 1$ such that*

$$e^{k/C - C\log(1/\gamma)} \leq \max_{p:\deg(p)\leq k} \frac{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[p^2(\mathbf{x})\mathbb{1}\{f(\mathbf{x}) = +1\}]}{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[p^2(\mathbf{x})\mathbb{1}\{f(\mathbf{x}) = -1\}]} \leq e^{C\ k\log k - \log(1/\gamma)/C}\,.$$

The lower bound of Lemma 2.2.3 implies that there exists a polynomial certificate, namely a polynomial that makes Equation (2.1) true, of degree $O(\log(1/\gamma))$. The upper bound implies that degree $\Omega(\log(1/\gamma))$ is essentially necessary.

It is not hard to prove the upper bound using the anti-concentration of Gaussian polynomials; see Lemma 2.3.8. To prove the lower bound of Lemma 2.2.3, we construct a low-degree polynomial using the Taylor approximation to the exponential-shift certificate that we defined previously. For $c = \Theta(\sqrt{\log(1/\gamma)})$, we consider the *square of the Taylor expansion* $S_k(cx)$ of the function $e^{cx}$. Using the Gaussian concentration and the fast convergence rate of the Taylor polynomial of $e^x$, in Claim 2.3.7, we show that for

degree $k = \Theta(c^2) = \Theta(\log(1/\gamma))$, $S_k^2(c\ \mathbf{w}^* \cdot \mathbf{x})$ is very close to $e^{2c\ (\mathbf{w}^* \cdot \mathbf{x})}$ in the $L_1$ sense. Therefore, we can use $T(\mathbf{x}) = S_k^2(c\ \mathbf{w}^* \cdot \mathbf{x})$ in the ratio of Equation (2.1) and obtain the same guarantees (up to constant factors) with the exponential shift $e^{2c\ (\mathbf{w}^* \cdot \mathbf{x})}$ that we discussed previously.

Our certificate against general (non-constant) hypotheses is the product of (the square of) a polynomial and a band around the current guess $\ell(\mathbf{x})$, i.e., $T(\mathbf{x}) = \mathbb{1}\{r_1 \leq \ell(\mathbf{x}) \leq r_2\} q^2(\mathbf{x})$. Notice that, since the current guess $\ell(\mathbf{x})$ is known to the certificate algorithm, in order to find a band of the form $\mathbb{1}\{r_1 \leq \ell(\mathbf{x}) \leq r_2\}$, we simply need to perform a brute-force search over the two thresholds $r_1, r_2$. Our technical contribution here is the following proposition, showing the existence of low-degree polynomial certificates for general halfspaces with constant-bounded Massart noise (see Subsection 2.3.1 for a more detailed statement and proof).

**Proposition 2.2.4** (Polynomial Certificate for Halfspaces with Constant-Bounded Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal. Assume that $\mathcal{D}$ satisfies the constant-bounded Massart noise condition with respect to some at most $(1-\gamma)$-biased target halfspace $f(\mathbf{x})$. Let $\ell(\mathbf{x})$ be any linear function such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] \geq \mathrm{OPT} + \epsilon$. There exist $r_1, r_2 \in \mathbb{R}$ and polynomial $q(\mathbf{x})$ of degree $k = \Theta(\log(1/\gamma))$ with $\|q(\mathbf{x})\|_2 = 1$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\ell(\mathbf{x}) y\ \mathbb{1}\{r_1 \leq \ell(\mathbf{x}) \leq r_2\} q^2(\mathbf{x})\right] \leq -\epsilon^2 \mathrm{poly}(\gamma)\ \|\ell(\mathbf{x})\|_2 \,.$$

### 2.2.2 Learning Halfspaces with General Massart Noise: Theorem 2.1.5

For the purpose of this description, we will assume that the target halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$ is homogeneous. This special case captures the key ideas of our algorithm; given such a result, the generalization to general halfspaces (under general Massart noise) is fairly straightforward. (This generalization is carried out in Section A.4.) To facilitate the intuition, we present the main ideas behind the algorithm of Theorem 2.1.5 in the following paragraphs. Our main technical contribution in this context is the construction of a low-degree sign-matching polynomial. This is presented at the end of this subsection and in full detail in Subsection 2.4.1.

At a high-level, our algorithm for learning halfspaces in the $\eta = 1/2$ regime consists of a random walk on the unit $d$-dimensional sphere, where we iteratively perform a random step in order to update our current guess $\mathbf{w}^{(t)}$, i.e.,

$$\mathbf{w}^{(t+1)} \leftarrow \frac{\mathbf{w}^{(t)} + \lambda \mathbf{v}}{\|\mathbf{w}^{(t)} + \lambda \mathbf{v}\|_2} \,. \tag{2.2}$$

Our goal is to find a way to sample the update vector $\mathbf{v}$, so that at every step there is some non-trivial probability that we make progress towards the optimal direction $\mathbf{w}^*$. In order to make progress towards $\mathbf{w}^*$, it suffices to have an update vector $\mathbf{v}$ that correlates with $\mathbf{w}^*$ and belongs in the orthogonal complement, $\mathbf{w}^\perp$, of $\mathbf{w}$. In what follows, we will denote by $(\mathbf{w}^*)^{\perp \mathbf{w}}$ the normalized projection of $\mathbf{w}^*$ onto the subspace $\mathbf{w}^\perp$; see Figure 2.2a. Given such a vector $\mathbf{v}$, we can show that there exists a step size $\lambda$ such that the update rule of Equation (2.2) moves $\mathbf{w}^{(t)}$ closer to $\mathbf{w}^*$ by a non-trivial amount with constant probability; see Lemma 2.4.14. Observe that a uniformly random unit direction

in $\mathbb{R}^d$ has roughly $1/\sqrt{d}$ correlation with $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ with constant probability. However, in order to hit $\mathbf{w}^*$, we need to perform roughly $d$ consecutive successful updates (see Lemma 2.4.14), resulting in an algorithm with $2^{d^{\Omega(1)}}$ runtime.

The main algorithmic result of this section is the following proposition, which shows that we can efficiently sample an update vector $\mathbf{v}$ that improves the current guess with non-trivial probability.

**Proposition 2.2.5** (Correlated Update Oracle)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta = 1/2$, with respect to a target halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Let $\mathbf{w} \in \mathbb{R}^d$ be a unit vector such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y] \geq \mathrm{OPT} + \epsilon$, for some $\epsilon \in (0, 1]$. There exists an algorithm that draws $N = d^{O(\log(1/\epsilon))} \log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and with probability at least $1 - \delta$ returns a distribution $\mathcal{V}$ on $\mathbb{R}^d$ such that*

$$\mathop{\mathbf{Pr}}_{\mathbf{v}\sim\mathcal{V}}\left[(\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{v} \geq \mathrm{poly}(\epsilon)\right] \geq \frac{1}{3}\,.$$

*Moreover, $\mathcal{V}$ has description size $\mathrm{poly}(d/\epsilon)$ and can be sampled in $\mathrm{poly}(d/\epsilon)$ time.*

(See Proposition 2.4.2 for a more detailed statement.) Our plan is to construct a subspace $V$ of $\mathbb{R}^d$ such that $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp_{\mathbf{w}}})\|_2 \geq \mathrm{poly}(\epsilon)$. To sample good update vectors $\mathbf{v}$, as claimed in Proposition 2.2.5, we can generate a random vector $\mathbf{v}$ on the unit sphere of $V$. However, we need to make sure that the dimension of $V$ is sufficiently small, namely at most $\mathrm{poly}(1/\epsilon)$.

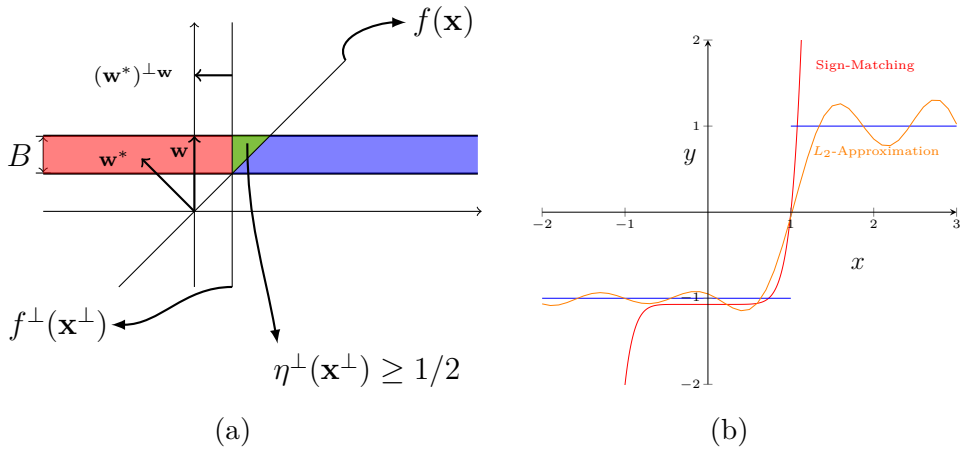
**Improving the Constant Guess** Let us assume for now that our current guess is $\mathbf{w} = \mathbf{0}$. Then, in order to make progress towards $\mathbf{w}^*$, one can simply use the degree-one Chow parameters of $y$, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}]$. Observe that the degree-one Chow parameters

have positive correlation with the optimal direction $\mathbf{w}^*$, since $(\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}]) \cdot \mathbf{w}^* = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(1-2\eta(\mathbf{x}))\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\ (\mathbf{x}\cdot\mathbf{w}^*)] > 0$. Therefore, the degree-one Chow parameters are a good first update to the guess $\mathbf{w} = \mathbf{0}$.

**Projecting onto $\mathbf{w}^\perp$: Subsection 2.4.2** In order to further improve a non-trivial guess $\mathbf{w}$, we need to find a good update direction $\mathbf{v}$ that correlates non-trivially with $(\mathbf{w}^*)^{\perp_\mathbf{w}}$. A natural attempt to do so would be to project $\mathbf{x}$ onto the orthogonal complement of the current guess, i.e., $\mathbf{w}^\perp$, and then compute the Chow parameters of the projected points. However, by doing so, the optimal classifier of the projected examples will no longer be a halfspace; see Figure 2.2a. In particular, the noise function $\eta^\perp(\mathbf{x}^\perp)$ after projection will be larger than $1/2$ for a large fraction of the points. To make the dataset nearly separable by a halfspace with normal vector $(\mathbf{w}^*)^{\perp_\mathbf{w}}$, we condition on a thin band and then project $\mathbf{x}$ onto $\mathbf{w}^\perp$. Doing so, apart from a small region close to the optimal classifier, $f^\perp$ (see Figure 2.2a), the noise function will be at most $1/2$. By making the band sufficiently thin, we can control the probability of this "high-noise" area. Notice that in the projected instance the optimal halfspace is no longer homogeneous. When the band (that we condition on) is far from the origin, the resulting optimal halfspace $f^\perp$ of the projected instance will be potentially very biased; see Figure 2.2a.

Assuming that our current halfspace $\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})$ is at least $\epsilon$ suboptimal compared to $\mathbf{w}^*$, we show that there exists a thin band conditional on which the current hypothesis is roughly $\epsilon/\sqrt{\log(1/\epsilon)}$-suboptimal. Moreover, this band is not very far from the origin, which implies that the optimal halfspace $f^\perp$ conditional on the band will not be very biased: its threshold will be at most $O(\sqrt{\log(1/\epsilon)})$. It is worth noting that a similar orthogonal projection step onto $\mathbf{w}^\perp$ was used in [DKK+20] to learn homogeneous halfspaces with Tsybakov noise. The major difference between the setting of the current

work and [DKK+20] is that in the general Massart regime it is not possible to control the distance of the band from the origin. Specifically, it could be the case that $\eta(\mathbf{x}) = 1/2$ for all $\mathbf{x}$ close to $\mathbf{w} \cdot \mathbf{x} = 0$, forcing us to pick a band whose optimal halfspace $f^\perp$ in the subspace $\mathbf{w}^\perp$ actually has threshold $\Omega(\sqrt{\log(1/\epsilon)})$, see Figure 2.2a. In contrast, in [DKK+20], the "soft" Tsybakov noise condition allows for the band to be picked arbitrarily close to origin resulting in nearly homogeneous halfspaces $f^\perp$. For the details of this projection step, see Lemma 2.4.6. In what follows, we denote the distribution of the projected instance over $\mathbf{w}^\perp \times \{\pm 1\}$ by $\mathcal{D}^\perp$. We elaborate further on this step in Subsection 2.4.2.




(a) (b)

Figure 2.2: **(a)** After we condition on the band $B$, we project $\mathbf{x}$ to the subspace $(\mathbf{w}^*)^{\perp \mathbf{w}}$ and nearly maintain the Massart noise property with respect to the biased halfspace $f^\perp(\mathbf{x}^\perp) = \mathrm{sign}((\mathbf{w}^*)^{\perp \mathbf{w}} \cdot \mathbf{x}^\perp + b)$. In particular, it holds that $\eta^\perp(\mathbf{x}^\perp) \leq 1/2$ everywhere apart from a small area (green). Since the underlying distribution is the standard Gaussian, a band with large negative mass (blue) cannot be very far from the origin, and therefore we have that $|b| = O(\sqrt{\log(1/\epsilon)})$.
**(b)** The sign-matching polynomial that corresponds to the red curve does not need to closely approximate the threshold function. Its degree scales as $\Theta(b^2)$, which is at most $O(\log(1/\epsilon))$ for an $\epsilon$-biased halfspace. On the other hand, to get an $L_2$ or $L_1$ approximation to error $\epsilon$ (orange curve), it is known that $\mathrm{poly}(1/\epsilon)$ degree is necessary.

**Using the Low-Order Chow Tensors: Subsection 2.4.3** To obtain a good update vector $\mathbf{v}$, a natural approach is to use the degree-one Chow parameters. However, since the optimal halfspace $f^\perp$ is biased, the degree-one Chow parameters are not guaranteed to correlate well with the direction of $f^\perp$. We thus need to look at higher-order Chow parameter tensors of $\mathcal{D}^\perp$. Recall that, to show Proposition 2.2.5, we want to construct a subspace $V$ of $\mathbb{R}^d$ such that $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp_\mathbf{w}})\|_2 \geq \mathrm{poly}(\epsilon)$. We now show how to find such a subspace using the low-order Chow tensors of $\mathcal{D}^\perp$.

Since we are working in Gaussian space, instead of considering the moment-tensor $\mathbf{x}^{\otimes m}$, we use the degree-$m$ Hermite moment tensor $\mathbf{H}^m(\mathbf{x})$; this corresponds to replacing all the monomials of the tensor $\mathbf{x}^{\otimes m}$ by their corresponding Hermite monomials. (For tensor notation, we refer to Section 1.4.)

**Definition 2.2.6** (Hermite Moment-Tensor)**.** *Let $\mathcal{H}$ be the linear operator that maps any $d$-variate monomial to the corresponding (normalized) $d$-variate polynomial in the Hermite basis, i.e., $\mathcal{H}(\mathbf{x}^\alpha) = h_\alpha(\mathbf{x})$. We define the degree-$m$ Hermite moment tensor as $(\mathbf{H}^m)_\alpha = \mathcal{H}((\mathbf{x}^{\otimes m})_\alpha)$.*

Using the Hermite moment-tensors, we also define the order-$m$ Chow parameter tensors of a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{\pm 1\}$.

**Definition 2.2.7** (Order-$m$ Chow Tensor of $\mathcal{D}$)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal distribution. We define the order-$m$ Chow tensor of $\mathcal{D}$ to be*

$$\mathbf{T}^m(\mathcal{D}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{H}^m(\mathbf{x})y]\,.$$

*When it is clear from the context, we shall omit the distribution $\mathcal{D}$ and simply write $\mathbf{T}^m$.*

Our high-level plan is to treat the above tensors as $d \times d^{m-1}$ matrices and perform

SVD to find the top few left singular vectors, i.e., the singular vectors whose singular values are larger than some threshold. To show that the subspace $V$ spanned by such eigenvectors contains a non-trivial part of $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$, we need to construct a *mean-zero polynomial* that only depends on the direction $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ and correlates well with the label $y$ (Lemma 2.4.10). This leads us to our main structural result.

**The Sign-Matching Polynomial: Subsection 2.4.1** We show that there exists a mean-zero polynomial $p$ that achieves non-trivial correlation with $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$, i.e., $\mathbf{E}_{(\mathbf{x}^\perp,y)\sim\mathcal{D}^\perp}[p(\mathbf{x})y] \geq \mathrm{poly}(\epsilon)$. Even though the noise of $\mathcal{D}^\perp$ is not exactly Massart – recall that there exists a small region where $\eta^\perp(\mathbf{x}^\perp) > 1/2$ – let us assume that $\eta^\perp(\mathbf{x}^\perp) \leq 1/2$ everywhere for simplicity. In this case, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x}^\perp,y)\sim\mathcal{D}^\perp}[p(\mathbf{x})y] = \mathop{\mathbf{E}}_{(\mathbf{x}^\perp,y)\sim\mathcal{D}^\perp_{\mathbf{x}}}\left[p(\mathbf{x})\mathrm{sign}((\mathbf{w}^*)^{\perp_{\mathbf{w}}}\cdot\mathbf{x}^\perp + b)(1-2\eta^\perp(\mathbf{x}^\perp))\right] .$$

*Since* $1-2\eta^\perp(\mathbf{x}^\perp) \geq 0$*, in order to achieve non-trivial correlation it suffices to find* $p$ *such that* $p(\mathbf{x})$ *matches the sign of* $f^\perp(\mathbf{x}^\perp)$*.*

At this point, we made crucial use of the Massart noise condition; in particular, this is not possible in the agnostic model. In the agnostic model, to achieve non-trivial correlation, one needs to actually approximate the threshold function; see Figure 2.2b. More specifically, to guarantee positive correlation in the agnostic model, we need a polynomial whose $L_1$ error with $f(\mathbf{x})$ is $O(\epsilon)$. Unfortunately, this cannot be done for any polynomial with degree $o(1/\epsilon^2)$; see, e.g., Proposition 2.1 of [DKPZ21]. In the Massart noise setting, we show that we can construct a *zero-mean sign-matching* polynomial of degree only $\log(1/\epsilon)$ that achieves $\mathrm{poly}(\epsilon)$ correlation with $y$. We remark that the mean-zero condition, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{x})] = 0$ is crucial here. Non-zero mean polynomials,

like the linear polynomial $(\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{x}^{\perp} + b$, might give constant correlation, but do not reveal any information about the optimal direction. More concretely, we establish the following proposition; see Proposition 2.4.3 and Lemma 2.4.4 for the corresponding formal statements.

**(Informal) Proposition 2.2.8** (Sign-Matching Polynomial)**.** *Let $b \in \mathbb{R}$. There exists a zero mean, unit variance polynomial $p : \mathbb{R} \mapsto \mathbb{R}$ of degree $k = \Theta(b^2+1)$ such that the sign of $p$ matches the sign of the threshold function* $\mathrm{sign}(z-b)$*, i.e.,* $\mathrm{sign}(p(z)) = \mathrm{sign}(z-b)$*, for all $z \in \mathbb{R}$.*

Notice that when we apply the above proposition, the threshold $b$ of the corresponding halfspace $f^{\perp}$ will be at most $O(\sqrt{\log(1/\epsilon)})$ resulting in a polynomial of degree $O(\log(1/\epsilon))$; see also Figure 2.2a. In the case of homogeneous halfspaces, this happens because we use the orthogonal projection step; see Subsection 2.4.2. In the case of general halfspaces, we may have such thresholds to start with.

### 2.2.3 SQ Lower Bounds: Theorems 2.1.3 and 2.1.6

**Learning General Halfspaces with Constant-Bounded Massart Noise: Theorem 2.1.3** Our SQ lower bounds make essential use of the "hidden-direction" framework developed in [DKS17]. Using this framework, we construct SQ lower bounds for learning halfspaces in high dimensions using a carefully constructed one-dimensional Massart noise instance. In particular, if we can construct a one-dimensional Massart noise distribution $\mathcal{D}$ on $\mathbb{R} \times \{\pm 1\}$ such that $\mathbf{E}_{(z,y)\sim\mathcal{D}}[z^k y] = \mathbf{E}_{z\sim\mathcal{D}_z}[z^i]\,\mathbf{E}_{y\sim\mathcal{D}_y}[y]$, for all $i \in [k]$, then we can obtain a family of $2^{d^{\Omega(1)}}$ distributions on $\mathbb{R}^d \times \{\pm 1\}$ whose pairwise correlation is $d^{-\Omega(k)}$. Using standard SQ lower bound arguments, having such a family

of pairwise correlated distributions implies a $d^{\Omega(k)}$ SQ lower bound for learning the halfspace.

Our main technical result is the following proposition (also see Proposition 11.7.3).

**Proposition 2.2.9.** *Fix $\eta \in (0, 1/2)$ such that $\eta = 1/2 - \Omega(1)$ and $\gamma \in (0, 1/2)$. There exists a distribution $\mathcal{D}$ on $(z, y) \in \mathbb{R} \times \{\pm 1\}$ whose $z$-marginal is the standard normal distribution with the following properties.*

- *$\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to a halfspace $f(z)$ with $\mathbf{Pr}_{z\sim\mathcal{D}_z}[f(z) = +1] = \gamma$.*
- *There exists an absolute constant $C$ such that for any integer $k \leq C \log(1/\gamma)$, it holds $\mathbf{E}_{(z,y)\sim\mathcal{D}}[yz^k] = \mathbf{E}_{y\sim\mathcal{D}_y}[y]\,\mathbf{E}_{z\sim\mathcal{D}_z}[z^k]$.*

To prove the existence of the (constant-bounded) Massart noise instance of Proposition 2.2.9, we use (infinite-dimensional) LP duality, where our variable is the signal function $\beta(z) = 1 - 2\eta(z)$. We have the following pair of primal and dual linear programs. We denote $\mathcal{P}_k^0$ the linear space of zero-mean polynomials of degree at most $k$.

| **Primal** | **Dual** |
|---|---|
| Find $\beta(z) \in L^\infty(\mathbb{R})$ | Find $p(z) \in \mathcal{P}_k^0$ |
| s.t. $\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)p(z)\beta(z)] = 0 \ \forall p \in \mathcal{P}_k^0$ | s.t. $\beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^+] > \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^-]$ |
| $\mathop{\mathbf{Pr}}_{z\sim\mathcal{N}}[\beta \leq \beta(z) \leq 1] = 1$ | |

Using an infinite-dimensional variant of the theorem of the alternative for linear programming, to show that the primal problem is feasible, it suffices to show that the

dual is infeasible. To show that the dual is infeasible, we prove a stronger statement: mean-zero polynomials of low-degree cannot match the sign of very biased Boolean functions. We prove the following (see Lemma 11.7.10 for the formal statement and proof).

**Lemma 2.2.10.** *Let $f : \mathbb{R} \mapsto \{\pm 1\}$ be any one-dimensional Boolean function, $\beta \in (0, 1)$ and $k \in \mathbb{Z}_+$. There exists a universal constant $C > 0$ such that if $\mathbf{Pr}_{z\sim\mathcal{N}}[f(z) = 1] \leq 2^{-Ck}(1-\beta)$, then for any mean-zero polynomial of degree at most $k$ it holds $\beta\,\mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^+] < \mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^-]$.*

We remark that in the above lemma we do not require $f$ to be a halfspace. Consequently, it is easy to adapt our argument to work for other Boolean concept classes that depend on a low-dimensional subspace $V$. For example, using the above lemma, we can obtain an SQ lower bound for learning intersections of 2 homogeneous halfspaces with constant-bounded Massart noise.

**Learning Homogeneous Halfspaces with General Massart Noise: Theorem 2.1.6** Our main idea to show an SQ lower bound for learning homogeneous halfspaces with general Massart noise is to use the fact that by setting $\eta(\mathbf{x}) = 1/2$, we essentially remove all the useful signal from some parts of the space. Our "reduction" to learning halfspaces with constant-bounded Massart noise works as follows: We create a $(d+1)$-dimensional high-noise instance, i.e., with $\eta = 1/2$, over $((\mathbf{x}, z), y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$ from many $d$-dimensional constant-bounded Massart noise instances. For every $(\mathbf{x}, z)$ outside of a thin slice (see Figure 2.3), we set $\eta(\mathbf{x}) = 1/2$. For $(\mathbf{x}, z)$ in the slice, we set the conditional distribution of $(\mathbf{x}, y)$ on $z$ to satisfy the constant-bounded Massart noise condition with respect to some non-homogeneous optimal halfspace (denoted

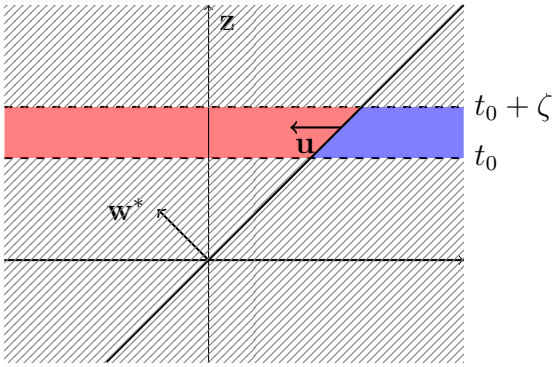



Figure 2.3: The "high-noise" ($\eta = 1/2$) Massart distribution $\mathcal{D}$ on $((\mathbf{x}, z), y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$ that we construct in our "reduction" from learning a biased halfspace with Massart noise. The $\mathbf{x}$-marginal of $\mathcal{D}$ is the standard normal distribution. Conditional on $z$, $(x, y)$ has constant-bounded Massart noise when $z \in [t_0, t_0 + \zeta]$; this corresponds to the blue/red area. When $z \notin [t_0, t_0 + \zeta]$ we set $y$ to be $\pm 1$ with probability $1/2$ independently of $\mathbf{x}$, i.e., $\eta((\mathbf{x}, z)) = 1/2$ in the gray area.

by $\mathbf{u}$ in Figure 2.3). We show that any hypothesis that achieves error OPT $+ \epsilon$ in the $(d+1)$-dimensional "high-noise" instance will perform well (on average) on the $d$-dimensional constant-bounded Massart noise instances — a problem that we have already showed to be hard in the SQ model. We refer to Subsection 11.7.2 for the detailed proof.

**Technical Preliminaries** We collect some technical preliminaries that we will use throughout this chapter.

**Gaussian Measure** We use the following fact providing upper and lower bounds on Gaussian tails and the Gaussian anticoncentation property for intervals.

**Fact 2.2.11** (Gaussian Density Properties)**.** *Let $\mathcal{N}$ be the standard one-dimensional normal distribution. Then, the following properties hold:*

1. *For any $t > 0$, it holds $e^{-t^2}/4 \leq \mathbf{Pr}_{z \sim \mathcal{N}}[z > t] \leq e^{-t^2/2}/2$.*

*2. For any $a, b \in \mathbb{R}$ with $a \leq b$, it holds $\mathbf{Pr}_{z\sim\mathcal{N}}[a \leq z \leq b] \leq (b-a)/\sqrt{2\pi}$.*

**Function Families (LTFs and Polynomials)** We use $\mathcal{C}_V$ to denote the set of Linear Threshold Functions (LTFs) with normal vector contained in $V \subseteq \mathbb{R}^d$, i.e., $\mathcal{C}_V = \{\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t) : \mathbf{v} \in V, \|\mathbf{v}\|_2 = 1, t \in \mathbb{R}\}$; when $V = \mathbb{R}^d$, we simply write $\mathcal{C}$. Moreover, we define $\mathcal{C}_0$ to be the set of homogeneous LTFs, i.e., $\mathcal{C}_0 = \{\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}) : \mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\|_2 = 1\}$. We denote by $\mathcal{P}_{k,d}$ the space of polynomials on $\mathbb{R}^d$ of degree at most $k$. We will use the following fact shows that for the probability that two halfspaces disagree can be upper bounded by the difference of their thresholds plus the angle of their corresponding normal vectors.

**Fact 2.2.12** (see, e.g., Fact 3.5 of [DKK+21a])**.** *Let $\mathcal{N}$ be the standard normal distribution in $\mathbb{R}^d$. Let $\mathbf{v}, \mathbf{u}$ be unit vectors in $\mathbb{R}^d$ and $t_1, t_2 \in \mathbb{R}$. It holds $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[\mathrm{sign}(\mathbf{u}\cdot\mathbf{x}+t_1) \neq \mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t_2)] \leq O(\theta(\mathbf{u},\mathbf{v})) + O(|t_1 - t_2|)$.*

## 2.3 A Certificate-Based Algorithm For Learning General Halfspaces with Constant-Bounded Massart Noise

In this section we show that, when the noise upper bound $\eta$ is bounded away from $1/2$, we can learn at most $(1-\gamma)$-biased halfspaces with sample complexity and runtime $d^{O(\log(1/\gamma))}\mathrm{poly}(1/\epsilon)$. Our algorithmic result matches our SQ lower bound, given in Theorem 2.1.3. We first state the formal version of Theorem 2.1.2.

**Theorem 2.3.1** (Learning $(1-\gamma)$-biased Halfspaces with Constant-Bounded Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard Gaussian.*

*Assume that $\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to some $(1-\gamma)$-biased optimal halfspace and $\beta = 1-2\eta$. Let $\epsilon, \delta \in (0,1]$. There exists an algorithm that draws $N = d^{O(\log(1/(\gamma\beta)))}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N,d)$, and computes a halfspace $h \in \mathcal{C}$ such that with probability at least $1-\delta$ we have that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$.*

The main technical ingredient in our proof is Proposition 2.3.2 where we show that, given a linear function $\ell(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} - t$ whose classification error is greater than $\mathrm{OPT}+\epsilon$, there exists a low-degree polynomial certifying function. We prove the existence of such a certificate in Subsection 2.3.1. In Subsection 2.3.2, we show that we can efficiently compute the low-degree polynomial certificate, and we bound the sample complexity and runtime of the corresponding SDP of the optimization problem. Finally, in Subsection 2.3.3, we show that given an efficient certificate oracle, we can use online gradient descent in order to learn the optimal halfspace, i.e., we provide the formal version and proof of Proposition 2.2.2.

### 2.3.1 The Low-Degree Polynomial Certificate

Here we show that given any $\epsilon$ suboptimal linear hypothesis $\ell(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} - t$ there exists a low degree polynomial certifying function. We establish the following.

**Proposition 2.3.2** (Certificate for General Halfspaces with Massart Noise: $\eta < 1/2$.)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to some at most $(1-\gamma)$-biased optimal halfspace $f(\mathbf{x})$, where $\gamma \in (0, 1/2]$. Let $\ell(\mathbf{x})$ be any linear function such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] \geq \mathrm{OPT} + \epsilon$. There exist $r_1, r_2 \in \mathbb{R}$ and polynomial*

$q(\mathbf{x}) = \sum_{|\alpha|\leq k} c_\alpha \mathbf{x}^\alpha$ *of degree* $k = \Theta(\log(\frac{1}{\gamma\beta}))$ *and* $\|q(\mathbf{x})\|_2 = 1$ *such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y\ \mathbb{1}\{r_1 \leq \ell(\mathbf{x}) \leq r_2\}q^2(\mathbf{x})] \leq -\epsilon^4 \mathrm{poly}(\beta\gamma)\|\ell(\mathbf{x})\|_2\,.$$

*Moreover, the coefficients of* $q$ *are bounded, specifically* $\sum_{|\alpha|\leq k} |c_\alpha| \leq d^{O(k)}$.

To prove the proposition, we distinguish different cases for the hypothesis $h(\mathbf{x})$. First we assume that $\mathbf{w} = 0$, i.e., the guess is not formally a halfspace but a constant function. This case may look easy to handle but captures many difficulties of designing certificates for biased halfspaces under Massart noise: when the constant guess is $-1$, we have to construct a polynomial that puts a large weight to the positive region of $f(\mathbf{x})$, i.e., $\{\mathbf{x} : f(\mathbf{x}) = +1\}$. Since the probability of this region can be as small as $\gamma$, it may be far from the origin (i.e., $t^* = \Theta(\sqrt{\log(1/\gamma)})$); see Figure 2.1a. It can be seen that it is challenging to isolate such regions far from the origin and the required polynomial degree is roughly $\log(\frac{1}{\beta\gamma})$. The argument and the details for this case are given in Subsection 2.3.1. In Subsection 2.3.1, we show that, similarly to the case of constant hypotheses, we can construct polynomial certificates against very biased hypotheses, i.e., when $\ell(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} - t$ and it holds $t/\|\mathbf{w}\|_2 \geq C\sqrt{\log(1/(\beta\gamma))}$. Finally, in Subsection 2.3.1, we handle the remaining hypotheses, namely we provide certificates when $t/\|\mathbf{w}\|_2 \leq C\sqrt{\log(1/(\beta\gamma))}$. Having Lemma 2.3.3, Lemma 2.3.12 and Lemma 2.3.16, we immediately obtain Proposition 2.3.2.

### Certificate Against Constant Hypotheses

Here we show that we can construct a polynomial certificate that certifies the non-optimality of any constant hypothesis, i.e., $h(\mathbf{x})$ corresponds to the constant function $+1$ (or $-1$) for every $\mathbf{x} \in \mathbb{R}^d$. In this case, we do not require a "band" $\mathbb{1}\{r_1 \leq \mathbf{w}\cdot\mathbf{x} \leq r_2\}$ as

part of the certificate, and therefore we may set $r_1 = -\infty$, $r_2 = +\infty$ in Proposition 2.3.2. We prove the following lemma:

**Lemma 2.3.3** (Certificate against Constant Hypotheses)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to some (at least) $(1-\gamma)$-biased optimal halfspace. For every constant hypothesis $s \in \{-1, +1\}$, there exists a polynomial $q(\mathbf{x}) = \sum_{|\alpha| \leq k} c_\alpha \mathbf{x}^\alpha$ of degree $k = \Theta(\log(\frac{1}{\beta\gamma}))$ with $\|q\|_2 = 1$, and sum of (absolute) coefficients $\sum_{|\alpha| \leq k} |c_\alpha| \leq d^{O(k)}$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[sy\ q^2(\mathbf{x})] \leq -\mathrm{poly}(\beta\gamma)\,.$$

*Proof.* Our plan is to pick the polynomial $q(\mathbf{x})$ so that it takes larger values to the side of the optimal halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} - t^*)$ that has the opposite sign of $s$. In the example of Figure 2.1a, in order to have a certificate for the constant hypothesis $s = -1$, we want to have a polynomial that takes much larger values in the blue area than the red area. We shall set $q(\mathbf{x}) = p(\mathbf{w}^* \cdot \mathbf{x})$ for some one-dimensional polynomial $p : \mathbb{R} \to \mathbb{R}$. Since the certificate that we construct only depends on the direction of the optimal halfspace, it follows that we can project the points on the subspace spanned by $\mathbf{w}^*$. The following claim shows that the projection of a distribution with $\eta$-Massart noise onto a lower dimensional subspace that contains the direction of the optimal halfspace also satisfies the $\eta$-Massart noise condition with respect to the same optimal halfspace.

**Claim 2.3.4** (Projections preserve Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ satisfying the $\eta$-Massart noise condition with respect to some optimal halfspace $f(\mathbf{x}) : \mathbb{R}^d \mapsto \{\pm 1\}$. Let $V$ be any subspace of $\mathbb{R}^d$ that contains the normal vector of the optimal*

*halfspace* $f(\mathbf{x})$*. Then, for every function* $g : \mathbb{R}^d \mapsto \mathbb{R}$ *it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[g(\mathrm{proj}_V(\mathbf{x}))y] = \mathop{\mathbf{E}}_{\mathbf{v}\sim(\mathcal{D}_\mathbf{x})_V}[g(\mathbf{v})\beta_V(\mathbf{v})f(\mathbf{v})]\,,$$

*where* $\beta_V(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ *satisfies* $\beta_V(\mathbf{x}) \in [1-2\eta, 1]$*.*

*Proof.* Assume that the optimal halfspace is $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} - t^*)$. From the fact that the distribution $\mathcal{D}$ satisfies the $\eta$-Massart noise condition, it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y|\mathbf{x}] = f(\mathbf{x})(1-\eta(\mathbf{x})) - f(\mathbf{x})\eta(\mathbf{x}) = f(\mathbf{x})(1-2\eta(\mathbf{x})) = f(\mathbf{x})\beta(\mathbf{x})$. Therefore, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[g(\mathrm{proj}_V(\mathbf{x}))y] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[g(\mathrm{proj}_V(\mathbf{x}))\beta(\mathbf{x})f(\mathbf{x})] \\
&= \mathop{\mathbf{E}}_{\mathbf{v}\sim(\mathcal{D}_\mathbf{x})_V}\left[\mathop{\mathbf{E}}_{\mathbf{u}\sim(\mathcal{D}_\mathbf{x})_{V^\perp}}[g(\mathbf{v})\beta(\mathbf{u}+\mathbf{v})\mathrm{sign}(\mathbf{w}^*\cdot(\mathbf{u}+\mathbf{v}) - t^*)\right] \\
&= \mathop{\mathbf{E}}_{\mathbf{v}\sim(\mathcal{D}_\mathbf{x})_V}\left[\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{v} - t^*)\; g(\mathbf{v}) \mathop{\mathbf{E}}_{\mathbf{v}\sim(\mathcal{D}_\mathbf{x})_{V^\perp}}[\beta(\mathbf{u}+\mathbf{v})]\right],
\end{aligned}$$

where we used the fact that the subspace $V$ contains the normal vector of the optimal halfspace $\mathbf{w}^*$. Observe that the "projected" noise function $\beta_V(\mathbf{v}) := \mathbf{E}_{\mathbf{u}\sim(\mathcal{D}_\mathbf{x})_{V^\perp}}[\beta(\mathbf{u}+\mathbf{v})]$ again satisfies the $\eta$-Massart noise condition, i.e., $\beta_V(\mathbf{v}) \in [1-2\eta, 1] = [\beta, 1]$. This concludes the proof of Claim 2.3.4. □

Since $q(\mathbf{x}) = p(\mathbf{w}^*\cdot\mathbf{x})$, we can use Claim 2.3.4 to project onto the subspace spanned by $\mathbf{w}^*$ and obtain that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[syq^2(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{v}\sim(\mathcal{D}_\mathbf{x})_{\mathbf{w}^*}}\left[s\;\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{v} - t^*)\; p^2(\mathbf{w}^*\cdot\mathbf{v})\; \beta_{\mathbf{w}^*}(\mathbf{v})\right], \tag{2.3}$$

where the "projected" noise function $\beta_{\mathbf{w}^*}(\mathbf{v}) \in [\beta, 1]$. For simplicity, in what follows, we

will continue denoting $\beta$ the projected noise function $\beta_{\mathbf{w}^*}$. Without loss of generality, we may assume that $\mathbf{w}^* = \mathbf{e}_1$ and in that case, using the fact that the projection of $\mathcal{D}_{\mathbf{x}}$ onto $\mathbf{w}^*$ is a one-dimensional standard normal distribution, we have that the expression of Equation (2.3) can be simplified as

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[syq^2(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[s\beta(\mathbf{x}_1)\mathrm{sign}(\mathbf{x}_1 - t^*)\ p^2(\mathbf{x}_1)]\,. \tag{2.4}$$

Moreover, to simplify the notation, assume that the constant guess is $s = -1$ (the case of $s = +1$ is similar) and that the halfspace puts $\gamma$ mass on the positive side, i.e., $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[f(\mathbf{x}) = +1] = \gamma$; see also Figure 2.1a. In that case, we want to construct a univariate polynomial $p(\mathbf{x}_1)$ so that $\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[\beta(\mathbf{x}_1)\mathrm{sign}(\mathbf{x}_1 - t^*)\ p^2(\mathbf{x}_1)]$ is sufficiently positive. Observe that the worst case noise function $\beta(\mathbf{x}_1)$ is to set $\beta(\mathbf{x}_1) = \beta$ for all points $\mathbf{x}_1 \geq t^*$ and $\beta(\mathbf{x}_1) = 1$ for all $\mathbf{x}_1 < t^*$. In that case, we want to find a univariate polynomial $p$ such that

$$\beta \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_1 \geq t^*\}p^2(\mathbf{x}_1)] \geq \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_1 \leq t^*\}p^2(\mathbf{x}_1)]\,. \tag{2.5}$$

To do that, we show our main structural result, i.e., that as long as the degree $k$ is larger than $(t^*)^2$, there exists a polynomial that can make the above inequality true. In fact, we can make the ratio of $\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_1 \geq t^*\}p^2(\mathbf{x}_1)]/\,\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_1 \leq t^*\}p^2(\mathbf{x}_1)]$ grow exponentially fast with respect to the degree-$k$ of the polynomial $p$. We prove the following lemma.

**Lemma 2.3.5** (Polynomial Shift)**.** *Let $t \in \mathbb{R}$ and $b \geq 1$. There exists a univariate polynomial $p(x) = \sum_{i=0}^{k} a_i x^i$ of degree $k = \Theta(t^2 + b^2)$, $L_2$ norm $\|p(x)\|_2 = 1$, and sum*

*of (absolute) coefficients* $\sum_{i=0}^{k} |a_i| = O(1)$*, such that*

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \geq t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \leq t\}]} \geq e^{b^2} \quad \textit{and} \quad \mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \geq t\}] \geq e^{-O(k)}\,.$$

*Proof.* We first show that there exists a positive function that makes the ratio of Lemma 2.3.5 large. We shall consider the exponential function $x \mapsto e^{cx}$ for some constant $c \geq |t|$ to be determined later. We have that

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[e^{cx}\mathbb{1}\{x \geq t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[e^{cx}\mathbb{1}\{x \leq t\}]} = \frac{\mathbf{Pr}_{x\sim\mathcal{N}}[x \geq t-c]}{\mathbf{Pr}_{x\sim\mathcal{N}}[x \leq t-c]} \geq e^{(c-t)^2/2}\,, \tag{2.6}$$

where the last inequality follows from the fact that $c \geq t$ and standard lower bounds on the tail probability of the standard normal distribution. Therefore, for $c = \Theta(|t| + b)$ we immediately obtain the claimed bound. We now have to replace the function $e^{cx}$ by the square of a polynomial $p$. To do so we use the Taylor expansion of the exponential function. Denote by $S_k(x)$ the degree-$k$ Taylor expansion of $e^x$, i.e., $S_k(x) = \sum_{i=0}^{k} x^i/i!$. Using Taylor's remainder theorem and the fact that $e^x = \sum_{i=0}^{\infty} x^i/i!$ for all $x \in \mathbb{R}$, we obtain the following fact.

**Fact 2.3.6** (Taylor Expansion of $e^x$)**.** *Fix* $R > 0$ *and let* $S_k(x)$ *be the degree-*$k$ *Taylor expansion of* $e^x$*.*

1. *For all* $x \in [-R, R]$ *we have that* $|S_k(x) - e^x| \leq e^R \frac{R^{k+1}}{(k+1)!}$*.*
2. *For all* $x \in \mathbb{R}$ *it holds that* $|S_k(x)| \leq e^{|x|}$*.*

Since we have to construct the square of a polynomial $p$ we cannot simply use $S_k(x)$. However, we can consider the function $e^{2cx}$ and use $S_k^2(cx)$ as our approximation. We

first show that as long as $k$ is at least $\Omega(c^2)$, we have that $S_k^2(cx)$ is a good approximation of $e^{2cx}$ with respect to the $L_1$ norm.

**Claim 2.3.7** ($L_1$ error of $S_k(x)$)**.** *Assume that $k \geq 32c^2$. It holds*

$$\|S_k^2(cx) - e^{2cx}\|_1 \lesssim e^{-k/32}\,.$$

*Proof.* For every $x \in [-R, R]$ it holds that

$$|S_k^2(x) - (e^x)^2| = |S_k(x) - e^x||S_k(x) + e^x| \leq e^R \frac{R^{k+1}}{(k+1)!} \quad 2e^R = 2e^{2R}\frac{R^{k+1}}{(k+1)!}\,, \tag{2.7}$$

where we used the second item of Fact 2.3.6, i.e, that $|S_k(x)| \leq e^R$. The above pointwise approximation guarantee together with the strong concentration properties of the Gaussian distribution (see, e.g., Fact 2.2.11) allow us to get a bound for the $L_1$-approximation error of $S_k^2(x)$. Using the error bound of Equation (2.7), we bound the $L_1$-approximation error as follows:

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|S_k^2(cx) - e^{2cx}|] &\leq \max_{|x|\leq R/c} |S_k^2(cx) - e^{2cx}| + \mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|S_k^2(cx) + e^{2cx}|\ \mathbb{1}\{|x| \geq R/c\}] \\
&\leq 2e^{2R}\frac{R^{k+1}}{(k+1)!} + (\|S_k^2(cx)\|_2 + \|e^{2cx}\|_2)\sqrt{\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}}[|x| \geq R/c]} \\
&\leq 2e^{2R}\frac{R^{k+1}}{(k+1)!} + 2\sqrt{2}e^{4c^2 - (R/c)^2/2}\,,
\end{aligned} \tag{2.8}$$

where for the last inequality we used the second item of Fact 2.3.6 to obtain that $\|S_k^2(x)\|_2 \leq \|e^{2c|x|}\|_2$, the fact that $\|e^{2c|x|}\|_2 \leq \sqrt{2}e^{4c^2}$, and the tail probability upper bound for the normal distribution, i.e., $\mathbf{Pr}_{x\sim\mathcal{N}}[|x| \geq (R/c)] \leq e^{-(R/c)^2/2}$, see Fact 2.2.11. By choosing $k = 32mc^2$, for any $m \geq 1$, and $R = k/8$ the estimate of Equation (2.8)

becomes

$$\|S_k^2(cx) - e^{2cx}\|_1 \lesssim e^{-mc^2} .$$

□

We are now ready to finish the proof of Lemma 2.3.5. Recall that to make Inequality (2.6) true, we chose $c = \Theta(b+|t|)$. Let $k = 32mc^2$, for some sufficiently large absolute constant $m > 0$. Given Inequality (2.6), we show that by replacing $e^{2cx}$ with $S_k^2(cx)$, i.e., the Taylor expansion of $e^{cx}$ squared, we get a similar formula for the lower-bound of the ratio. It suffices to show that

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \le t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[S_k^2(cx)\mathbb{1}\{x \le t\}]} \ge \frac{1}{2} \quad \text{and} \quad \frac{\mathbf{E}_{x\sim\mathcal{N}}[S_k^2(cx)\mathbb{1}\{x \ge t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \ge t\}]} \ge \frac{1}{2} . \tag{2.9}$$

We start from the first inequality of Equation (2.9). To prove it, it suffices to show that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|S_k^2(x) - e^{2cx}|\ \mathbb{1}\{x \le t\}] \le \mathop{\mathbf{E}}_{x\sim\mathcal{N}}[e^{2cx}\ \mathbb{1}\{x \le t\}] . \tag{2.10}$$

From the $L_1$-bound of Claim 2.3.7, we have that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|S_k^2(x) - e^{2cx}|\ \mathbb{1}\{x \le t\}] \le \mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|S_k^2(x) - e^{2cx}|] \lesssim e^{-mc^2} .$$

Moreover, to bound the $\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\ \mathbb{1}\{x \le t\}]$ from below, we use the lower bound for the tails of a Gaussian distribution (Fact 2.2.11). First, note that $\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\ \mathbb{1}\{x \le t\}] = e^{2c^2}\,\mathbf{Pr}_{x\sim\mathcal{N}}[x \le t-2c]$. Therefore, we have that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[e^{2cx}\ \mathbb{1}\{x \le t\}] \ge \frac{1}{4}e^{2c^2}e^{-(t-2c)^2} \ge \frac{1}{4}e^{-7c^2} ,$$

where we used that $c \geq |t|$. Therefore, Inequality (2.10) is true for $m$ being a large enough absolute constant.

For the second inequality of Equation (2.9), using the $L_1$-bound of Claim 2.3.7 we obtain that $\mathbf{E}_{x\sim\mathcal{N}}[S_k^2(cx)\mathbb{1}\{x \geq t\}] \geq \mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \geq t\}] - \|S_k^2(x) - e^{2cx}\|_1$, and therefore, taking $m$ to be larger than an absolute constant we have that, there exists an absolute constant $C > 0$, such that

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[S_k^2(cx)\mathbb{1}\{x \geq t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \geq t\}]} \geq 1 - \frac{Ce^{-mc^2}}{\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \geq t\}]} \geq \frac{1}{2},$$

where we used that $\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \geq t\}] = e^{2c^2}\,\mathbf{Pr}_{x\sim\mathcal{N}}[x \geq t - 2c] \geq e^{2c^2}/2$ since $c \geq t$.

We next show that we can normalize the polynomial $p(x) = S_k(cx)$ without making the expectation of $p^2$ over $x \geq t$, i.e., $\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x \geq t\}]$, too small. From the second item of Fact 2.3.6 we obtain that $\|S_k^2(cx)\|_2 \leq \|e^{2c|x|}\|_2 = e^{O(c^2)}$. The result follows from the second inequality of Equation (2.9) and the fact that $\mathbf{E}_{x\sim\mathcal{N}}[e^{2cx}\mathbb{1}\{x \geq t\}] \geq e^{2c^2}/2$.

Finally, we bound the coefficients of the polynomial $S_k^2(cx)$. We have $S_k^2(cx) = \left(\sum_{i=0}^k \frac{c^i}{i!}x^i\right)^2$. It holds that $\left(\sum_{i=0}^k \frac{c^i}{i!}\right)^2 = e^{2c}$. From the $L_1$ approximation guarantee of Claim 2.3.7 we obtain that $\|S_k^2(cx)\|_2 \geq \|S_k^2(cx)\|_1 \geq \|e^{2cx}\|_1 - 1 \geq e^{2c^2}/2$, since $c \geq 1$. We conclude that the sum of the absolute coefficients of $S_k^2(cx)/\|S_k^2(cx\|_2$ is at most 2, which implies that the sum of the absolute coefficients of $S_k(cx)$ is at most $\sqrt{2}$.

□

For the upper bound:

**Lemma 2.3.8.** *Let $t \geq 0$. There exists an absolute constant $C' \geq 1$ such that for any*

*univariate polynomial of degree $k$ it holds*

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x\geq t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x\leq t\}]}\leq e^{C'k\log k-t^2/C'}\,.$$

*Proof.* We start by bounding from above the $\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x\geq t\}]$. Using the Cauchy-Schwarz inequality, we get

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x\geq t\}]\leq(\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^4(x)])^{1/2}(\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}}[x\geq t])^{1/2}\lesssim(\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^4(x)])^{1/2}e^{-t^2/4}\,. \quad (2.11)$$

In order to bound the $\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x\leq t\}]$ from below, we are going to use Lemma 2.3.9. By setting $q=4k, n=k$ and $\gamma=(\frac{1}{12Ck})^k(\mathbf{E}_{x\sim\mathcal{N}}[p^4(x)])^{1/4}$, on Lemma 2.3.9. We get that

**Lemma 2.3.9** (Theorem 8 of [CW01])**.** *Let $p:\mathbb{R}^d\mapsto R$ be a polynomial of degree at most $n$. Then there is an absolute constant $C>0$ such that for any $0<q<\infty$ and $t\geq 0$, it holds $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}_d}[|p(\mathbf{x})|\leq\gamma]\leq Cq\gamma^{1/n}(\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_d}[|p(\mathbf{x})|^{q/n}])^{-1/q}$.*

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}}\left[|p(x)|\leq\left(\frac{1}{12Ck}\right)^k\left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^4(x)]\right)^{1/4}\right]\leq 1/3\,.$$

Therefore, by squaring we get

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}}\left[p^2(x)\geq\left(\frac{1}{12Ck}\right)^{2k}\left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^4(x)]\right)^{1/2}\right]\geq 2/3\,. \quad (2.12)$$

Furthermore, using the assumption that $t\geq 0$, we have $\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x\leq t\}]\geq$

$\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x\le 0\}]$, hence, combing the with Equation (2.12), we get

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x\le t\}] \gtrsim \left(\frac{1}{12Ck}\right)^{2k}\left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[p^4(x)]\right)^{1/2}. \tag{2.13}$$

Combining eq. (2.11) and eq. (2.13), we get

$$\frac{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x\ge t\}]}{\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)\mathbb{1}\{x\le t\}]} \lesssim (12Ck)^{2k}e^{-t^2/4} \le e^{C'k\log k - t^2/C'},$$

for some $C'\ge 1$ absolute constant. □

We prove below Fact 2.3.18, we restate it for convenience.

**Fact 2.3.10** (Estimation of $\mathbf{M}$)**.** *Let $\Omega=\{\mathbf{A}\in\mathcal{S}^m:\ \|\mathbf{A}\|_F\le 1\}$ and $\epsilon,\delta\in(0,1)$. Let $\ell(\mathbf{x})=\mathbf{w}\cdot\mathbf{x}+t$ with $|\ell(\mathbf{x})|^2\le C$ and $\widetilde{\mathbf{M}}=\frac{1}{N}\sum_{i=1}^N \mathbf{m}(\mathbf{x}^{(i)})\mathbf{m}(\mathbf{x}^{(i)})^\intercal\mathbb{1}_B(\mathbf{x}^{(i)})y^{(i)}\ell(\mathbf{x}^{(i)})$. There exists an algorithm that draws $N=\frac{d^{O(\log\frac{1}{\gamma\beta})}}{C\epsilon^2}\log(1/\delta)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(N,d)$ time and with probability at least $1-\delta$ outputs a matrix $\widetilde{\mathbf{M}}$ such that*

$$\mathbf{Pr}\left[\sup_{\mathbf{A}\in\Omega}\left|\mathrm{tr}(\mathbf{A}\widetilde{\mathbf{M}})-\mathrm{tr}(\mathbf{A}\mathbf{M})\right|\ge\epsilon\right]\le 1-\delta\,.$$

To conclude the proof of Lemma 2.3.3, notice that using Lemma 2.3.5, we obtain that there exists a polynomial $p$ of degree $(t^*)^2+\log(1/\beta)$ such that Equation (2.5) is true. Moreover, from the same lemma, we obtain that $\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[p^2(\mathbf{x}_1)]=e^{-O((t^*)^2+\log(1/\beta))}$. Since the halfspace is at least $(1-\gamma)$-biased, from standard upper bounds on the Gaussian tails, we obtain that $|t^*|\lesssim\sqrt{\log(1/\gamma)}$ obtaining the claimed bound for the degree of the certifying polynomial $p$. It remains to bound the coefficients of the multivariate polynomial $q(\mathbf{x})=p(\mathbf{w}^*\cdot x)$. We are going to use the following fact.

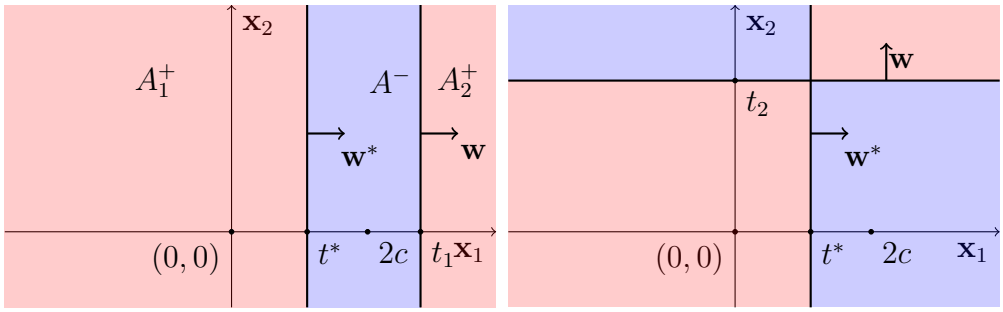



Figure 2.4: **(a)** The guess $\mathbf{w} = \mathbf{e}_1 - t_1$ and $t_1$ is larger than $C\sqrt{\log(1/(\beta\gamma))}$ for some sufficiently large constant $C$. We can pick $c$ such that $2c - t^*$ and $t_1 - 2c$ are both large constants making the exponential shift certificate $e^{2c\mathbf{x}_1}$ work, similarly to the constant case of Figure 2.1a.
**(b)** The guess $\mathbf{w} = \mathbf{e}_2 - t_2$ and $t_1$ is larger than $C\sqrt{\log(1/(\beta\gamma))}$ for some sufficiently large constant $C$. The region $\mathbf{x}_2 \geq t_2$ has probability at most $e^{-O(t_2^2)}$ and for $\mathbf{x}_2 \leq t_2$ we can treat $\mathbf{x}_2 - t_2$ as a constant negative guess.

**Fact 2.3.11** (Lemma 3.6 of [DKTZ20b])**.** *Let $p(t) = \sum_{i=0}^k c_i t^i$ be a degree-$k$ univariate polynomial. Given $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq 1$, define the multivariate polynomial $q(\mathbf{x}) = p(\mathbf{w} \cdot \mathbf{x}) = \sum_{S:|S|\leq k} C_S \mathbf{x}^S$. Then we have that $\sum_{S:|S|\leq k} C_S^2 \leq d^{2k} \sum_{i=0}^k c_i^2$.*

From Lemma 2.3.5 we know that the sum of the absolute coefficients of $p$ is $O(1)$. Thus using Fact 2.3.11 we obtain that the sum of absolute coefficients of $p(\mathbf{w} \cdot \mathbf{x})$ is at most $d^{O(k)}$.

□

### Certificate Against "Large Threshold" Halfspaces

We now deal with the case where the halfspace has a large threshold but is not the constant hypothesis. In particular, we assume that $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} - t)$ with $t/\|\mathbf{w}\| \geq C\sqrt{\log(1/(\beta\gamma))}$ for some sufficiently large absolute constant $C$. This case is, in fact, a generalization of the constant hypothesis case that corresponds to $\mathbf{w} = \mathbf{0}$.

In this case we will show that roughly the same polynomial that we used for constant guesses in Subsection 2.3.1 can also be made to work for very biased guesses. We prove the next lemma.

**Lemma 2.3.12** (Certificate against "Large Threshold" Hypotheses)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to some (at most) $(1-\gamma)$-biased optimal halfspace. Define the linear function $\ell(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} - t)$ and assume that $t/\|\mathbf{w}\|_2 \geq C\sqrt{\log(1/(\beta\gamma))}$ for some sufficiently large absolute constant $C$. Then, there exists polynomial $q(\mathbf{x}) = \sum_{|\alpha| \leq k} c_\alpha \mathbf{x}^\alpha$ of degree $\Theta(\log(\frac{1}{\beta\gamma}))$, $L_2$ norm $\|q(\mathbf{x})\|_2 = 1$, and sum of absolute coefficients $\sum_{|\alpha| \leq k} |c_\alpha| = d^{O(k)}$, such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x}) y q^2(\mathbf{x})] \leq -\|\ell(\mathbf{x})\|_2 \ \mathrm{poly}(\beta\gamma)\,.$$

*Proof.* Using Claim 2.3.4, we can project the distribution $\mathcal{D}$ on the 2-dimensional subspace $V$ spanned by $\mathbf{w}$ and $\mathbf{w}^*$. We have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x}) y p^2(\mathbf{w}^* \cdot \mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}_V}[\ell(\mathbf{v})\beta_V(\mathbf{v})\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{v} - t^*) p^2(\mathbf{w}^* \cdot \mathbf{v})]\,. \tag{2.14}$$

In what follows, we again abuse notation and denote $\beta_V(\mathbf{v})$ simply as $\beta(\mathbf{v})$. By the spherical symmetry of the Gaussian distribution, we can, without loss of generality, assume that $\mathbf{w}^* = \mathbf{e}_1$ and $\mathbf{w} = \|\mathbf{w}\|_2(\cos\theta \mathbf{e}_1 + \sin\theta \mathbf{e}_2)$. Moreover, we may assume that $t > 0$ and $\theta \in [0, \pi/2]$ (see, e.g., Figure 2.4) since the other cases are similar. Our certificate will be the same polynomial as in the previous case of Lemma 2.3.3. In particular, we will again use the square of the Taylor approximation $S_k^2(c\mathbf{x}_1)$ of the exponential $e^{2c\mathbf{x}_1}$.

We decompose the problem of proving that this polynomial is a certificate for general halfspaces to the problem of showing that it is a certificate for halfspaces with large thresholds simultaneously in both orthogonal directions. In particular, using Equation (2.14), we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})yp^2(\mathbf{w}^*\cdot\mathbf{x})] \\
&\quad= \|\mathbf{w}\|_2\cos\theta \mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}\left[\left(\mathbf{x}_1 - \frac{t}{2\|\mathbf{w}\|_2\cos\theta}\right)\beta(\mathbf{x}_1,\mathbf{x}_2)p^2(\mathbf{x}_1)\mathrm{sign}(\mathbf{x}_1 - t^*)\right] \\
&\quad+ \|\mathbf{w}\|_2\sin\theta \mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}\left[\left(\mathbf{x}_2 - \frac{t}{2\|\mathbf{w}\|_2\sin\theta}\right)\beta(\mathbf{x}_1,\mathbf{x}_2)p^2(\mathbf{x}_1)\mathrm{sign}(\mathbf{x}_1 - t^*)\right].
\end{aligned} \tag{2.15}$$

To simplify the notation, set $t_1 = t/(2\|\mathbf{w}\|_2\cos\theta)$ and $t_2 = t/(2\|\mathbf{w}\|_2\sin\theta)$ and notice that by the assumptions of Lemma 2.3.12, it holds that both $t_1$ and $t_2$ are larger than $C/2\sqrt{\log(1/(\beta\gamma))}$, see Figure 2.4. We will show that we can pick $c = \Theta(\log(1/(\beta\gamma))$ and $k = \Theta(c^2)$ so that the polynomial $S_k^2(c\mathbf{x}_1)$ simultaneously satisfies for $i = 1, 2$, the following inequality

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[(\mathbf{x}_i - t_i)\beta(\mathbf{x}_1,\mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \le -t_i\ \mathrm{poly}(\beta\gamma)\,.$$

Since $t_i \ge 1$, it follows that we can replace the $t_i\ \mathrm{poly}(\beta\gamma)$ above by $(2t_i+1)\mathrm{poly}(\beta\gamma)$. Having these bounds and using Equation (2.15), we obtain

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})yp^2(\mathbf{w}^*\cdot\mathbf{x})] &= -(\|\mathbf{w}\|_2(\cos\theta+\sin\theta)+t)\ \mathrm{poly}(\gamma\beta) \\
&= -(\|\mathbf{w}\|_2 + t)\ \mathrm{poly}(\gamma\beta) = -\|\ell(\mathbf{x})\|_2\ \mathrm{poly}(\beta\gamma)\,,
\end{aligned}$$

where for the last inequality we used the fact that $\cos\theta + \sin\theta \geq 1$, for $\theta \in [0, \pi/2]$ and that $\sqrt{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\ell(\mathbf{x})^2]} = \sqrt{\|\mathbf{w}\|_2^2 + t^2}$. We bound the first term in claim Claim 2.3.15 and the second in Claim 2.3.13.

We first bound the contribution of the direction that is orthogonal to the optimal, i.e., $\mathbf{e}_2$. At a high-level, we have that, since $t_2$ is a large multiple of $\sqrt{\log(1/(\beta\gamma))}$, "for most" values of $\mathbf{x}_2$, the quantity $\mathbf{x}_2 - t_2$ will be negative. Therefore, $\mathbf{x}_2 - t_2$ roughly corresponds to the constant guess $-1$, that we covered in Lemma 2.3.3.

**Claim 2.3.13.** *It holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1,\mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \leq -t_2 \ \mathrm{poly}(\beta\gamma) \ .$$

*Proof.* When $\mathbf{x}_2 - t_2 < 0$, we can treat $\mathbf{x}_2 - t_2$ as a constant negative guess, and use directly the bound obtained in the proof of Lemma 2.3.3, for $s = -1$, to make the term $\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[(-1)\beta(\mathbf{x}_1)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \leq -C'(\beta\gamma)^{\rho_1}$, for some absolute constants $\rho_1, C' > 0$. In particular, since in this case it holds $\mathbf{x}_2 - t_2 \leq 0$, the "worst-case" noise function $\beta(\mathbf{x}_1,\mathbf{x}_2)$ is $\beta(\mathbf{x}_1,\mathbf{x}_2) = \beta$ for all $(\mathbf{x}_1,\mathbf{x}_2)$ such that $\mathbf{x}_1 \geq t^*$ and $\beta(\mathbf{x}_1,\mathbf{x}_2) = 1$ for $\mathbf{x}_1 \leq t^*$, see Figure 2.4b. Notice that this worst case $\beta(\mathbf{x}_1,\mathbf{x}_2)$ does not depend on $\mathbf{x}_2$. Therefore, we obtain

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\mathbb{1}\{\mathbf{x}_2 \leq t_2\}(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1,\mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \\
&\qquad \leq \mathop{\mathbf{E}}_{\mathbf{x}_2\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_2 \leq t_2\}(\mathbf{x}_2 - t_2)] \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[(-1)\beta(\mathbf{x}_1)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \\
&\qquad \leq -\mathop{\mathbf{E}}_{\mathbf{x}_2\sim\mathcal{N}}[\mathbb{1}\{\mathbf{x}_2 \leq t_2\}|\mathbf{x}_2 - t_2|] \ \ C'(\beta\gamma)^{\rho_1} \\
&\qquad \leq -C'/2 \ t_2 \ \ (\beta\gamma)^{\rho_1} \ , \qquad (2.16)
\end{aligned}$$

where the last inequality follows by our assumption that $t_2 \geq 1$.

On the other hand, the probability of the region $\{\mathbf{x}_2 \in \mathbb{R} : \mathbf{x}_2 - t_2 \geq 0\}$ is exponentially small which allows us to control the contribution of the region where $\mathbf{x}_2 \geq t_2$. To bound the expectation of $p^4(\mathbf{x}_1)$, we use the following lemma known as Bonami-Beckner inequality or simply Gaussian hypercontractivity.

**Lemma 2.3.14** (Gaussian Hypercontractivity)**.** *Let $p : \mathbb{R} \to \mathbb{R}$ be any polynomial of degree at most $\ell$. Then, for every $q \geq 2$, it holds $\mathbf{E}_{x\sim\mathcal{N}}[|p(x)|^q] \leq (q-1)^{q\ell/2}\Big(\mathbf{E}_{x\sim\mathcal{N}}[p^2(x)]\Big)^{q/2}$.*

Using Cauchy-Schwarz and the fact that $\beta(\mathbf{x}) \leq 1$, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\mathbb{1}\{\mathbf{x}_2 \geq t_2\}(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1,\mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \leq \sqrt{\mathop{\mathbf{Pr}}_{\mathbf{x}_2\sim\mathcal{N}}[\mathbf{x}_2 \geq t_2]}\left(\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[(\mathbf{x}_2 - t_2)^4]\right)^{1/4}\left(\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[p^8(\mathbf{x}_1)]\right)^{1/4}. \tag{2.17}$$

Using Gaussian hypercontractivity, Lemma 2.3.14, we obtain that $(\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[p^8(\mathbf{x}_1)])^{1/4} = e^{O(k)}$. Moreover, from standard bounds of Gaussian tails, Fact 2.2.11, we have that $\mathbf{Pr}_{\mathbf{x}_2\sim\mathcal{N}}[\mathbf{x}_2 \geq t_2] \leq e^{-t_2^2/2}$. Finally, we have that $(\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[(\mathbf{x}_2 - t_2)^4])^{1/4} = (3 + 6t_2^2 + t_2^4)^{1/4} \leq 2t_2$, for all $t_2 \geq 1$. Combining the above bounds, we obtain that

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\mathbb{1}\{\mathbf{x}_2 \geq t_2\}(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1,\mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \leq t_2 e^{-t_2^2/4}e^{O(k)}. \tag{2.18}$$

Recall that $k = O(\log(1/(\beta\gamma)))$, therefore there exists an absolute constant $\rho_2 > 0$ such that $e^{O(k)} \leq (1/(\beta\gamma))^{\rho_2}$. Therefore, combining eq. (2.16) and eq. (2.18), we get that

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1,\mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \leq -Ct_2(\beta\gamma)^{\rho_1} + t_2 e^{-t_2^2/4}(\beta\gamma)^{\rho_2}.$$

Hence, because $t_2 \geq C\sqrt{\log(1/(\beta\gamma))}/2$, for $C$ sufficiently large positive absolute constant, we have that $\mathbf{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[(\mathbf{x}_2 - t_2)\beta(\mathbf{x}_1,\mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] = -t_2\ \mathrm{poly}(\beta\gamma)$. □

We now bound the contribution of the direction parallel to the optimal vector, i.e., $\mathbf{e}_1$.

**Claim 2.3.15.** *It holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[(\mathbf{x}_1 - t_1)\beta(\mathbf{x}_1,\mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)p^2(\mathbf{x}_1)] \leq -t_1\ \mathrm{poly}(\beta\gamma)\ .$$

*Proof.* Notice that we can marginalize out the direction $\mathbf{e}_2$ in this case. The analysis is similar to that of Lemma 2.3.3. However, we now have to distinguish three different intervals for $\mathbf{x}_1$ inside which $(\mathbf{x}_1 - t_1)\mathrm{sign}(\mathbf{x}_1 - t^*)$ has the same sign. We will crucially use the fact that $|t^*| = O(\sqrt{\log(1/\gamma)})$, since the Gaussian puts at least $\gamma$ mass on the positive side of the halfspace, and that the constant $C$ is sufficiently large so that $t_1$ is at least a large constant multiple of $t^*$. We have the region $A^+ = A_1^+ \cup A_2^+$, where $A_1^+ = \{\mathbf{x}_1 : \mathbf{x}_1 \leq t^*\}$ and $A_2^+ = \{\mathbf{x}_1 : \mathbf{x}_1 \geq t_1\}$. In $A^+$ it holds $(\mathbf{x}_1 - t_1)\mathrm{sign}(\mathbf{x}_1 - t^*) \geq 0$. We also define the disagreement region $A^- = \{\mathbf{x}_1 : t^* \leq \mathbf{x}_1 \leq t_1\}$, where $(\mathbf{x}_1 - t_1)\mathrm{sign}(\mathbf{x}_1 - t^*) \leq 0$. The worst case noise function puts $\beta(\mathbf{x}_1,\mathbf{x}_2) = \beta$ when $\mathbf{x}_1 \in A^-$ and $\beta(\mathbf{x}_1,\mathbf{x}_2) = 1$ otherwise. Thus, in order to show that $p$ is a certifying polynomial, we have to show that

$$\frac{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[|\mathbf{x}_1 - t_1|p^2(\mathbf{x}_1)\mathbb{1}\{\mathbf{x}_1 \in A^-\}]}{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[|\mathbf{x}_1 - t_1|p^2(\mathbf{x}_1)\mathbb{1}\{\mathbf{x}_1 \in A^+\}]} \geq \frac{1}{\beta}\,. \tag{2.19}$$

Similarly to the proof of Lemma 2.3.5 we will first use an exponential function to shift the mean of the Gaussian, i.e., use $e^{2c\mathbf{x}_1}$ instead of $p^2(\mathbf{x}_1)$ in the ratio of Inequality (2.19). Our goal is to show that there exists some $c \geq t^*$ such that the probability of the region

$A^-$ is much larger than the probability of $A^+$. Since $|t_1|$ is larger than some sufficiently large constant multiple of $\sqrt{\log(1/(\beta\gamma))}$ the interval $[t^*, t_1]$ will contain most of the mass of the shifted normal $\mathcal{N}(c, 1)$ for $c$ some constant multiple of $\sqrt{\log(1/(\beta\gamma))}$, see Figure 2.1b.

We start by bounding from above the denominator of Inequality (2.19).

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[|\mathbf{x}_1 - t_1|e^{2c\mathbf{x}_1}\mathbb{1}\{\mathbf{x}_1 \in A^+\}] &\le \|\mathbf{x}_1 - t_1\|_2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[e^{4c\mathbf{x}_1}\mathbb{1}\{\mathbf{x}_1 \in A^+\}]} \\
&\le 2t_1e^{4c^2}\left(\sqrt{\mathop{\mathbf{Pr}}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbf{x}_1 \le t^* - 4c]} + \sqrt{\mathop{\mathbf{Pr}}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbf{x}_1 \ge t_1 - 4c]}\right) \\
&\le 2t_1e^{4c^2}(e^{-(t^*-4c)^2/4} + e^{-(t_1-4c)^2)/4}) \\
&\le t_1e^{4c^2}\beta/16\,, \qquad (2.20)
\end{aligned}$$

where we used that $\|\mathbf{x}_1 - t_1\|_2 = (\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[(\mathbf{x}_1 - t_1)^2]^{1/2} = (1 + t_1^2)^{1/2} \le 2t_1$ for all $t_1 \ge 1$ and, for the last inequality we used the Gaussian tail upper bound, see Fact 2.2.11, and the fact that $c = \Theta(\sqrt{\log(1/(\beta\gamma))})$, i.e., we get the constant $1/16$ by appropriately choosing some $c = \Theta(\sqrt{\log(1/(\beta\gamma))})$. We next bound the numerator from below. We have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[|\mathbf{x}_1 - t_1|e^{2c\mathbf{x}_1}\mathbb{1}\{\mathbf{x}_1 \in A^-\}] &\ge \frac{t_1}{2}\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[e^{2c\mathbf{x}_1}\mathbb{1}\{t^* \le \mathbf{x}_1 \le t_1/2\}] \\
&= \frac{t_1}{2}e^{4c^2}\mathop{\mathbf{Pr}}_{\mathbf{x}_1\sim\mathcal{N}}[t^* \le \mathbf{x}_1 + 2c \le t_1/2] \ge \frac{t_1e^{4c^2}}{4}\,, \qquad (2.21)
\end{aligned}$$

where for the last inequality we used the fact that $\mathbf{Pr}_{\mathbf{x}_1\sim\mathcal{N}}[t^* \le \mathbf{x}_1 + 2c \le t_1/2]$ is at least $1/2$ since $t_1/2 - t^*$ is at least a large absolute constant and $c = \Theta(\sqrt{\log(1/(\beta\gamma))})$. In particular, both $c - t^*$ and $t_1/2 - c$ are at least large absolute constants. See Figure 2.4a.

Thus, it holds that

$$\frac{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^-\}e^{2c\mathbf{x}_1}]}{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^+\}e^{2c\mathbf{x}_1}]} \geq \frac{4}{\beta}\,.$$

Similarly to the proof of Lemma 2.3.5, we now replace the exponential function $e^{2cx}$ by the square of the Taylor expansion of $e^{cx}$, i.e., $S_k^2(cx)$. It suffices to bound the following ratios

$$\frac{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[e^{2c\mathbf{x}_1}|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^+\}]}{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[(S_k(c\mathbf{x}_1))^2|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^+\}]} \geq \frac{1}{2} \quad \text{and} \quad \frac{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[(S_k(c\mathbf{x}_1))^2|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^-\}]}{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[e^{2c\mathbf{x}_1}|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^-\}]} \geq \frac{1}{2}\,. \tag{2.22}$$

We start by showing the first inequality of Equation (2.22). It suffices to show that

$$\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[|e^{2c\mathbf{x}_1} - S_k^2(c\mathbf{x}_1)||\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^+\}] \leq \mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[e^{2c\mathbf{x}_1}|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^+\}]\,.$$

Using the $L_1$ approximation error bound for the Taylor polynomial of Claim 2.3.7 and Hölder's inequality we can get an $L_2$ approximation guarantee. Assuming that $k$ is some sufficiently large absolute constant multiple of $c^2$, i.e., $k = C'c^2$, it holds

$$\begin{aligned}\|e^{2c\mathbf{x}_1} - S_k^2(c\mathbf{x}_1)\|_2 &\leq \|e^{2c\mathbf{x}_1} - S_k^2(c\mathbf{x}_1)\|_1^{1/3}\|e^{2c\mathbf{x}_1} - S_k^2(c\mathbf{x}_1)\|_4^{2/3} \\ &\leq e^{-C'k}e^{O(c^2)} \leq e^{-\Omega(k)}\,,\end{aligned} \tag{2.23}$$

where the last inequality follows from the fact that $S_k^2(c\mathbf{x}_1) \leq e^{2c|\mathbf{x}_1|}$, $\|e^{2c|\mathbf{x}_1|}\|_4 \leq e^{O(c^2)}$ and $C'$ is sufficiently large absolute constant. We can now use Cauchy-Schwarz and the

$L_2$ approximation guarantee of $S_k^2(c\mathbf{x}_1)$ to obtain

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[|e^{2c\mathbf{x}_1} - S_k^2(c\mathbf{x}_1)||\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^+\}] \\
&\quad\leq \|\mathbf{x}_1 - t_1\|_2 \|e^{2c\mathbf{x}_1} - S_k^2(c\mathbf{x}_1)\|_2 \\
&\quad\leq 2t_1 e^{-\Omega(k)}\,, \qquad (2.24)
\end{aligned}$$

where we used that $\|\mathbf{x}_1 - t_1\|_2 \leq 2t_1$ for $t_1 \geq 1$. We have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[e^{2c\mathbf{x}_1}|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^+\}] &\geq \mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}}[e^{2c\mathbf{x}_1}|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \leq t^*\}] \\
&\geq (t_1 - t^*)e^{-(2c-t^*)^2}/4 \geq t_1 e^{-4c^2}/8\,, \quad \triangleright\ t^* \leq c \text{ and } t^* \leq t_1/2
\end{aligned} \qquad (2.25)$$

where for the second inequality we used the Gaussian tail lower bounds of Fact 2.2.11. Combining Inequality (2.24) and Inequality (2.25), we obtain that for $k$ larger than some constant multiple of $c^2$ the first inequality of Equation (2.22) holds. The proof of the second inequality of Equation (2.22) is similar. We obtain that Inequality (2.19) holds for the polynomial $S_k^2(c\mathbf{x}_1)$ for $k = \Theta(1/\log(1/(\beta\gamma)))$ and $c = \Theta(1/\sqrt{\log(1/(\beta\gamma))})$.

Finally, using Fact 2.3.6 we have that $\|S_k^2(c\mathbf{x}_1)\|_2 \leq \|e^{2c|\mathbf{x}_1|}\|_2 = e^{O(c^2)}$. Using the estimate of Equation (2.21) and the $L_2$ approximation guarantee of Equation (2.23) we obtain that

$$\frac{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[S_k^2(c\mathbf{x}_1)|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^-\}]}{\|S_k^2(c\mathbf{x}_1)\|_2} = \frac{e^{-\Omega(c^2)} - e^{-\Omega(k)}}{e^{O(c^2)}} = e^{-O(c^2)} = \mathrm{poly}(\gamma\beta)\,,$$

where we used the fact that $k$ is a sufficiently large constant multiple of $c^2$ and $c^2 =$

$\Theta(\log(1/(\beta\gamma)))$. We conclude that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}\left[(\mathbf{x}_1 - t_1)\beta(\mathbf{x}_1,\mathbf{x}_2)\mathrm{sign}(\mathbf{x}_1 - t^*)\frac{S_k^2(c\mathbf{x}_1)}{\|S_k^2(c\mathbf{x}_1)\|_2}\right] \\
&\qquad \le \beta\frac{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[S_k^2(c\mathbf{x}_1)|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^-\}]}{\|S_k^2(c\mathbf{x}_1)\|_2} - \frac{\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[S_k^2(c\mathbf{x}_1)|\mathbf{x}_1 - t_1|\mathbb{1}\{\mathbf{x}_1 \in A^+\}]}{\|S_k^2(c\mathbf{x}_1)\|_2} \\
&\qquad \le -t_1\mathrm{poly}(\beta\gamma)\,.
\end{aligned}$$

The proof of the upper bound on the coefficients of the polynomial $S_k^2(c\mathbf{x}_1)/\|S_k^2(c\mathbf{x}_1)\|_2$ is similar to that of the constant hypothesis case, see Lemma 2.3.3.

□

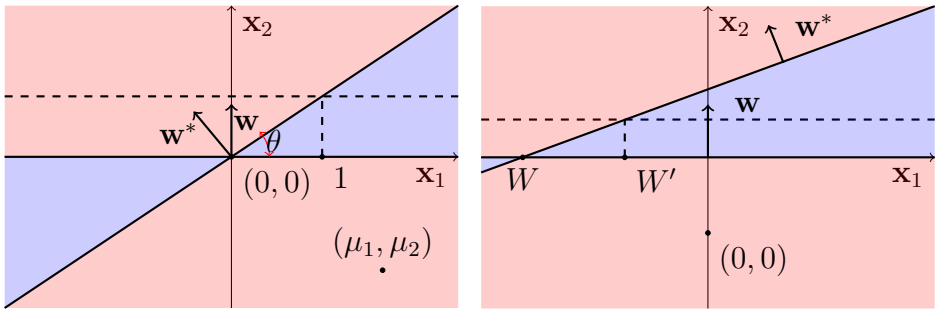
This completes the proof of Lemma 2.3.12. □

### Certificate Against "Small Threshold" Halfspaces

**Lemma 2.3.16** (Certificate against "Small Threshold" Hypotheses)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to some at most $(1-\gamma)$-biased optimal halfspace. Define the linear function $\ell(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} - t)$ and assume that $t/\|\mathbf{w}\|_2 \le C\sqrt{\log(1/(\beta\gamma))}$ for some absolute constant $C > 0$. Moreover, assume that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] \ge \mathrm{OPT} + \epsilon$. Then, there exists polynomial $q(\mathbf{x}) = \sum_{|\alpha|\le k} c_\alpha \mathbf{x}^\alpha$ of degree $\Theta(\log(\frac{1}{\beta}))$, norm $\|q(\mathbf{x})\|_2 = 1$, and sum of (absolute) coefficients $\sum_{|\alpha|\le k}|c_\alpha| \le d^{O(k)}$, and $r_1, r_2 \in \mathbb{R}$ with $|r_1 - r_2| \ge \mathrm{poly}(\epsilon\beta\gamma)$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y\mathbb{1}\{r_1 \le \ell(\mathbf{x}) \le r_2\}q^2(\mathbf{x})] \le -\epsilon^4\mathrm{poly}(\beta\gamma)\ \|\ell(\mathbf{x})\|_2\,.$$

*Proof.* Denote $\ell^*(\mathbf{x}) = \mathbf{w}^*\cdot\mathbf{x} - t^*$ the optimal halfspace and denote $\ell(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} - t$ for




(a) (b)

Figure 2.5: **(a)** Our certificate when the angle $\theta$ between the two halfspaces is large. Notice that we have changed coordinates so that the point $(0,0)$ is their crossing point. The mean of the Gaussian is now moved to the point $(\mu_1, \mu_2) = (-t/\tan(\theta) + t^*/\sin\theta, -t)$. Notice that since in this case we have assumed that $|t| \leq C\sqrt{\log(1/(\beta\gamma))}$, the $\mathbf{e}_2$-coordinate of the mean of the Gaussian cannot be very far from the origin. In this case, we overestimate the agreement region between the two halfspaces (red area) considering to be $\{\mathbf{x}_1 \leq 1\}$. In order to put more weight to the disagreement area, we again use the polynomial shift (see Lemma 2.3.5) in the direction $\mathbf{e}_1 = (\mathbf{w}^*)^{\perp_{\mathbf{w}}}$. Observe that this differs from the cases of Subsection 2.3.1 and Subsection 2.3.1, where the polynomial shift was along the direction of $\mathbf{w}^*$.
**(b)** The case where the angle $\theta$ between the two halfspaces is small, $\theta = O(\epsilon^2\gamma\beta)$. In that case, the crossing point $W$ of the two halfspaces is very far from the origin, $|W| \geq \Omega(1/(\epsilon\gamma\beta))$. In this case, simply taking a band around $\ell(\mathbf{x})$ works as a certificate.

some vectors $\mathbf{w} \in \mathbb{R}^d$ and some threshold $t \in \mathbb{R}$. Moreover, denote $\theta = \theta(\mathbf{w}^*, \mathbf{w})$ the angle between $\mathbf{w}^*$ and $\mathbf{w}$. We first observe that since $\mathrm{OPT} = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell^*(\mathbf{x}) \neq y]$ it holds that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] - \mathrm{OPT} = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq \mathrm{sign}(\ell^*(\mathbf{x}))\}\beta(\mathbf{x})] \,.$$

Thus, the disagreement probability between $\ell(\mathbf{x})$ and $\ell(\mathbf{x}^*)$ $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq \mathrm{sign}(\ell^*(\mathbf{x}))\}] \geq \epsilon$ from our assumption that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] \geq \mathrm{OPT} + \epsilon$. From

Fact 2.2.12 we know that

$$\Pr_{\mathbf{x}\sim\mathcal{D}_x}[\mathrm{sign}(\ell(\mathbf{x})) \neq \mathrm{sign}(\ell^*(\mathbf{x}))] \leq O(\theta) + O(|t^* - t|)\,.$$

Thus, we either have that the angle of the two halfspaces is large $\theta = \Omega(\epsilon)$ or the difference of their thresholds $|t^* - t| = \Omega(\epsilon)$.

Since we can always write

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathbf{w}\cdot\mathbf{x} - t)y\ \mathbb{1}\{r_1 \leq \ell(\mathbf{x}) \leq r_2\}q^2(\mathbf{x})] \\ &\qquad = \|\mathbf{w}\|_2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left(\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\cdot\mathbf{x} - \frac{t}{\|\mathbf{w}\|_2}\right) y\ \mathbb{1}\{r_1 \leq \ell(\mathbf{x}) \leq r_2\}q^2(\mathbf{x})\right],\end{aligned}$$

we will assume for simplicity that $\|\mathbf{w}\|_2 = 1$ and $|t| \leq C\sqrt{\log(1/(\beta\gamma))}$. We are going to construct a polynomial $q(\mathbf{x})$ that only depends on the subspace $V$ spanned by $\mathbf{w}, \mathbf{w}^*$. Therefore, as in the case of Lemma 2.3.3, we can project $\mathbf{x}$ to the subspace $V$ spanned by $\mathbf{w}, \mathbf{w}^*$ and preserve the $\eta$-Massart noise assumption, see Claim 2.3.4. Let $\theta = \theta(\mathbf{w}, \mathbf{w}^*)$ be the angle between $\mathbf{w}, \mathbf{w}^*$. Without loss of generality, we may assume that $\mathbf{w} = \mathbf{e}_2$ and $\mathbf{w}^* = -\sin\theta\mathbf{e}_1 + \cos\theta\mathbf{e}_2$, see Figure 2.5a. Therefore, in this case we have that $\ell(\mathbf{x}) = \mathbf{x}_2 - t$ and $\ell^*(\mathbf{x}) = -\sin\theta\mathbf{e}_1 + \cos\theta\mathbf{e}_2 - t^*$.

We first provide a certificate that works when the angle $\theta$ is non-trivial, that is when $\theta \geq C'\epsilon^2\gamma\beta$, for some small enough constant $C'$ chosen appropriately. We change the coordinates so that the crossing point of the two linear functions $\ell, \ell^*$ is the origin $(0, 0)$. This will move the mean of the Gaussian to the point $(\mu_1, \mu_2) = (-t/\tan\theta + t^*/\sin\theta, -t)$, see Figure 2.5a.

Assume first that $C\epsilon^2\gamma\beta = \theta \leq \pi/2$ and $\mu_1 \geq 0$, we show later that the other cases are symmetric. Set $r_1 = 0$, $r_2 = \tan(\theta) = \Theta(\theta)$, and let $q(\mathbf{x})$ depend only on $\mathbf{x}_1$, we

have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y\ \mathbb{1}\{r_1 \le \ell(\mathbf{x}) \le r_2\}q^2(\mathbf{x})] \\
&\quad = \mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}(\mu_1,\mu_2)}[\mathbf{x}_2\mathrm{sign}(-\sin\theta\mathbf{x}_1 + \cos\theta\mathbf{x}_2)\ \beta(\mathbf{x}_1,\mathbf{x}_2)\mathbb{1}\{0 \le \mathbf{x}_2 \le r_2\}q^2(\mathbf{x}_1)] \\
&\quad \le \mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}(\mu_1,\mu_2)}[\mathbf{x}_2\mathrm{sign}(1-\mathbf{x}_1)\ \beta(\mathbf{x}_1,\mathbf{x}_2)\ \mathbb{1}\{0 \le \mathbf{x}_2 \le r_2\}q^2(\mathbf{x}_1)]\,, \qquad (2.26)
\end{aligned}$$

where for the inequality we used the fact that $0 \le \mathbf{x}_2 \le r_2 = \tan(\theta)$, and therefore it holds that if $-\sin\theta\mathbf{x}_1 + \cos\theta\mathbf{x}_2 \ge 0$ then it holds that $1 - \mathbf{x}_1 \ge 0$, for all $(\mathbf{x}_1, \mathbf{x}_2)$ such that $\mathbf{x}_2 \in [0, r_2]$. Observe now that in order to maximize the quantity of Equation (2.26) the "worst-case" noise function $\beta(\mathbf{x}_1, \mathbf{x}_2)$ is equal to $\beta$ for all points where the integral is negative, and $+1$ when the integral is positive, that is $\beta(\mathbf{x}_1, \mathbf{x}_2) = \beta\mathbb{1}\{\mathbf{x}_1 \ge 1\} + \mathbb{1}\{\mathbf{x}_1 < 1\}$. Since this "worst-case" noise $\beta(\mathbf{x}_1, \mathbf{x}_2)$ is independent of $\mathbf{x}_2$, it follows that we can decompose the expectation of Equation (2.26), i.e.,

$$\mathop{\mathbf{E}}_{\mathbf{x}_1\sim\mathcal{N}(\mu_1)}[q^2(\mathbf{x}_1)\mathrm{sign}(1-\mathbf{x}_1)\beta(\mathbf{x}_1)] \mathop{\mathbf{E}}_{\mathbf{x}_2\sim\mathcal{N}(\mu_2)}[\mathbf{x}_2\mathbb{1}\{0 \le \mathbf{x}_2 \le r_2\}]\,.$$

Since $|\mu_2| = |t| \le C\sqrt{\log(1/(\beta\gamma))}$ from standards bounds of the Gaussian tail probability (Fact 2.2.11), we obtain

$$\mathop{\mathbf{E}}_{\mathbf{x}_2\sim\mathcal{N}(\mu_2)}[\mathbf{x}_2\mathbb{1}\{0 \le \mathbf{x}_2 \le r_2\}] \ge \theta^2\mathrm{poly}(\beta\gamma)\,.$$

It remains to bound the term $\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}(\mu_1)}[q^2(\mathbf{x}_1)\mathrm{sign}(1-\mathbf{x}_1)\beta(\mathbf{x}_1)]$. We observe that the "worst-case" value of $\mu_1 \ge 0$ in order to maximize this expectation is $\mu_1 = 0$. Now, we can use the same argument as in the proof of Lemma 2.3.3; notice that in this case the threshold is 1 instead of $\sqrt{\log(1/\gamma)}$ and therefore, by picking $k = \Theta(\log(1/(\beta)))$ we have

that there exists polynomial $q$ such that $\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}(\mu_1)}[q^2(\mathbf{x}_1)\mathrm{sign}(1-\mathbf{x}_1)\beta(\mathbf{x}_1)] \leq -\mathrm{poly}(\beta\gamma)$ and thus, we obtain the bound:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y\ \mathbb{1}\{r_1 \leq \mathbf{w}\cdot\mathbf{x} \leq r_2\}q^2(\mathbf{x})] \leq -\theta^2\mathrm{poly}(\beta\gamma)\|\ell(\mathbf{x})\|_2\,.$$

In the case where $\theta \in [0,\pi/2]$ and $\mu_1 < 0$, we may pick $r_1 = -\tan(\theta)$ and $r_2 = 0$ resulting in a completely symmetric case to the previous one. Finally, the case where $\theta \in [\pi/2,\pi]$ is easier than the previous two cases since the disagreement region between the two halfspaces is now a superset of the corresponding region in the previous cases.

We now handle the case where the angle between the two halfspaces is small, i.e., $\theta \leq C'\epsilon^2\beta\gamma$. We know that the disagreement between two halfspaces is upper bounded by their angle, i.e., $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})] \leq O(\theta)+O(|t-t^*|)$. Since $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq f(\mathbf{x})] \geq \epsilon$, we obtain that the two thresholds $t, t^*$ cannot be very close, i.e., $|t-t^*| = \Omega(\epsilon)$. As in the previous case, we may assume that $\mathbf{w} = \mathbf{e}_2$ and $\mathbf{w}^* = -\sin\theta\mathbf{e}_1 + \cos\theta\mathbf{e}_2$. Recall that the intersection point of the two halfspaces has coordinates $(t/\tan\theta - t^*/\sin\theta, t)$. This means that when $\theta = O(\epsilon^2\gamma\beta)$ the intersection point of the two halfspaces is very far from the origin: its first coordinate is $|t/\tan\theta - t^*/\sin\theta| = \Omega(1/(\epsilon\gamma\beta))$, since by the triangle inequality, it follows

$$\begin{aligned}|t\cos\theta - t^*| &\geq |t-t^*|\cos\theta - |t^*||\cos\theta - 1| \\ &\geq |t-t^*|(1-\epsilon^4\gamma^2\beta^2) - O(\sqrt{\log(1/(\beta\gamma))})(\epsilon^4\gamma^2\beta^2 - 2\epsilon^8\gamma^4\beta^4) = \Omega(\epsilon)\,,\end{aligned}$$

where we used the inequality $1-x^2 \leq \cos x \leq 1 - x^2/2 + x^4$, for all $x \in [0,\pi/2]$ and the fact that $\epsilon \leq 1/2$. Combining, the above with $\sin(\theta) = O(\epsilon^2\beta\gamma)$, we get that $|t/\tan\theta - t^*/\sin\theta| = \Omega(1/(\epsilon\gamma^2\beta^2))$. In this case, we do not require a polynomial for

the certificate, therefore the certificate is going to be simply a band, see Figure 2.5b. Let $W$ be the $\mathbf{e}_1$-coordinate of the crossing point of the optimal halfspace with $\ell(\mathbf{x})$, see Figure 2.5b. We have that $|W| = \Omega(1/(\epsilon\beta\gamma))$. Without loss of generality, we may assume that $W < 0$ along with the band $\epsilon^2/2 \leq \ell(\mathbf{x}) \leq \epsilon^2$ since the proof of the other case is similar: we just need to consider the band $-\epsilon^2 \leq \ell(\mathbf{x}) \leq -\epsilon^2/2$ instead, see Figure 2.5b. Moreover, denote $W'$ the $\mathbf{e}_1$-coordinate of the point of $\ell^*(x) = \ell(\mathbf{x}) + \epsilon^2$, see Figure 2.5b. It holds $W' = W + \epsilon^2/\tan\theta$, and note that $|W'| = \Omega(1/(\epsilon\beta\gamma))$ since $|W + \epsilon^2/\tan\theta| \geq (|t\cos\theta - t^*| - |\epsilon^2|)/\sin\theta \geq \Omega(1/(\epsilon\beta\gamma))$. We have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\ell(\mathbf{x})\mathrm{sign}(-\mathbf{x}_1\sin\theta + \mathbf{x}_2\cos\theta - t^*)\mathbb{1}\{\epsilon^2/2 \leq \ell(\mathbf{x}) \leq \epsilon^2\}\beta(\mathbf{x}_1,\mathbf{x}_2)] \\
&\qquad \leq \mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\ell(\mathbf{x})\mathrm{sign}(-\mathbf{x}_1 + W + \epsilon^2/\tan\theta)\mathbb{1}\{\epsilon^2/2 \leq \ell(\mathbf{x}) \leq \epsilon^2\}\beta(\mathbf{x}_1,\mathbf{x}_2)]\,,
\end{aligned} \tag{2.27}$$

where we overestimated the contribution of the agreement area (red region in Figure 2.5b) by the region $\mathbf{x}_1 \leq W'$. Since $W'$ is still very far from the origin the contribution of the region $\mathbf{x}_1 \leq W'$ is going to be very small. To bound the quantity of Equation (2.27), we first bound from above the region where $\mathrm{sign}(-\mathbf{x}_1 + W')$ is positive:

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\ell(\mathbf{x})\mathbb{1}\{\epsilon^2/2 \leq \ell(\mathbf{x}) \leq \epsilon^2\}\beta(\mathbf{x})\mathbb{1}\{\mathbf{x}_1 \leq W'\}] \leq \mathop{\mathbf{Pr}}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbf{x}_1 \leq W'] \leq e^{-1/(\epsilon\beta\gamma)^2}\,, \tag{2.28}$$

where we used the fact that $|W'| = \Omega(1/(\epsilon\gamma\beta))$. Next, we bound from below the region

where $\mathrm{sign}(-\mathbf{x}_1 + W')$ is negative:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\ell(\mathbf{x})\mathbb{1}\{\epsilon^2/2 \le \ell(\mathbf{x}) \le \epsilon^2\}\beta(\mathbf{x})\mathbb{1}\{\mathbf{x}_1 \ge W'\}] \\
&\qquad \ge \beta \mathop{\mathbf{E}}_{\mathbf{x}_2\sim\mathcal{N}}[\ell(\mathbf{x}_2)\mathbb{1}\{\epsilon^2/2 \le \ell(\mathbf{x}_2) \le \epsilon^2\}] \mathop{\mathbf{Pr}}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbf{x}_1 \ge W'] \\
&\qquad \ge \frac{\epsilon^2}{4}\beta \mathop{\mathbf{Pr}}_{\mathbf{x}_2\sim\mathcal{N}}[\epsilon^2/2 \le \ell(\mathbf{x}_2) \le \epsilon^2] \ge \epsilon^4\beta\mathrm{poly}(\beta\gamma)\ , \qquad (2.29)
\end{aligned}$$

where we used that $\mathbf{Pr}_{\mathbf{x}_1\sim\mathcal{N}}[\mathbf{x}_1 \ge W'] \ge 1/2$ and that $\beta(\mathbf{x}) \ge \beta$. Using Equation (2.28) and Equation (2.29) to Equation (2.27), we get

$$\mathop{\mathbf{E}}_{(\mathbf{x}_1,\mathbf{x}_2)\sim\mathcal{N}_2}[\ell(\mathbf{x})\mathrm{sign}(-\mathbf{x}_1\sin\theta+\mathbf{x}_2\cos\theta-t^*)\mathbb{1}\{\epsilon^2/2 \le \ell(\mathbf{x}) \le \epsilon^2\}\beta(\mathbf{x}_1,\mathbf{x}_2)] \le -\epsilon^4\mathrm{poly}(\beta\gamma)\ .$$

We combine the above cases to obtain the claimed bound. □

### 2.3.2 Efficiently Computing the Certificate via SDP

In this section, we show that we can efficiently compute our polynomial certificate given labeled examples from the target distribution. The following is the main proposition of this subsection, where we bound the number of samples and the runtime needed to compute the certificate given samples from the distribution $\mathcal{D}$. The proof is similar to that given in [DKTZ20b]; we adapt it to work for our certifying function and for general (as opposed to homogeneous) halfspaces.

**Proposition 2.3.17** (Certificate Oracle)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to some (at least) $(1-\gamma)$-biased optimal halfspace $f(\mathbf{x})$. Let $\ell(\mathbf{x})$ be any linear function such that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathrm{sign}(\ell(\mathbf{x})) \ne f(\mathbf{x})] \ge \epsilon$, for $\epsilon \in (0,1)$. There exists*

*an algorithm that draws* $N = d^{O(\log(1/(\beta\gamma)))}\log(1/\delta)/\epsilon^2$ *samples from* $\mathcal{D}$*, runs in time* $\mathrm{poly}(N,d)$*, and with probability* $1-\delta$ *returns a positive function* $T(\mathbf{x})$ *with* $\|T(\mathbf{x})\|_4 \le 1$ *such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T(\mathbf{x})\ell(\mathbf{x})y] \le -\epsilon^4 d^{-O(\log(1/(\beta\gamma)))}\|\ell(\mathbf{x})\|_2\,.$$

*Proof.* From Proposition 2.3.2, we know that we are looking for a certificate function $T(\mathbf{x})$ of the form $T(\mathbf{x}) = \mathbb{1}_B(\mathbf{x})q^2(\mathbf{x})$, where $B$ is a band with respect $\ell(\mathbf{x})$, i.e., $B = \{r_1 \le \ell(\mathbf{x}) \le r_2\}$ for some $r_1, r_2 \in \mathbb{R}\cup\{\infty\}$ and $q(\mathbf{x})$ is a $k = \Theta(\log(1/\gamma\beta))$ degree polynomial. We illustrate how we can formulate the search of such function as an SDP. For the rest of this section, let $\mathbb{1}_B(\mathbf{x})$ be the indicator function of the region $B = \{\mathbf{x} : r_1 \le \ell(\mathbf{x}) \le r_2\}$, for some appropriate choices $r_1, r_2 \in \mathbb{R}\cup\{\infty\}$ and $\lambda = \epsilon^4 d^{-O(k)}\|\ell(\mathbf{x})\|_2$. Denote by $\mathbf{m}(\mathbf{x})$ the vector containing all monomials up to degree $k$, such that $\mathbf{m}_S(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{x}^S$, indexed by the multi-index $S$ satisfying $|S| \le k$. Recall that if $S = (s_1, \ldots, s_d)$, then $\mathbf{x}^S = \prod_{i=1}^d \mathbf{x}_i^{s_i}$. The dimension of $\mathbf{m}(\mathbf{x}) \in \mathbb{R}^m$ is $m = \binom{d+k}{k}$. Let $\mathbf{M} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{m}(\mathbf{x})\mathbf{m}(\mathbf{x})^\intercal \mathbb{1}_B(\mathbf{x})\ell(\mathbf{x})y]$, for any real matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$, we define the following function

$$\mathcal{L}_\ell(\mathbf{A}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{m}(\mathbf{x})^\intercal \mathbf{A}\ \mathbf{m}(\mathbf{x})\mathbb{1}_B(\mathbf{x})\ell(\mathbf{x})y] = \mathrm{tr}\left(\mathbf{A}\mathbf{M}\right)\ . \tag{2.30}$$

Notice that $\mathcal{L}_\ell$ is linear in its variable $\mathbf{A}$. From Proposition 2.3.2, we know that if $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathrm{sign}(\ell(\mathbf{x})) \ne f(\mathbf{x})] \ge \epsilon$, then there exists a (normalized) polynomial $q(\mathbf{x}) = \mathbf{m}(\mathbf{x})\cdot\mathbf{a}$, with $\|\mathbf{a}\|_1 = 1$ such that $\mathbf{E}_{(\mathbf{x},y)}[\ell(\mathbf{x})y\mathbb{1}_B(\mathbf{x})q(\mathbf{x})] \le -\lambda$. Therefore, for $\mathbf{A} = \mathbf{a}\mathbf{a}^\intercal$, we have $q^2(\mathbf{x}) = \mathbf{m}(\mathbf{x})^\intercal\mathbf{A}\ \mathbf{m}(\mathbf{x})$ and hence, $\mathcal{L}_\ell(\mathbf{A}) \le -\lambda$. It follows that there exists a positive semi-definite rank-1 matrix $\mathbf{A}$ such that $\mathcal{L}_\ell(\mathbf{A}) \le -\lambda$. Moreover, $\|\mathbf{A}\|_F^2 = \left\|\mathbf{a}\mathbf{a}^T\right\|_F = \|\mathbf{a}\|_2^4 \le 1$. Recall that, $\mathcal{S}^m$ is the set of (symmetric) positive semi-

definite matrices of $m$-dimension. We formulate the following semi-definite program

$$\begin{aligned} \text{Find} \quad & \mathbf{A} \in \mathcal{S}^m \\ \text{s. t.} \quad & \operatorname{tr}(\mathbf{AM}) \leq -\lambda \\ & \|\mathbf{A}\|_F^2 \leq 1 \end{aligned} \tag{2.31}$$

Moreover, from Proposition 2.3.2, the SDP (2.31) is feasible if $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\operatorname{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})] \geq \epsilon$. We define $\widetilde{\mathbf{M}} = \frac{1}{N}\sum_{i=1}^N \mathbf{m}(\mathbf{x}^{(i)})\mathbf{m}(\mathbf{x}^{(i)})^\intercal \mathbb{1}_B(\mathbf{x}^{(i)}) y^{(i)} \ell(\mathbf{x}^{(i)})$, the empirical estimate of $\mathbf{M}$ using $N$ samples from $\mathcal{D}$. Using the following fact, we bound the sample size required so that $\widetilde{\mathbf{M}}$ is sufficiently close to $\mathbf{M}$ which is similar to Lemma 3.8 of [DKTZ20b].

**Fact 2.3.18** (Estimation of $\mathbf{M}$)**.** *Let* $\Omega = \{\mathbf{A} \in \mathcal{S}^m : \|\mathbf{A}\|_F \leq 1\}$ *and* $\epsilon, \delta \in (0,1)$*. Let* $\ell(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + t$ *with* $|\ell(\mathbf{x})|^2 \leq C$ *and* $\widetilde{\mathbf{M}} = \frac{1}{N}\sum_{i=1}^N \mathbf{m}(\mathbf{x}^{(i)})\mathbf{m}(\mathbf{x}^{(i)})^\intercal \mathbb{1}_B(\mathbf{x}^{(i)}) y^{(i)} \ell(\mathbf{x}^{(i)})$*. There exists an algorithm that draws* $N = \frac{d^{O(\log \frac{1}{\gamma\beta})}}{C\epsilon^2}\log(1/\delta)$ *samples from* $\mathcal{D}$*, runs in* $\operatorname{poly}(N,d)$ *time and with probability at least* $1-\delta$ *outputs a matrix* $\widetilde{\mathbf{M}}$ *such that*

$$\mathbf{Pr}\left[\sup_{\mathbf{A}\in\Omega} \left|\operatorname{tr}(\mathbf{A}\widetilde{\mathbf{M}}) - \operatorname{tr}(\mathbf{AM})\right| \geq \epsilon\right] \leq 1-\delta\,.$$

*Proof.* Using the Cauchy-Schwarz inequality, we get

$$\operatorname{tr}\left(\mathbf{A}(\mathbf{M}-\widetilde{\mathbf{M}})\right) \leq \|\mathbf{A}\|_F \left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F\,.$$

Therefore, it suffices to bound the probability that $\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F \geq \epsilon$. From Markov's

inequality, we have

$$\mathbf{Pr}\left[\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F \geq \epsilon\right] \leq \frac{1}{\epsilon^2}\,\mathbf{E}\left[\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F^2\right]. \tag{2.32}$$

Using multi-indices $S_1$, $S_2$ that correspond to the monomials $\mathbf{x}^{S_1}, \mathbf{x}^{S_2}$ (as indices of the matrix $\mathbf{M}$), we have

$$\mathbf{E}\left[\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F^2\right] = \sum_{S_1,S_2:|S_1|,|S_2|\leq k} (\mathbf{M}_{S_1,S_2}-\widetilde{\mathbf{M}}_{S_1,S_2})^2 = \sum_{S_1,S_2:|S_1|,|S_2|\leq k} \mathbf{Var}[\widetilde{\mathbf{M}}_{S_1,S_2}]\,.$$

Using the fact that the samples $(\mathbf{x}^{(i)}, y^{(i)})$ are independent, we can bound from above the variance of each entry $(S_1, S_2)$ of $\widetilde{\mathbf{M}}$

$$\begin{aligned}\mathbf{Var}[\widetilde{\mathbf{M}}_{S_1,S_2}] &\leq \frac{1}{N}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbf{x}^{2(S_1+S_2)}\left(\mathbb{1}_B(\mathbf{x})\ell(\mathbf{x})y\right)^2\right] \leq \frac{2}{N}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}^{2(S_1+S_2)}(\|\mathbf{x}\|_2^2+t^2)\right]\\ &\leq \frac{2C}{N}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\|\mathbf{x}\|_2^2)^{|S_1+S_2|+1}\right].\end{aligned}$$

For every $n \geq 1$, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\|\mathbf{x}\|_2^2)^n\right] = d^{O(n)}$. Using the above bound for the variance and summing over all pairs $S_1, S_2$ with $|S_1|, |S_2| \leq k$, we obtain

$$\mathbf{E}\left[\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F^2\right] = \frac{2C}{N}d^{O(k)} \tag{2.33}$$

Combining Equation (2.32) and Equation (2.33) we obtain that with $N = d^{O(m)}/(C\epsilon^2)$ samples we can estimate $\mathbf{M}$ within the target accuracy with probability at least $3/4$. To amplify the probability to $1-\delta$, we can simply use the above empirical estimate $\ell$ times to obtain estimates $\widetilde{\mathbf{M}}^{(1)}, \ldots, \widetilde{\mathbf{M}}^{(\ell)}$ and keep the coordinate-wise median as our final estimate. It follows that $O(\log(m/\delta))$ repetitions suffice to guarantee confidence

probability at least $1 - \delta$.

□

Using Fact 2.3.18, we replace the matrix $\mathbf{M}$ in Equation (2.30) with the estimate $\widetilde{\mathbf{M}}$ and define the following "empirical" SDP

$$\begin{aligned} \text{Find} \quad & \mathbf{A} \in \mathcal{S}^m \\ \text{such that} \quad & \operatorname{tr}(\mathbf{A}\widetilde{\mathbf{M}}) \leq -\lambda/2 \\ & \|\mathbf{A}\|_F^2 \leq 1 \end{aligned} \tag{2.34}$$

From Fact 2.3.18, we obtain that with $N$ samples we can get a matrix $\widetilde{\mathbf{M}}$ such that $|\operatorname{tr}(\mathbf{A}\widetilde{\mathbf{M}} - \operatorname{tr}(\mathbf{A}\mathbf{M})| \leq -\lambda/2$ with probability at least $1 - \delta'$. From Proposition 2.3.2, we know that with the given bound for $k$ and $\|\mathbf{A}\|_F$, there exists $\mathbf{A}^*$ such that

$$\operatorname{tr}(\mathbf{A}^*\mathbf{M}) \leq -\lambda.$$

Therefore, the SDP (2.31) is feasible. Moreover, from Fact 2.3.18 we get that

$$\operatorname{tr}(\mathbf{A}^*\widetilde{\mathbf{M}}) \leq -\lambda/2 \, .$$

Thus, the SDP (2.34) is feasible. Since the dimension of the matrix $\mathbf{A}$ is smaller than the number of samples, we have that the runtime of the SDP is polynomial in the number of samples. Solving the SDP with tolerance $\lambda/4$, we obtain an almost feasible $\widetilde{\mathbf{A}}$, in the sense that $\operatorname{tr}(\widetilde{\mathbf{A}}\widetilde{\mathbf{M}}) \leq -\lambda/4$. Using again the guarantee of Fact 2.3.18, we get that solving the SDP (2.34), we obtain a positive-semi definite matrix $\widetilde{\mathbf{A}}$ such that $\operatorname{tr}(\widetilde{\mathbf{A}}\mathbf{M}) \leq -\lambda/4$. Moreover, we have that for the matrix $\mathbf{A}$ returned by our SDP it

holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{m}(\mathbf{x})^T\mathbf{A}\mathbf{m}(\mathbf{x}))^4] \leq \|\mathbf{A}\|_F^4 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{m}(\mathbf{x})^T\mathbf{m}(\mathbf{x}))^4] = d^{\log(1/(\beta\gamma))} \tag{2.35}$$

To complete the proof, we need to show how to guess the band $B$. For some large enough constant $C > 0$, let $\mathcal{T} = \{\pm\lambda^2, \pm 2\lambda^2, \ldots, C\sqrt{\log(1/\lambda)}\}$. Assume that for some $B = \{\mathbf{x} : r_1 \leq \ell(\mathbf{x}) \leq r_2\}$, with $r_1, r_2 \in \mathbb{R}$ and some polynomial $q(\mathbf{x})$ with $\|q^2(\mathbf{x})\|_2 \leq 1$, such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}_B(\mathbf{x})q^2(\mathbf{x})\ell(\mathbf{x})y\right] \leq -\lambda/4$. Then, there exists $\tilde{r}_1, \tilde{r}_2 \in \mathcal{T}$, with $|r_1 - \tilde{r}_1| \leq \lambda^2$ and $|r_2 - \tilde{r}_2| \leq \lambda^2$, such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}\{\tilde{r}_1 \leq \ell(\mathbf{x}) \leq \tilde{r}_2\}q^2(\mathbf{x})\ell(\mathbf{x})y\right] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}_B(\mathbf{x})q^2(\mathbf{x})\ell(\mathbf{x})y\right] + 2\lambda^2 \leq -\lambda/4\,,$$

where we used Cauchy–Schwarz inequality and the Gaussian concentration. Thus, by setting $\delta' = \Theta(\delta\lambda^4/\log(1/\lambda))$ solving the SDP (2.34) for all the different choices of $r_1, r_2 \in \mathcal{T}$, with $r_1 \leq r_2$, we guarantee that the algorithm will return a polynomial $q(\mathbf{x})$ and some thresholds $\tilde{r}_1, \tilde{r}_2 \in \mathcal{T}$, such that $T(\mathbf{x}) = \mathbb{1}\{\tilde{r}_1 \leq \ell(\mathbf{x}) \leq \tilde{r}_2\}q^2(\mathbf{x})$ is a certifying function, i.e., it holds

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}\{\tilde{r}_1 \leq \ell(\mathbf{x}) \leq \tilde{r}_2\}q^2(\mathbf{x})\ell(\mathbf{x})y\right] \leq -\lambda/4\,,$$

with probability $1-\delta$. Finally, using Inequality (2.35) we obtain that the $L_4$ norm of $T(\mathbf{x})$ is bounded from above as $\|T(\mathbf{x})\|_4 \leq d^{O(\log(1/(\beta\gamma)))}$. Thus, the function $T(\mathbf{x})/\|T(\mathbf{x})\|_4$ is an $\epsilon^4 d^{-O(\log(1/(\beta\gamma)))}$-certificate.

□

### 2.3.3 Learning a Near-Optimal Halfspace via Online Convex Optimization

In this subsection, we present a black-box approach to learn halfspaces with $\eta$-Massart Noise given a $\rho$-certifying oracle. A similar reduction for homogeneous halfspaces based on online-convex optimization was given in [DKK+20]. Here we adapt it so that it handles non-homogeneous halfspaces. Another difference is that in [DKK+20] the certificate function was bounded in the $L_\infty$ sense. Here we have certificates bounded in the $L_4$ norm. The arguments however are similar, and we provide them here for completeness. Formally, we prove:

**Proposition 2.3.19.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to some halfspace. Fix $\epsilon, \delta \in (0,1)$. Given a $\rho$-certificate oracle $\mathcal{O}$ that returns certifying functions with bounded $\ell_4$-norm. There exists an algorithm that makes $T = O(\frac{d \log(1/\epsilon)}{\rho^2 \epsilon^4})$ calls to $\mathcal{O}$, draws $N = \tilde{O}(\frac{d}{\rho^2 \epsilon^4} \log(1/\delta))$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(d, N, T)$ time and computes a hypothesis $h$, such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, with probability $1 - \delta$.*

We will require the following standard regret bound from online convex optimization.

**Lemma 2.3.20** (see, e.g., Theorem 3.1 of [Haz16])**.** *Let $\mathcal{V} \subseteq \mathbb{R}^n$ be a non-empty closed convex set with diameter $K$. Let $r_1, \ldots, r_T$ be a sequence of $T$ convex functions $r_i : \mathcal{V} \mapsto \mathbb{R}$ differentiable in open sets containing $\mathcal{V}$, and let $G = \max_{i\in[T]} \|\nabla_{\mathbf{w}} r_i\|_2$. Pick any $\mathbf{w}^{(1)} \in \mathcal{V}$ and set $\eta_i = \frac{K}{G\sqrt{t}}$ for $i \in [T]$. Then, for all $\mathbf{u} \in \mathcal{V}$, we have that $\sum_{i=1}^T (r_i(\mathbf{w}^{(t)}) - r_i(\mathbf{u})) \leq \frac{3}{2} GK\sqrt{T}$.*

**Algorithm 1** Learning Halfspaces with $\eta$-Massart noise.

**Input:**

1. $\epsilon, \delta > 0$.
2. A distribution $\mathcal{D}$ that satisfies the $\eta$-Massart Noise condition.
3. Access to a $\rho$-certificate oracle $\mathcal{O}$

**Output:** A vector $\mathbf{w}$ and threshold $t \in \mathbb{R}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t) \neq y] \leq \mathrm{OPT} + \epsilon$

**Define:** $T = C\ d\log(1/\epsilon)/(\rho\epsilon)^2$, $N = C\ d/\epsilon^2 \log(1/\epsilon)\log(T/\delta)$, for some large enough constant $C > 0$, $\mathcal{V} = \{(\mathbf{w}, t) \in \mathbb{R}^{d+1} : \|\mathbf{w}\|_2 \leq 1\}, |t| \leq 4\sqrt{\log(1/\epsilon)}\}$

1. $\mathbf{w}^{(0)} \leftarrow \mathbf{e}_1$, $t^{(0)} \leftarrow 0$
2. For $i \in [T]$ do
   (a) $\eta_i \leftarrow 1/(\sqrt{i} + \rho)$
   (b) If $(\mathbf{w}^{(i)}, t^{(i)}) = \mathbf{0}$ then $T^{(i)}(\mathbf{x}) = 0$.
   (c) Else let ANS $\leftarrow \mathcal{O}((\mathbf{w}^{(i)}, t^{(i)}))$.
   (d) If ANS = FAIL then **return** $(\mathbf{w}^{(i)}, t^{(i)})$ else $T^{(i)} \leftarrow$ ANS
   (e) Let $\nabla\hat{r}_i(\mathbf{w}, t)$ be an estimator of $-\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left(T^{(i)}(\mathbf{x}) + \rho\right) y(\mathbf{x}, 1)\right]$ (Lemma 2.3.22)
   (f) $(\mathbf{w}^{(i+1)}, t^{(i+1)}) \leftarrow \mathrm{proj}_{\mathcal{V}}\left((\mathbf{w}^{(i)}, t^{(i)}) - \eta_i \nabla\hat{r}_i(\mathbf{w}^{(i)}, t^{(i)})\right)$
3. **return** $(\mathbf{w}^{(T+1)}, t^{(T+1)})$

We show below that the optimal vector $\mathbf{w}^*$ and threshold $t^*$ and our current candidate vector $\mathbf{w}^{(i)}$ and threshold $t^{(i)}$ have a separation in the value of $r_i$.

**Lemma 2.3.21** (Error of $r_i$)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Assume that $\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to the optimal halfspace* $\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x} + t^*)$*. Let $\mathbf{w}^{(i)}$ with $\|\mathbf{w}^{(i)}\| \leq 1$ and $t^{(i)} \in \mathbb{R}$. Fix*

*$\rho \in (0,1)$ and let $r_i(\mathbf{w},t) = -\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(T^{(i)}(\mathbf{x})+\rho)y(\mathbf{x},1)]\cdot(\mathbf{w},t)$, where $T^{(i)}(\mathbf{x})$ is a non-negative function returned from a $(2\rho)$-certificate oracle, we have that*

$$r_i(\mathbf{w}^{(i)},t^{(i)}) - r_i(\mathbf{w}^*,t^*) \geq \rho\epsilon^2/2 \,.$$

*Proof.* Let $\ell^{(i)}(\mathbf{x}) = \mathbf{w}^{(i)}\cdot\mathbf{x}+t^{(i)}$. Using the fact that $T^{(i)}(\mathbf{x})\geq 0$, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T^{(i)}(\mathbf{x})(\mathbf{w}^*\cdot\mathbf{x}+t^*)y] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[T^{(i)}(\mathbf{x})|\mathbf{w}^*\cdot\mathbf{x}+t^*|\beta(\mathbf{x})] > 0 \,.$$

Therefore, we have that for every $i\in[T]$, it holds $r_i(\mathbf{w}^*,t^\star) \leq -\rho\,\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^*\cdot\mathbf{x}+t^*|\beta(\mathbf{x})]$. Let $I = \{\mathbf{x} : \mathbf{w}^*\cdot\mathbf{x}\in(-t^*-\epsilon/2, -t^*+\epsilon/2)\}$ and note that it should hold that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})\neq y] = \mathrm{OPT} \leq 1/2-\epsilon$, otherwise the $(2\rho)$-certifying oracle would not be able to return a function, and therefore $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})]\geq 2\epsilon$. We have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^*\cdot\mathbf{x}+t^*|\beta(\mathbf{x})] &\geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^*\cdot\mathbf{x}+t^*|\{\mathbf{x}\notin I\}\beta(\mathbf{x})] \geq \frac{\epsilon}{2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\{\mathbf{x}\notin I\}\beta(\mathbf{x})] \\ &= \frac{\epsilon}{2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})] - \frac{\epsilon}{2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\{\mathbf{x}\in I\}\beta(\mathbf{x})] \geq \epsilon^2/2 \,.\end{aligned}$$

where in the last inequality we used that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})]\geq 2\epsilon$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\{\mathbf{x}\in I\}]\leq\epsilon$ from Fact 2.2.11.

It remains to bound from below $r_i(\mathbf{w}^{(i)},t^{(i)})$. Using the fact that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[T^{(i)}(\mathbf{x})(\mathbf{w}^{(i)}\cdot\mathbf{x}+t^{(i)})y] \leq -2\rho\|\ell^{(i)}(\mathbf{x})\|_2$, we have

$$\begin{aligned}r_i(\mathbf{w}^{(i)},t^{(i)}) &= -\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(T^{(i)}(\mathbf{x})+\rho)(\mathbf{w}^{(i)}\cdot\mathbf{x}+t^{(i)})y\right] \\ &\geq 2\rho\|\ell^{(i)}(\mathbf{x})\|_2 - \rho\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}^{(i)}\cdot\mathbf{x}+t^{(i)})y\right] \geq \rho\|\ell^{(i)}(\mathbf{x})\|_2 \geq 0 \,,\end{aligned}$$

where we used the Cauchy-Schwarz inequality and the fact that $\mathbf{x}$ is standard normal. $\square$

Since we do not have access to $r_i$ precisely, we need a function $\hat{r}_i$, which is close to $r_i$ with high probability. The following simple lemma gives us an efficient way to compute an approximation $\hat{r}_i$ of $r_i$.

**Lemma 2.3.22** (Estimating the function $r_i$)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal and let $T^{(i)}(\mathbf{x})$ be a non-negative function returned by a $(2\rho)$-certificate oracle. Moreover, assume that $T^{(i)}(\mathbf{x})$ has bounded $\ell_4$ norm, i.e., $\|T^{(i)}(\mathbf{x})\|_4 \leq 1$. Then after drawing $O(d\log(1/\epsilon)/\epsilon^2 \log(d/\delta))$ samples from $\mathcal{D}$, with probability at least $1-\delta$, we can compute an estimator $\hat{r}_i$ that satisfies the following conditions:*

- $\left\|\nabla \hat{r}_i(\mathbf{w},t) - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(T^{(i)}(\mathbf{x})+\rho)y(\mathbf{x},1)]\right\|_2 \leq \epsilon/\sqrt{\log(1/\epsilon)}$
- $\|\nabla \hat{r}_i(\mathbf{w},t)\|_2 \leq 2\sqrt{d}$.

*Proof.* Let $\mathcal{D}_N$ be the empirical distribution of $\mathcal{D}$ with $N$ samples. It suffices to find $N$, such that with probability $1-\delta$ the

$$\left\| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}_N}[(T(\mathbf{x})+\rho)y\mathbf{x}] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(T(\mathbf{x})+\rho)y\mathbf{x}] \right\|_2 \leq \epsilon \,.$$

Let $\widehat{\mathbf{R}} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}_N}[(T(\mathbf{x})+\rho)y\mathbf{x}]$ and $\mathbf{R} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(T(\mathbf{x})+\rho)y\mathbf{x}]$. From Markov's inequality, we have that

$$\mathbf{Pr}\left[\left\|\widehat{\mathbf{R}} - \mathbf{R}\right\|_2 \geq \epsilon\right] \leq \frac{1}{\epsilon^2}\,\mathbf{E}[\|\widehat{\mathbf{R}} - \mathbf{R}\|_2^2] \leq \frac{1}{N\epsilon^2}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|(T(\mathbf{x})+\rho)y\mathbf{x}\|_2^2] \,.$$

From Cauchy-Schwarz, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\|(T(\mathbf{x})+\rho)\mathbf{x}\|_2^2] \leq (\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(T(\mathbf{x})+\rho)^4])^{1/2}(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\|\mathbf{x}\|_2^4])^{1/2} .$$

From the fact that $\|T\|_4 \leq 1$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(T(\mathbf{x})+\rho)^4] \leq 4(1+\rho^4)$. Moreover, $(\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\|\mathbf{x}\|_2^4])^{1/2} \lesssim d$, hence

$$\mathbf{Pr}\left[\left\|\widehat{\mathbf{R}}-\mathbf{R}\right\|_2 \geq \epsilon\right] \lesssim \frac{4d}{N\epsilon^2} .$$

For $N' = O(\frac{d}{\epsilon^2})$ samples, we get that the above holds with constant probability. Like in Fact 2.3.18, we can amplify the probability to $1-\delta$, using $N = N'\log(d/\delta)$ samples.

Therefore, let $\widetilde{\mathbf{R}}$ be the estimator of $\mathbf{R}$, we have

$$\left\|\widetilde{\mathbf{R}}\right\|_2 \leq \|\mathbf{R}\|_2 + \epsilon \leq 4\sqrt{(1+\rho^2)}\sqrt{d} + \epsilon \leq 8\sqrt{d} . \tag{2.36}$$

Finally, it remains to show that we can compute an estimator $\widetilde{\mathbf{M}} = (\widetilde{\mathbf{R}}, \tilde{m})$ such that for any unit vector $\mathbf{v} \in \mathbb{R}^d$ and any $|t| \leq 4\sqrt{\log(1/\epsilon)}$, we have with probability $1-\delta$ that

$$|\widetilde{\mathbf{R}}\cdot\mathbf{v} + \tilde{m}t - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(T(\mathbf{x})+\rho)y(\mathbf{x}\cdot\mathbf{v}+t)]| \leq \epsilon .$$

From the above, for any unit vector $\mathbf{v} \in \mathbb{R}^d$, we have with probability $1-\delta$ that $|\widetilde{\mathbf{R}}\cdot\mathbf{v} - \mathbf{R}\cdot\mathbf{v}| \leq \|\widetilde{\mathbf{R}}-\mathbf{R}\|_2 \leq \epsilon$. Therefore, it remains to estimate $\widetilde{m}$. Let $\mathcal{D}_{N'}$ be the empirical distribution of $\mathcal{D}$ with $N'$ samples. From Markov's inequality we have for any $\epsilon' > 0$ that

$$\mathbf{Pr}[|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}_{N'}}[(T(\mathbf{x})+\rho)yt] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(T(\mathbf{x})+\rho)yt]| \geq \epsilon'] \leq \frac{\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(T(\mathbf{x})+\rho)^2]t^2}{N'\epsilon^2} \leq \frac{4t^2}{N'\epsilon^2} .$$

Therefore, using $N' = O(\log(1/\epsilon)/\epsilon^2)$ we get an estimator with constant probability and hence by using the boosting technique as before, we get that with overall $N' \log(1/\delta)$ samples, we can boost the probability to $1 - \delta$. □

We now proceed with the proof of Proposition 2.3.19.

*Proof of Proposition 2.3.19.* Note for the proof for simplicity, we assume that we have access to $(2\rho)$-certifying oracle, but the same argument works for $\rho' = \rho/2$ and access to $\rho'$-certifying oracle. Let $\ell^{(i)}(\mathbf{x}) = \mathbf{w}^{(i)} \cdot \mathbf{x} + t^{(i)}$. We define $\mathcal{V} = \{(\mathbf{w}, t) \in \mathbb{R}^{d+1} : \|\mathbf{w}\|_2 \leq 1, |t| \leq 4\sqrt{\log(1/\epsilon)}\}$. Let $T$ be the number of optimization steps that our algorithm runs. Assume, in order to reach a contradiction, that for all steps $i \in [T]$ it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell^{(i)}(\mathbf{x})) \neq y] \geq \mathrm{OPT} + \epsilon$. Let $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + t^*)$. For each step $i$, let define $T^{(i)}(\mathbf{x})$ to be the non-negative function outputted by the $(2\rho)$-certifying oracle $\mathcal{O}$. Thus, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[T^{(i)}(\mathbf{x}) y(\mathbf{w}^{(i)} \cdot \mathbf{x} + t^{(i)})] \leq -\rho \|\ell^{(i)}(\mathbf{x})\|_2 \, .$$

Let $\nabla \hat{r}_i(\mathbf{w}, t)$ be an estimator of $\nabla r_i(\mathbf{w}, t) = -\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\left(T^{(i)}(\mathbf{x}) + \rho\right) y(\mathbf{x}, 1)]$ such that with probability at least $1 - \frac{\delta}{T}$ it holds $\|\nabla \hat{r}_i(\mathbf{w}, t) - \nabla r_i(\mathbf{w}, t)\|_2 \leq \frac{1}{2}\rho\epsilon^2$ for all $(\mathbf{w}, t) \in \mathcal{V}$. Moreover, from Lemma 2.3.22, for $N = \tilde{O}(d/\log(T/\delta)/(\epsilon^4\rho^2))$ samples, we can achieve that.

Using the separation between the optimal hypothesis and the current one, i.e., Lemma 2.3.21, for every step $i \in [T]$, we have that $r_i(\mathbf{w}^{(i)}, t^{(i)}) - r_i(\mathbf{w}^*, t^*) \geq \frac{1}{2}\rho\epsilon^2$. Therefore, it holds

$$\hat{r}_i(\mathbf{w}^{(i)}, t^{(i)}) - \hat{r}_i(\mathbf{w}^*, t^*) \geq \frac{1}{4}\rho\epsilon^2 \, , \tag{2.37}$$

with probability at least $1 - \frac{\delta}{T}$. In order to apply the online gradient descent algorithm,

i.e., Lemma 2.3.20, we need to define the parameters of the algorithm. First, the convex set we optimize over is $\mathcal{V}$, hence, the diameter of $\mathcal{V}$ is $K = O(\sqrt{\log(1/\epsilon)})$. Furthermore, from Lemma 2.3.22, we get that $\|\nabla \hat{r}_i(\mathbf{w}, t)\|_2 = O(\sqrt{d})$, and therefore, from Lemma 2.3.20, it holds

$$\frac{1}{T}\sum_{i=1}^{T}\left(\hat{r}_i(\mathbf{w}^{(i)}, t^{(i)}) - \hat{r}_i\left(\mathbf{w}^*, t^*\right)\right) \lesssim \frac{\sqrt{d} + \sqrt{\log(1/\epsilon)}}{\sqrt{T}} .$$

By the union bound, it follows that with probability at least $1 - \delta$, we have that

$$\frac{1}{4}\rho\epsilon^2 \leq \frac{1}{T}\sum_{i=1}^{T}\left(\hat{r}_i(\mathbf{w}^{(i)}, t^{(i)}) - \hat{r}_i(\mathbf{w}^*, t^*)\right) \lesssim \frac{\sqrt{d} + \sqrt{\log(1/\epsilon)}}{\sqrt{T}} ,$$

which leads to a contradiction for $T = \Theta(\frac{d\log(1/\epsilon)}{\rho^2\epsilon^4})$.

Thus, either there exists $i \in [T]$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell^{(i)}(\mathbf{x})) \neq y] \leq \mathrm{OPT} + \epsilon$, which the algorithm returns, or the $(2\rho)$-certifying oracle $\mathcal{O}$ did not provide a correct certificate, which happens with probability at most $\delta$. Moreover, the algorithm calls the certificate oracle $T$ times and the number of samples needed are $N$. The runtime is the number of steps and multiplies by the number of samples $T$ $N$. This completes the proof. □

Given Proposition 2.3.19, the proof of Theorem 2.3.1 follows by showing that the function returned by Proposition 2.3.17 is a $\rho$-certifying oracle, for an appropriate choice of $\rho$.

*Proof of Theorem 2.3.1.* In order to prove Theorem 2.3.1, we need to construct a $\rho$-certifying oracle. Using $N = O(d^{\log(1/(\gamma\beta))})\log(1/\delta)/\epsilon^2$ samples, Proposition 2.3.17 provides with probability $1 - \delta$ a $\rho = \epsilon^2\mathrm{poly}(\beta\gamma)$ certifying oracle in runtime $\mathrm{poly}(N, d)$.

Therefore, from Proposition 2.3.19 the proof follows. The overall samples complexity is $O(N \log(T))$ and the runtime is $\mathrm{poly}(N, d)$. □

## 2.4 Learning Halfspaces with General Massart Noise

In this section, we present our algorithm for learning halfspaces with general Massart noise. The main result of this section is the following theorem.

**Theorem 2.4.1** (Learning Halfspaces with General Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta = 1/2$ with respect to a target (possibly biased) halfspace $f \in \mathcal{C}$. Let $\epsilon, \delta \in (0, 1]$. There exists an algorithm that draws $N = d^{O(\log(1/\epsilon))} \log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d) 2^{\mathrm{poly}(1/\epsilon)}$, and computes a halfspace $h \in \mathcal{C}$ such that with probability at least $1 - \delta$ it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$.*

We remark that the algorithm of Theorem 2.4.1 works for any halfspace, regardless of whether it is biased or not. In the presentation that follows, we will focus on learning homogeneous halfspaces for the sake of simplicity. It is not hard to generalize the algorithm to work for arbitrary halfspaces. (We present the general algorithm for arbitrary halfspaces in Section A.4.) The proof of Theorem 2.4.1 consists of three main parts; see Section (2.2.2) for a roadmap of the proof. First, in Subsection 2.4.2 we show that by restricting on a thin slice and then projecting the distribution on the subspace $\mathbf{w}^{\perp}$, we obtain an instance where the optimal halfspace has again 1/2-Massart noise everywhere apart from a small region close to the halfspace, see Figure 2.2a. In Subsection 2.4.1, we present our main technical contribution, i.e., that there exists a low-degree, mean-zero, polynomial that achieves non-trivial correlation when $\eta(\mathbf{x}) \leq 1/2$

"almost everywhere". Finally, in Subsection 2.4.3 we show that the existence of such polynomials implies that by finding the left singular vectors of (a flattened version) of the Chow tensors of the distribution, we can construct a $\mathrm{poly}(1/\epsilon)$-dimensional subspace inside which the direction $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ has $\mathrm{poly}(\epsilon)$ projection. Recall that $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ is the normalized projection of $\mathbf{w}^*$ onto the orthogonal complement of $\mathbf{w}^{\perp}$:

$$(\mathbf{w}^*)^{\perp_{\mathbf{w}}} = \frac{\mathrm{proj}_{\mathbf{w}^{\perp}}(\mathbf{w}^*)}{\|\mathrm{proj}_{\mathbf{w}^{\perp}}(\mathbf{w}^*)\|_2}\,.$$

In Subsection 2.4.3, we combine everything together to prove Theorem 2.4.1.

We now state the main technical result of this section. (This is the formal version of Proposition 2.2.5.)

**Proposition 2.4.2.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta = 1/2$, with respect to some target halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Let $\mathbf{w} \in \mathbb{R}^d$ be a unit vector such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y] \geq \mathrm{OPT} + \epsilon$ and $\theta(\mathbf{w}, \mathbf{w}^*) \leq \pi - \epsilon$ for some $\epsilon \in (0, 1/2]$. There exists an algorithm that draws $N = d^{O(\log(1/\epsilon))}\log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and with probability at least $1-\delta$ returns a basis of a subspace $V \subseteq \mathbf{w}^{\perp}$ such that $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp_{\mathbf{w}}})\|_2 \geq \mathrm{poly}(\epsilon)$.*

Observe that given a subspace $V$ with the above guarantees, we can sample a uniformly random direction on the unit sphere of $V$ and obtain a unit update vector $v$ such that $\mathbf{v} \in \mathbf{w}^{\perp}$ and $\mathbf{v} \cdot (\mathbf{w}^*)^{\perp_{\mathbf{w}}} \geq \mathrm{poly}(\epsilon)$, which given Lemma 2.4.14 improves the current guess $\mathbf{w}$; see Subsection 2.4.3 for the details.

### 2.4.1 The Sign-Matching Polynomial

Here we give an explicit construction of a mean-zero low-degree polynomial that achieves non-trivial correlation with the labels $y$. The assumptions on the distribution $\mathcal{D}$ over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ are that it has Gaussian $\mathbf{x}$-marginal and "almost-Massart" noise, i.e., that is $\eta(\mathbf{x})$ is greater than $1/2$ only on a very small region close to the optimal halfspace. We prove the following:

**Proposition 2.4.3** (Correlation via the Sign-Matching Polynomial)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is the standard normal distribution. Let $f(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} - b)$ be such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(b)f(\mathbf{x}) > 0\}] \geq \zeta$ for some $\zeta \in (0, 1/2]$, and $\eta(\mathbf{x}) > 1/2$ only when $0 \leq \mathrm{sign}(b)(\mathbf{v}^* \cdot \mathbf{x} - b) \leq \xi$, where $\xi$ is a sufficiently small constant multiple of $\zeta^3$. There exists a univariate polynomial $p(z)$ of degree $\Theta(b^2 + 1)$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p(\mathbf{v}^* \cdot \mathbf{x})] = 0$, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[p^2(\mathbf{v}^* \cdot \mathbf{x})] = 1$, and $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[yp(\mathbf{v}^* \cdot \mathbf{x})] = \mathrm{poly}(\zeta)$.*

*Proof.* The main ingredient of the proof is the following lemma that shows the existence of a mean zero polynomial that matches the sign of the linear function $z - b$; see Figure 2.2b.

**Lemma 2.4.4** (Sign-Matching Polynomial)**.** *Let $b \in \mathbb{R}$. There exists a zero mean and unit variance polynomial $p : \mathbb{R} \mapsto \mathbb{R}$ of degree $k = \Theta(b^2 + 1)$ such that*

- *The sign of $p$ matches the sign of the threshold function $\mathrm{sign}(z-b)$, i.e., $\mathrm{sign}(p(z)) = \mathrm{sign}(z - b)$ all $z \in \mathbb{R}$,*
- *for any $\rho \in (0, 1)$, it holds $|p(b+\mathrm{sign}(b)\rho)| \geq c^k \mathrm{poly}(\rho)$, where $c > 0$ is a universal constant,*

- *$p(z)$ is increasing for all $|z| \geq |b|$.*

*Proof.* We first pick some odd integer $k$ large enough such that $2|b| \leq \sqrt{k} \leq 4\max(|b|, 1)$. Let $q(z) = z^{3k}, r(z) = z^{2k} - (2k-1)!!$. We consider the polynomial

$$p(z) = q(z) - \frac{q(b)}{r(b)} r(z)\,.$$

It holds $p(b) = 0$ and $\mathbf{E}_{z\sim\mathcal{N}}[p(z)] = 0$, since $\mathbf{E}_{z\sim\mathcal{N}}[z^{3k}] = 0$, for odd $k$ and $\mathbf{E}_{z\sim\mathcal{N}}[z^{2k}] = (2k-1)!!$.

For the rest of the proof, we need the following simple estimate for double factorials. Sharper bounds can be obtained via Stirling's approximation; for our purposes the following rough bounds suffice.

**Fact 2.4.5.** *Let $m$ be a positive integer. Then $(m/2)^m \leq (2m-1)!! \leq (2m)^m$.*

We next show that $p(z)$ is increasing for all $|z| \geq |b|$. We compute the derivative of $p(z)$, i.e.,

$$p'(z) = 3kz^{3k-1} - 2k\frac{q(b)}{r(b)} z^{2k-1} = kz^{2k-1}\left(3z^k - 2\frac{q(b)}{r(b)}\right)\,, \tag{2.38}$$

we observe that if $|z^k| \geq \frac{2}{3}|q(b)/r(b)|$, then $p(z)$ is increasing. We show that this is the case for all $|z| \geq |b|$. Since $|z|^k$ is increasing in $|z|$ it suffices to verify the inequality for $|z| = |b|$. In that case we obtain the inequality $(2k-1)!!/b^{2k} \geq 5/3$, which can be verified by using Fact 2.4.5 and our assumption that $|b| \leq \sqrt{k}/2$.

We now prove that $\mathrm{sign}(p(z)) = \mathrm{sign}(z-b)$. Note that $\mathrm{sign}(r(z)) < 0$ for any $z \in [-|b|, |b|]$ because $|b| \leq \sqrt{k}/2$ and thus, from Fact 2.4.5, it holds $b^{2k} - (2k-1)!! \leq (k/4)^k - (k/2)^k \leq 0$. Without loss of generality, assume that $b > 0$. Assume that there exists $s \in \mathbb{R}$ with $s \neq b$ such that $p(s) = 0$. First, assume that $s \geq 0$, by eq. (2.38), since $r(b) < 0$ we get that $p(z)$ is increasing for $z \geq 0$, therefore the polynomial $p(z)$

has only one root at $s = b$, therefore we have a contradiction. For the case that $s < 0$, first observe that for any $s = 0$, it holds that $q(s)r(b) = q(b)r(s)$ and in particular $\text{sign}(q(s)r(b)) = \text{sign}(q(b)r(s))$. Thus, if $-b > s$, then it holds $\text{sign}(q(b)r(s)) = -\text{sign}(b)$ and $\text{sign}(q(s)r(b)) = -\text{sign}(s)$, therefore for any such solution it holds $\text{sign}(s) = \text{sign}(b)$ but $s < 0$ and $b > 0$, so we have a contradiction. Finally, for the case that $s \leq -b$, recall that for $|z| \geq |b|$, $p(z)$ is increasing, so for $z < -b$, we have that $p(z)$ is increasing. But $p(-b) < 0$ because there is no other root in the interval $[-b, b)$, thus $p(s) < 0$ which leads to a contradiction. Therefore, the only root of the polynomial is at $s = b$.

We next prove that $\mathbf{E}_{z\sim\mathcal{N}}[p^2(z)] = (O(k))^{3k}$. By applying twice the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we get

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[p^2(z)] \lesssim \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[z^{6k}] + \frac{q^2(b)}{r^2(b)}\left(\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[z^{4k}] + ((2k-1)!!)^2\right) \lesssim (6k-1)!! + \frac{q^2(b)}{r^2(b)}(4k-1)!! \,.$$

Using that $|b| \leq \sqrt{k}/2$ and Fact 2.4.5, we have that $|b^{2k} - (2k-1)!!| \geq (k/2)^k/2$. Hence, we get that $(q(b)/r(b))^2 \lesssim (2|b|^3/k)^{2k} \lesssim k^k$. Therefore, using Fact 2.4.5, we get that $\mathbf{E}_{z\sim\mathcal{N}}[p^2(z)] = O((6k)^{3k})$.

Without loss of generality, assume that $b > 0$. Observe that, for $\rho \in (0,1)$, we have

$$\begin{aligned} p(b+\rho) &= (b+\rho)^{3k} - \frac{b^{3k}((b+\rho)^{2k} - (2k-1)!!)}{b^{2k} - (2k-1)!!} \\ &= (b+\rho)^{3k} - b^{3k} - \frac{b^{3k}\sum_{m=0}^{2k-1}\binom{2k}{m}b^m\rho^{2k-m}}{b^{2k} - (2k-1)!!} \,. \end{aligned}$$

Recall that using the fact that $|b| \leq \sqrt{k}/2$, it holds that $b^{2k} - (2k-1)!! < 0$, thus

$$p(b+\rho) \geq (b+\rho)^{3k} - b^{3k} = \sum_{m=0}^{3k-1}\binom{3k}{m}b^m\rho^{3k-m} \,.$$

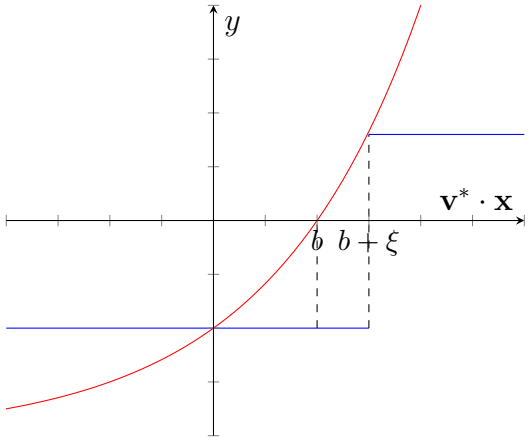



Figure 2.6: The sign-matching polynomial $p$ corresponds to the red curve and the threshold function $f(\mathbf{x})$ to the blue. Observe that, in the region $I^{\xi} = \{b \leq \mathbf{v}^* \cdot \mathbf{x} \leq b + \xi\}$, the sign of the polynomial $p$ does not agree with $f(\mathbf{x})$. This is the region where the Massart noise condition is (potentially) violated, as a result of the orthogonal projection step, Figure 2.2a. Even though the sign of $p$ does not agree with $f(\mathbf{x})$ everywhere, we can take $\xi$ to be small, i.e., $\xi$ is a sufficiently small constant multiple of $\zeta^3$, making the negative contribution of the region $I_{\xi}$ small. We also use crucially the fact that $p$ is monotone for $\mathbf{v}^* \cdot \mathbf{x} \geq b$ and not very flat in that region, see Item 2 of Lemma 2.4.4.

From the equation above, we get that $p(b+\rho) \gtrsim b^{3k-1}\rho + \rho^{3k}$. Let $\tilde{p}(z)$ be the polynomial $p(z)$ normalized, i.e., $\tilde{p}(z) = p(z)/(\mathbf{E}_{z\sim\mathcal{N}}[p^2(z)])^{1/2}$. Notice that normalizing $p$ does not affect its sign or monotonicity. It remains to prove that $\tilde{p}(b+\rho) = c^k \mathrm{poly}(\rho)$, where $c > 0$ is a small enough constant. If $b \geq 1$, recall that in that case it holds that $\sqrt{k} \leq 4b$, therefore it holds $\tilde{p}(b+\rho) \gtrsim b^{3k-1}\rho/(6k)^{3k/2} \gtrsim \rho/(24)^{2k}$. For the case where $0 < b < 1$, we have that $k = O(1)$ and similarly, we have $\tilde{p}(b+\rho) \gtrsim \rho^{2k}/(6k)^{3k/2} = \mathrm{poly}(\rho)$, hence $\tilde{p}(b+\rho) = c^k \mathrm{poly}(\rho)$, for some universal constant $c > 0$. □

It remains to show that the polynomial $p$ of the Lemma 2.4.4, achieves non-trivial correlation with $y$, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[yp(\mathbf{v}^* \cdot \mathbf{x})] = \mathrm{poly}(\zeta)$. To do this we are going to use the monotonicity properties of $p$, the fact that it matches the sign of $f(\mathbf{x})$, and the fact that it is not very flat close to its unique root at $b$, see Item 2 of Lemma 2.4.4. We

assume that $b \geq 0$ as the other case follows similarly. Recall that, from the assumptions of Proposition 2.4.3, there is an explicitly defined region where the noise is above $1/2$, i.e., $\eta(\mathbf{x}) > 1/2$. We denote this region $I^{\xi}$, see also Figure 2.6, that is

$$I^{\xi} = \{\mathbf{x} \in \mathbb{R}^d : \eta(\mathbf{x}) > 1/2\} \subseteq \{\mathbf{x} \in \mathbb{R}^d : b \leq \mathbf{v}^* \cdot \mathbf{x} \leq b + \xi\} \,.$$

Since $p(\mathbf{v}^* \cdot \mathbf{x})$ ($p$ from Lemma 2.4.4) matches the sign of $\mathbf{v}^* \cdot \mathbf{x} - b$ we have

$$\begin{aligned} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{v}^* \cdot \mathbf{x})y] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[p(\mathbf{v}^* \cdot \mathbf{x})\mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} - b)\beta(\mathbf{x})(\mathbb{1}\{\mathbf{x} \notin I^{\xi}\} - \mathbb{1}\{\mathbf{x} \in I^{\xi}\})] \\ &\geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|p(\mathbf{v}^* \cdot \mathbf{x})|\beta(\mathbf{x})\mathbb{1}\{\mathbf{x} \notin I^{\xi}\}] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|p(\mathbf{v}^* \cdot \mathbf{x})|\mathbb{1}\{\mathbf{x} \in I^{\xi}\}] \,. \end{aligned} \tag{2.39}$$

We first bound from below the correlation of the sign-matching polynomial outside the region $I^{\xi}$, i.e., the region where the Massart noise condition is true. From Lemma 2.4.4, we have that $p(z)$ is increasing for $z > b$, and therefore it holds

$$\begin{aligned} &\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|p(\mathbf{v}^* \cdot \mathbf{x})|\beta(\mathbf{x})]\mathbb{1}\{\mathbf{x} \notin I^{\xi}\}] \\ &\quad\geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|p(\mathbf{v}^* \cdot \mathbf{x})|\beta(\mathbf{x})\mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} > b + \zeta^2\}] \\ &\quad\geq |p(b + \zeta^2)| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\beta(\mathbf{x})\mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} > b + \zeta^2\}] \\ &\quad= |p(b + \zeta^2)| \left( \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\beta(\mathbf{x})\mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} > b\}] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{b \leq \mathbf{v}^* \cdot \mathbf{x} \leq b + \zeta^2\}] \right) \,, \end{aligned} \tag{2.40}$$

where for the first inequality we used the fact that $\mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} > b + \xi\} \geq \mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} > b + \zeta^2\}$, because $\xi = \Theta(\zeta^3) \leq \zeta^2$ from the assumptions of Proposition 2.4.3, and for the second, we used the monotonicity of $p$. We next bound from above the contribution of the

sign-matching polynomial on the region where the Massart noise condition is violated, i.e., $\mathbf{x} \in I^{\xi}$. Moreover, using again the fact that $p(z)$ is increasing for $z > b$, we get

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|p(\mathbf{v}^* \cdot \mathbf{x})|]\mathbb{1}\{\mathbf{x} \in I^{\xi}\}] &\leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|p(\mathbf{v}^* \cdot \mathbf{x})|\mathbb{1}\{b \leq \mathbf{v}^* \cdot \mathbf{x} \leq b + \zeta^2\}] \\ &\leq |p(b+\zeta^2)| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{b \leq \mathbf{v}^* \cdot \mathbf{x} \leq b + \zeta^2\}] \,. \end{aligned} \tag{2.41}$$

Substituting the bounds of eq. (2.40) and eq. (2.41) to eq. (2.39), we get that

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{v}^* \cdot \mathbf{x})y] &\geq |p(b+\zeta^2)|(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\beta(\mathbf{x})\mathbb{1}\{\mathbf{v}^* \cdot \mathbf{x} \geq b\}] - 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{b \leq \mathbf{v}^* \cdot \mathbf{x} \leq b + \zeta^2\}]) \\ &\gtrsim |p(b+\zeta^2)|(\zeta - \zeta^2) \gtrsim |p(b+\zeta^2)|\zeta \,,\end{aligned}$$

where we used that from the assumptions of Prop. 2.4.3 we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(b)f(\mathbf{x}) > 0\}] \gtrsim \zeta$, and the anti-concentration property of the Gaussian distribution from Fact 2.2.11, i.e., $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{b < \mathbf{v}^* \cdot \mathbf{x} < b + \zeta^2\}] \leq \zeta^2/(\sqrt{2\pi})$.

It remains to prove that $|p(b+\zeta^2)| = \mathrm{poly}(\zeta)$. From Lemma 2.4.4, we have that $|p(b+\zeta^2) = c^k \mathrm{poly}(\zeta)$, for some universal constant $c > 0$. Note that from the assumption of Proposition 2.4.3 we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(b)\mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} - b) > 0\}] \gtrsim \zeta$, hence, from Fact 2.2.11 we get $|b| \lesssim \sqrt{\log(1/\zeta)}$, therefore $c^k = \mathrm{poly}(\zeta)$, since $k = \Theta(b^2+1)$. Hence, we have $|p(b+\zeta^2)| = \mathrm{poly}(\zeta)$. This completes the proof of Proposition 2.4.3.

□

### 2.4.2 Projecting onto $\mathbf{w}^{\perp}$

In this subsection, we show that we can project $\mathcal{D}$ onto the subspace $(\mathbf{w}^*)^{\perp_{\mathbf{w}}}$ without violating the 1/2-Massart noise condition by a lot. At the same time we make sure

that the "optimal halfspace" of the projected instance is not very biased. We show the following lemma.

**Lemma 2.4.6** (Orthogonal Projection Inside a Band)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta = 1/2$ with respect to $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Let $\mathbf{w}$ be a unit vector such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y] \geq \mathrm{OPT}+\epsilon$, and $\theta(\mathbf{w},\mathbf{w}^*) \leq \pi-\epsilon$, for some $\epsilon \in (0,1]$. Fix some $\rho > 0$. For $t_1, t_2 \in \mathbb{R}$, consider the band $B = \{t_1 \leq \mathbf{w}\cdot\mathbf{x} \leq t_2\}$. Denote $\mathcal{D}^\perp = \mathcal{D}_B^{\perp_\mathbf{w}}$, i.e., $\mathcal{D}$ is the orthogonal projection onto $\mathbf{w}^\perp$ of the conditional distribution on $B$, and consider the halfspace $f^\perp : \mathbf{w}^\perp \mapsto \{\pm 1\}$, with $f^\perp(\mathbf{x}) = \mathrm{sign}(\mathbf{x}\cdot(\mathbf{w}^*)^{\perp_\mathbf{w}} - b)$, for some threshold $b \in \mathbb{R}$. Moreover, define the noise function $\eta^\perp(\mathbf{x}) = \mathbf{Pr}_{(\mathbf{z},y)\sim\mathcal{D}^\perp}[y \neq f^\perp(\mathbf{z})|\mathbf{z} = \mathbf{x}]$. There exist $t_1, t_2 \in \mathbb{R}$ multiples of $\rho$, with $|t_1 - t_2| = \rho$, such that:*

1. $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}^\perp}[(1-2\eta^\perp(\mathbf{x}))\mathbb{1}\{f^\perp(\mathbf{x})\mathrm{sign}(b) > 0\}] \gtrsim \epsilon/\sqrt{\log(1/\epsilon)} - \rho/\epsilon$,
2. *if* $\eta^\perp(\mathbf{x}) > 1/2$ *then* $0 < \mathrm{sign}(b)(\mathbf{x}\cdot(\mathbf{w}^*)^{\perp_\mathbf{w}} - b) \leq \rho/\epsilon$.

**Remark 2.4.7.** Observe that for Lemma 2.4.6 to give non-trivial guarantees we have to pick the size of the band, $\rho$, to be smaller than a sufficiently small constant multiple of $\epsilon/\sqrt{\log(1/\epsilon)}$. In fact, in what follows, we will be using bands of size $\rho = \mathrm{poly}(\epsilon)$. Moreover, the condition that $\theta(\mathbf{w},\mathbf{w}^*) \geq \pi - \epsilon$ is mostly technical. Having a vector $\mathbf{w}$ with $\theta(\mathbf{w},\mathbf{w}^*) \geq \pi - \epsilon$ we can always output $-\mathbf{w}$ and achieve classification error $O(\epsilon)$. See the proof of Theorem 2.4.1.

*Proof.* First, let us denote $\beta(\mathbf{x}) = 1-2\eta(\mathbf{x})$. Note that the assumption $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y] \geq \mathrm{OPT}+\epsilon$ is equivalent to $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\}\beta(\mathbf{x})] \geq \epsilon$. We consider the region $B = \{t_1 \leq \mathbf{w}\cdot\mathbf{x} \leq t_2\}$, for any $t_1, t_2 \in \mathbb{R}$ with $|t_1 - t_2| = \rho$. We denote by $\mathbf{x}^\perp$ the projection of $\mathbf{x}$ onto the subspace $\mathbf{w}^\perp$. Define the distribution $\mathcal{D}^\perp = \mathcal{D}_B^{\mathrm{proj}_{\mathbf{w}^\perp}}$, that

is the distribution $\mathcal{D}$ conditioned on the set $B$ and projected onto $\mathbf{w}^\perp$, the hypothesis $f^\perp(\mathbf{x}^\perp) = \mathrm{sign}(\mathbf{x}^\perp \cdot (\mathbf{w}^*)^{\perp_\mathbf{w}} + b)$ where $b \in \mathbb{R}$ is chosen appropriately below, and the noise function $\eta^\perp(\mathbf{x}^\perp) = \mathbf{Pr}_{(\mathbf{z},y)\sim\mathcal{D}^\perp}[y \neq f^\perp(\mathbf{z}) | \mathbf{z} = \mathbf{x}^\perp]$, see Figure 2.2a.

Note that the distribution $\mathcal{D}^\perp$ does not satisfy the 1/2-Massart noise condition anymore. We first illustrate how the noise function changes. The orthogonal projection on $\mathbf{w}^\perp$ creates a region where the 1/2-Massart noise condition is violated, i.e., a region where $\eta^\perp(\mathbf{x}^\perp)$ may be more than 1/2, but we can control the probability that we get points inside this region, i.e., the green area of Figure 2.2a. More formally, we show that

$$\mathop{\mathbf{E}}_{(\mathbf{x}_\mathbf{w},\mathbf{x}^\perp)\sim(\mathcal{D}_B)_\mathbf{x}} \left[\beta(\mathbf{x}) \mathbb{1}\{f(\mathbf{x}_\mathbf{w}, \mathbf{x}^\perp) \neq f^\perp(\mathbf{x}^\perp)\}\right] \leq \mathop{\mathbf{Pr}}_{\mathbf{x}^\perp\sim\mathcal{D}_\mathbf{x}^\perp}[\eta^\perp(\mathbf{x}^\perp) > 1/2] \lesssim \rho/\epsilon \,. \quad (2.42)$$

To show Inequality (2.42) notice that $\theta(\mathbf{w}, \mathbf{w}^*) \gtrsim \epsilon$, otherwise we would have $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}\cdot\mathbf{x})] \leq \epsilon$ from Fact 2.2.12. We can assume that $\mathbf{w}^* = \lambda_1\mathbf{w} + \lambda_2(\mathbf{w}^*)^{\perp_\mathbf{w}}$, where $\lambda_1 = \cos\theta$ and $\lambda_2 = \sin\theta$. Note that $\lambda_2 \gtrsim \epsilon$, since $\theta \in [\Omega(\epsilon), \pi - \epsilon]$. Next we set $\mathbf{x} = (\mathbf{x}_\mathbf{w}, \mathbf{x}^\perp)$, where $\mathbf{x}_\mathbf{w} = \mathbf{w}\cdot\mathbf{x}$. We show that the hypothesis $f^\perp(\mathbf{x}) = \mathrm{sign}((\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^\perp + t_1\lambda_1/\lambda_2)$ is almost as good as the $f(\mathbf{x})$ for the distribution $\mathcal{D}^\perp$. Note that in the statement of Lemma 2.4.6, we write $f^\perp(\mathbf{x}) = \mathrm{sign}((\mathbf{w}^*)^{\perp_\mathbf{w}} \cdot \mathbf{x}^\perp - b)$, where $b = -t_1\lambda_1/\lambda_2$.

Conditioned on $\mathbf{x} \in B$, i.e., $\mathbf{x}_\mathbf{w} = \mathbf{x}\cdot\mathbf{w} \in [t_1, t_1 + \rho]$, it holds that

$$\mathbf{w}^*\cdot\mathbf{x} = \lambda_1\mathbf{x}_\mathbf{w} + \lambda_2(\mathbf{w}^*)^{\perp_\mathbf{w}}\cdot\mathbf{x}^\perp = \lambda_1 t_1 + \lambda_2(\mathbf{w}^*)^{\perp_\mathbf{w}}\cdot\mathbf{x}^\perp + s\rho\,,$$

for some $s \in [-1, 1]$ (recall that $|\lambda_1| \leq 1$). Notice that when $0 < \mathrm{sign}(-\lambda_1 t_1)(\lambda_1 t_1 + \lambda_2(\mathbf{w}^*)^{\perp_\mathbf{w}}\cdot\mathbf{x}^\perp) < \rho$, $f^\perp(\mathbf{x}^\perp)$ is not necessary equal to the sign of $(\mathbf{w}^*\cdot\mathbf{x})$ (recall that $\lambda_2 > 0$), and therefore we are inside the region that the 1/2-Massart noise condition is violated, otherwise the $f^\perp(\mathbf{x}^\perp)$ always matches the sign of $\mathbf{w}^*\cdot\mathbf{x}$. This proves the

second statement of Lemma 2.4.6.

To prove Inequality (2.42), we need to bound the probability of the event that $f^{\perp}(\mathbf{x}^{\perp})$ is not necessary equal to the sign of $(\mathbf{w}^* \cdot \mathbf{x})$, or equivalently if $\mathrm{sign}(-\lambda_1 t_1) > 0$ then $(\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{x}^{\perp} \in [(-\lambda_1 t_1 - \rho)/\lambda_2, -\lambda_1 t_1/\lambda_2] =: I^{\rho}$, and otherwise $(\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{x}^{\perp} \in [-\lambda_1 t_1/\lambda_2, -(\lambda_1 t_1 + \rho)/\lambda_2] =: I^{\rho}$. We have that

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}^{\perp}\sim\mathcal{D}_{\mathbf{x}}^{\perp}}[\eta^{\perp}(\mathbf{x}^{\perp}) > 1/2] &\leq \mathop{\mathbf{Pr}}_{\mathbf{x}^{\perp}\sim\mathcal{D}_{\mathbf{x}}^{\perp}}\left[(\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{x}^{\perp} \in I^{\rho}\right] \\ &= \frac{\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{x} \in I^{\rho}\right]}{\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x} \in B]} \lesssim \rho/\lambda_2 \lesssim \rho/\epsilon\,,\end{aligned}$$

where we used the anti-concentration property of the Gaussian distribution and the last inequality holds because we have that $\lambda_2 \gtrsim \epsilon$. This proves Inequality (2.42).

It remains to prove that there exist thresholds $t_1, t_2$ and a band $B = \{t_1 \leq \mathbf{w}\cdot\mathbf{x} \leq t_2\}$ with respect the $t_1, t_2$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}^{\perp}}[(1 - 2\eta^{\perp}(\mathbf{x}))\mathbb{1}\{f^{\perp}(\mathbf{x}) \neq \mathrm{sign}(-b)\}] \gtrsim \frac{\epsilon}{\sqrt{\log(1/\epsilon)}}$, where $\mathcal{D}^{\perp}$ is the distribution $\mathcal{D}$ conditioned on the set $B$ and projected onto $\mathbf{w}^{\perp}$. We first show the following claim

**Claim 2.4.8.** *Let $\mathbf{v}$ be a unit vector and let $S$ be a subset of $\mathbb{R}^d$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{\mathbf{x} \in S\}\beta(\mathbf{x})] \geq \epsilon$, for some $\epsilon > 0$. Fix $\rho > 0$. There exist $t_1, t_2 \in \mathbb{R}$ such that $t_1, t_2$ are multiples of $\rho$, $|t_1 - t_2| = \rho$, and*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{\mathbf{x} \in S\}\mathbb{1}\{t_1 \leq \mathbf{v}\cdot\mathbf{x} \leq t_2\}\beta(\mathbf{x})] \gtrsim \frac{\epsilon\rho}{\sqrt{\log(1/\epsilon)}}\,.$$

*Proof.* Note that since $\beta(\mathbf{x})\mathbb{1}\{\mathbf{x} \in S\} \in [0,1]$, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{\mathbf{x} \in S\}\mathbb{1}\{C\sqrt{\log(1/\epsilon)} \leq |\mathbf{v}\cdot\mathbf{x}|\}\beta(\mathbf{x})] \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{C\sqrt{\log(1/\epsilon)} \leq |\mathbf{v}\cdot\mathbf{x}|\}] \leq \epsilon^2$, where $C$ is a large enough universal constant. To get this we used the upper bound on the Gaussian tails,

Fact 2.2.11. Therefore, we can ignore the region far away from the origin and split the interval $|\mathbf{v}\cdot\mathbf{x}| \leq C\sqrt{\log(1/\epsilon)}$, into slices of length $\rho$. There must exist one slice inside which the expectation of $\beta(\mathbf{x})\mathbb{1}\{\mathbf{x}\in S\}$ is roughly $\epsilon\rho/\sqrt{\log(1/\epsilon)}$. Let $t_i = i\rho$ and $t_{-i} = -i\rho$, for $0 \leq i \leq C\sqrt{\log(1/\epsilon)}/\rho$, we define $B_i = \{t_i \leq \mathbf{v}\cdot\mathbf{x} \leq t_{i+1}\}$ and $B_{-i} = \{-t_{i+1} \leq \mathbf{v}\cdot\mathbf{x} \leq -t_i\}$. We have that

$$\sum_{|i|\leq C\sqrt{\log(1/\epsilon)}/\rho} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathbf{x}\in S\}\mathbb{1}\{\mathbf{x}\in B_i\}\beta(\mathbf{x})] \geq \epsilon - \epsilon^2 \geq \epsilon/2\,, \tag{2.43}$$

where we used that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathbf{x}\in S\}\mathbb{1}\{C\sqrt{\log(1/\epsilon)} \leq |\mathbf{v}\cdot\mathbf{x}|\}\beta(\mathbf{x})] \leq \epsilon^2$. Using the fact that all the quantities we add in Equation (2.43) are positive, there exists $B_i'$, for some $i'$, such that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\mathbf{x}\in S\}\mathbb{1}\{\mathbf{x}\in B_{i'}\}\beta(\mathbf{x})] \gtrsim \frac{\epsilon\rho}{\sqrt{\log(1/\epsilon)}}\,,$$

which completes the proof. □

An application of Claim 2.4.8 to the set $\{\mathbf{x}\in\mathbb{R}^d : f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\}$, gives us that there exists a band $B = \{t_1 \leq \mathbf{w}\cdot\mathbf{x} \leq t_2\}$ with $|t_1 - t_2| = \rho$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\}\mathbb{1}\{\mathbf{x}\in B\}\beta(\mathbf{x})] \gtrsim \frac{\epsilon\rho}{\sqrt{\log(1/\epsilon)}}\,.$$

Moreover, note that for the distribution $\mathcal{D}_B$, that is the distribution $\mathcal{D}$ conditioned on $B$, it holds $\mathbf{E}_{\mathbf{x}\sim(\mathcal{D}_B)_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\}] \gtrsim \frac{\epsilon}{\sqrt{\log(1/\epsilon)}}$, where we used the Gaussian anti-concentration property, i.e., that $\mathbf{Pr}_{x\sim\mathcal{N}}[t_1 \leq x \leq t_2] \lesssim |t_2 - t_1|$.

We have that $f^\perp(\mathbf{x})$ agrees almost everywhere with $f(\mathbf{x})$ with respect to the distribution $\mathcal{D}_B$, i.e., we have that $\mathbf{E}_{(\mathbf{x}_\mathbf{w},\mathbf{x}^\perp)\sim(\mathcal{D}_B)_\mathbf{x}}[\mathbb{1}\{f((\mathbf{x}_\mathbf{w},\mathbf{x}^\perp)) \neq f^\perp(\mathbf{x}^\perp)\}] \leq \mathbf{Pr}_{\mathbf{x}^\perp\sim\mathcal{D}_\mathbf{x}^\perp}[\eta^\perp(\mathbf{x}^\perp) >$

$1/2] \lesssim \rho/\epsilon$ (Inequality (2.42)). Thus, using the triangle inequality, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x_w},\mathbf{x}^\perp)\sim(\mathcal{D}_B)_\mathbf{x}}\left[(1-2\eta^\perp(\mathbf{x}^\perp))\mathbb{1}\{\mathrm{sign}(\mathbf{x_w}) \neq f^\perp(\mathbf{x}^\perp)\}\right] \gtrsim \frac{\epsilon}{\sqrt{\log(1/\epsilon)}} - \frac{\rho}{\epsilon}\,.$$

The proof concludes by noting that it holds $\mathrm{sign}(\mathbf{x_w}) = \mathrm{sign}(-b)$, by the definition of $f^\perp(\mathbf{x})$. This completes the proof of Lemma 2.4.6. □

### 2.4.3 Using the Low-Order Chow Tensors

In this subsection, we use our structural result of Subsection 2.4.1 to construct the sampling oracle of Proposition 2.4.2 that provides us with good update vectors $\mathbf{v}$. To do so, we shall find a subspace $V$ of $\mathbb{R}^d$ that contains non-trivial part of $\mathbf{v}^*$. In order to get a good update with non-trivial probability, we will sample a unit vector uniformly at random from $V$. The description of the corresponding algorithm is given in Algorithm 2.

**Proposition 2.4.9.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Let $f(\mathbf{x}) = \mathrm{sign}(\mathbf{v}^* \cdot \mathbf{x} - b)$ be such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[y \neq f(\mathbf{x})|\mathbf{x}] = \eta(\mathbf{x})$, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(b)f(\mathbf{x}) > 0\}] \geq \zeta$ and $\eta(\mathbf{x}) > 1/2$ only when $\mathrm{sign}(b)(\mathbf{v}^* \cdot \mathbf{x} - b) \leq \xi$, where $\xi = \Theta(\zeta^3)$. There exists an algorithm with sample complexity and runtime $d^{O(\log(1/\zeta))}\log(1/\delta)$ that with probability at least $1-\delta$ returns a basis of a subspace $V$ of dimension $\mathrm{poly}(1/\zeta)$ such that $\|\mathrm{proj}_V(\mathbf{v}^*)\|_2 = \mathrm{poly}(\zeta)$.*

*Proof.* The following lemma shows that the existence of a mean-zero polynomial $p$ that achieves non-trivial correlation with $y$ implies that the subspace spanned by the top singular vectors of the Chow tensors of $\mathcal{D}$ will contain non-trivial part of $\mathbf{v}^*$. We remark that we do not rely on tensor SVD to obtain the singular vectors: in what follows we flatten the order-$m$ Chow tensors and treat them as $d \times d^{m-1}$ matrices. We show the

following lemma.

**Algorithm 2** Creating a Random Oracle with Good Correlation, see Proposition 2.4.2.

**Input:**

1. $\epsilon, \delta > 0$.
2. An empirical distribution $\widehat{\mathcal{D}}$ of the distribution $\mathcal{D}$ that satisfies the general Massart noise condition with respect to $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$.
3. A unit vector $\mathbf{v} \in \mathbb{R}^d$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y] \geq \mathrm{OPT} + \epsilon$.

**Output:** A subspace $V$ with $\dim(V) = \mathrm{poly}(1/\epsilon)$, and $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp_{\mathbf{w}}})\|_2 \geq \mathrm{poly}(\epsilon)$.

**Define:** $k = C\log(1/\epsilon)$, $\sigma = \mathrm{poly}(\epsilon)/C$, $\rho = \epsilon^3/C$, $l = C\log(1/\epsilon)/\rho$, for $C > 0$ be a sufficiently large constant.

1. For $i \in [l]$, set $B_i = \{i\rho \leq \mathbf{w}\cdot\mathbf{x} \leq (i+1)\rho\}$ and $B_{-i} = \{-(i+1)\rho \leq \mathbf{w}\cdot\mathbf{x} \leq -i\rho\}$.
2. For $i \in \{-l, \ldots, l\}$, repeat
   (a) Let $\widehat{\mathcal{D}}_i^{\perp}$ be the distribution $\widehat{\mathcal{D}}$ conditioned on $B_i$ and projected onto $\mathbf{w}^{\perp}$.
   (b) For any $m \in [k]$, calculate the $m$-order Chow tensor $\mathbf{T}^m(\widehat{\mathcal{D}}_i^{\perp})$, see Definition 2.2.7.
   (c) Flatten each $\mathbf{T}^m(\widehat{\mathcal{D}}_i^{\perp})$ for $m \in [k]$, to a $d \times d^{m-1}$ matrix $\mathbf{M}^m = (\mathbf{T}^m(\widehat{\mathcal{D}}_i^{\perp}))^{\flat}$, see Section 1.4.
   (d) Let $V_i^m$ be the set of the left singular vectors of $\mathbf{M}^m$, for all $m \in [k]$, whose singular values have absolute value at least $\sigma$.
3. **Return:** A basis of $V = \bigcup_{i=1}^{\ell}\bigcup_{m=1}^{k} V_i^m$.

**Lemma 2.4.10.** *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ with standard normal $\mathbf{x}$-marginal. Let $p : \mathbb{R} \mapsto \mathbb{R}$ be a univariate, mean zero, unit variance polynomial of degree $k$ such that for some unit vector $\mathbf{v}^* \in \mathbb{R}^d$ it holds $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[yp(\mathbf{v}^* \cdot \mathbf{x})] \geq \tau$ for some $\tau \in (0, 1]$. Let $\mathbf{T}'^m$ be an approximation of the order-$m$ Chow-parameter tensor $\mathbf{T}^m$ of $\mathcal{D}$ such that $\|\mathbf{T}'^m - \mathbf{T}^m\|_F \leq \tau/(4\sqrt{k})$. Denote by $V_m$ the subspace spanned by the left*

*singular vectors of* $(\mathbf{T}'^m)^\flat$ *whose singular values are greater than* $\tau/(4\sqrt{k})$. *Moreover, denote by* $V$ *the union of* $V_1, \ldots, V_k$. *Then we have that*

1. $\dim(V) \leq 4k/\tau^2$, *and*
2. $\|\mathrm{proj}_V(\mathbf{v}^*)\|_2 \geq \tau/(4\sqrt{k})$.

*Proof.* Note $\|\mathbf{T}^m\|_F = \sup_{p \in \mathcal{H}_k} \mathbf{E}[yp(\mathbf{x})/\sqrt{\mathbf{E}[p^2(\mathbf{x})]}] \leq \sup_{p \in \mathcal{H}_k} \mathbf{E}[p^2(\mathbf{x})]^{1/2}/\mathbf{E}[p^2(\mathbf{x})]^{1/2} = 1$, where $\mathcal{H}_k$ is the set of polynomials formed from a linear combination of $k$-degree Hermite polynomials. Therefore, by the assumptions of Lemma 2.4.10 and the triangle inequality we obtain that $\|\mathbf{T}'^m\|_F \leq \|\mathbf{T}^m\|_F + 1 \leq 2$.

Recall that by $(\mathbf{T}'^m)^\flat$ we denote the $d \times d^{m-1}$ flattening of $\mathbf{T}'^m$. To simplify notation write $\mathbf{M}^m = (\mathbf{T}'^m)^\flat \in R^{d \times d^{m-1}}$. We have that $V_m$ is the span of the left singular vectors of $\mathbf{M}^m$ with singular value at least $\tau$ and $V$ be the union of all the $V_m$. We first show that the dimension of $V$ is not very large.

**Claim 2.4.11.** *It holds that* $\dim(V) \leq 4k/\tau^2$.

*Proof.* We have that $\|\mathbf{M}^m\|_F = \|\mathbf{T}'^m\|_F \leq 2$. Therefore, since $\|\mathbf{M}^m\|_F$ is equal to the sum of the squares of the singular values $\sigma_1, \ldots, \sigma_d$ of $\mathbf{M}^m$ we obtain that the number of singular values $\sigma_i$ with $|\sigma_i| \geq \tau$ is

$$\frac{1}{\tau^2} \sum_{i=1}^{d} \mathbb{1}\{\sigma_i^2 \geq \tau^2\}\tau^2 \leq \frac{1}{\tau^2} \sum_{i=1}^{d} \sigma_i^2 = \frac{\|\mathbf{M}^m\|_F^2}{\tau^2} \leq \frac{4}{\tau^2}\,.$$

We can do the same calculation for all $\mathbf{T}^m$ to conclude that the dimension of the union of all subspaces is at most $4k/\tau^2$. □

Next, we show that $V$ contains a non-trivial part of the optimal direction $\mathbf{v}^*$, i.e., it holds that $\|\mathrm{proj}_V(\mathbf{v}^*)\|_2 \geq \tau/(4\sqrt{k})$. We first note that since the univariate polynomial

$p$ is mean-zero, it can be written as a linear combination of the univariate Hermite polynomials $h_i(z)$ of degree greater than or equal to 1. Write $p(z) = \sum_{i=1}^{k} c_i h_i(z)$. We note that since $\mathbf{E}_{z\sim\mathcal{N}}[p(z)^2] = 1$, we have that $\sum_{i=1}^{k} c_i^2 = 1$. Moreover, since the order-$i$ Chow tensor is defined as $\mathbf{T}^i = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{H}^i(\mathbf{x})y]$, it holds

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[yp(\mathbf{v}^*\cdot\mathbf{x})] = \sum_{i=1}^{k} c_i \mathbf{T}^i\cdot(\mathbf{v}^*)^{\otimes i} \geq \tau\,.$$

Therefore, by Cauchy-Schwarz, there exists $j\in\{1,\ldots,k\}$ such that $\mathbf{T}^j\cdot(\mathbf{v}^*)^{\otimes j}\geq\tau/\sqrt{k}$. Write $\mathbf{v}^*$ as $\mathbf{v}+\mathbf{u}$ where $\mathbf{v}\in V$ and $\mathbf{u}\in V^\perp$. Moreover, denote $\mathbf{r}\in\mathbb{R}^{d^{j-1}}$ to be the flattening of the tensor $\mathbf{v}^{\otimes(j-1)}$. We note that

$$\begin{aligned}
\mathbf{T}^j\cdot(\mathbf{v}^*)^{\otimes j} &\leq \|\mathbf{T}^j-\mathbf{T}'^j\|_F + |\mathbf{T}'^j\cdot(\mathbf{v}\otimes(\mathbf{v}^*)^{\otimes j-1})| + |\mathbf{T}'^j\cdot(\mathbf{u}\otimes(\mathbf{v}^*)^{\otimes j-1})|\\
&\leq \|\mathbf{T}^j-\mathbf{T}'^j\|_F + |\mathbf{v}^\intercal\mathbf{M}^j\mathbf{r}| + |\mathbf{u}^\intercal\mathbf{M}^j\mathbf{r}|\\
&\leq 2\frac{\tau}{4\sqrt{k}} + \|\mathbf{v}\|_2\|\mathbf{M}^j\|_F \leq \frac{\tau}{2\sqrt{k}} + 2\|\mathbf{v}\|_2\,,
\end{aligned}$$

where we used that $\mathbf{u}$ belongs on the subspace spanned from the left singular vectors of $\mathbf{M}^j$ whose singular values are less than $\tau/(4\sqrt{k})$, and therefore it holds $|\mathbf{u}^\intercal\mathbf{M}^j\mathbf{v}|\leq\tau/(4\sqrt{k})$. Thus, we have $\|\mathbf{v}\|_2\geq\frac{\tau}{4\sqrt{k}}$ and therefore, $\|\mathrm{proj}_V(\mathbf{v}^*)\|_2\geq\tau/(4\sqrt{k})$.

□

The proof of Proposition 2.4.9 follows from an application of Proposition 2.4.3 and Lemma 2.4.10 with the appropriate parameters, which we state below. To conclude the proof, we need the following lemma which bounds the number of samples needed to estimate the order-$m$ Chow parameters.

**Lemma 2.4.12.** *Fix $m\in\mathbb{Z}_+$ and $\epsilon,\delta\in(0,1)$. Let $\mathcal{D}$ be a distribution in $\mathbb{R}^d\times\{\pm1\}$ with*

*standard normal* $\mathbf{x}$*-marginals. There is an algorithm that with* $N = d^{O(m)} \log(1/\delta)/\epsilon^2$ *samples and* $\mathrm{poly}(d, N)$ *runtime, outputs an approximation* $\mathbf{T}'^m$ *of the order-*$m$ *Chow-parameter tensor* $\mathbf{T}^m$ *of* $\mathcal{D}$ *such that with probability* $1 - \delta$*, it holds* $\|\mathbf{T}'^m - \mathbf{T}^m\|_F \leq \epsilon$ .

*Proof.*

**Fact 2.4.13.** *Fix* $m \in \mathbb{Z}_+$ *and* $\epsilon, \delta \in (0, 1)$*. Let* $\mathcal{D}$ *be a distribution in* $\mathbb{R}^d \times \{\pm 1\}$ *and let* $N = d^{O(m)} \log(1/\delta)/\epsilon^2$*. Let* $\alpha$ *be a multi-index satisfying* $|\alpha| \leq m$*. Then there is an algorithm that with* $N$ *samples and runtime poly*$(d, N)$*, computes with probability* $1 - \delta$ *estimates* $\hat{h}_\alpha$ *such that*

$$|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y h_\alpha(\mathbf{x})] - \hat{h}_\alpha| \leq \epsilon \ ,$$

*for all* $|\alpha| \leq m$*.*

*Proof.* Let $\widehat{\mathcal{D}}_N$ be the empirical distribution of $\mathcal{D}$ with $N$ samples. Using Markov's inequality, we have

$$\begin{aligned}
\mathbf{Pr}[|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[h_\alpha(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[h_\alpha(\mathbf{x})y]| \geq \epsilon] &\leq \frac{1}{N\epsilon^2}\mathbf{Var}[h_\alpha(\mathbf{x})y] \\
&\leq \frac{1}{N\epsilon^2}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[h_\alpha^2(\mathbf{x})] \\
&\leq O\left(\frac{1}{N\epsilon^2}\right) \ ,
\end{aligned}$$

where we used the fact that for the Hermite polynomials it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_d}[h_\alpha^2(\mathbf{x})] = 1$. Moreover, all possible choices of $\alpha$ are at most $d^m$. Hence, using the fact that $N = O(d^m/\epsilon^2)$ and applying a union bound for all different $h_\alpha$, we get that with constant probability it holds

$$|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y h_\alpha(\mathbf{x})] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[y h_\alpha(\mathbf{x})]| \leq \epsilon \ ,$$

for all $|\alpha| \leq m$. By applying a standard probability amplification technique (see, e.g., Ex. 1, Chapter 13 of [SSBD14]), we can boost the confidence to $1-\delta$ with $N' = O(N \log(1/\delta))$ samples. □

In order to estimate up to order-$m$ Chow tensors, with accuracy $\epsilon$, we need to learn each order-$m$ Chow parameters up to accuracy $\epsilon^2 d^{-m}$, therefore we need to $d^{O(m)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ samples. □

Recall that from our assumptions $|b| \leq \sqrt{\log(1/\zeta)}$. Moreover, from Proposition 2.4.3, we have that for $k = O(\log(1/\zeta))$, there exists a degree-$k$ polynomial $p$, such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[yp(\mathbf{v}^* \cdot \mathbf{x})] \geq \mathrm{poly}(\zeta)$. Thus, in order to apply Lemma 2.4.10, we need to estimate the first order-$k$ Chow tensors. Hence, the sample complexity is $N = d^{O(\log(1/\zeta))}\mathrm{poly}(1/\zeta)\log(1/\delta)$ and the runtime is $\mathrm{poly}(N, d)$. This completes the proof of Proposition 2.4.9.

□

Combining Lemma 2.4.6 and Proposition 2.4.9, we can prove Proposition 2.4.2.

*Proof of Proposition 2.4.2.* From the assumption that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y] \geq \mathrm{OPT} + \epsilon$, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq f(\mathbf{x})\}] \geq \epsilon$. To prove this, note that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y] = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq f(\mathbf{x})\}] + \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})]$ and $\mathrm{OPT} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})]$, by combining with $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y] \geq \mathrm{OPT} + \epsilon$, it follows that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq f(\mathbf{x})\}] \geq \epsilon\,.$$

Let $\widehat{\mathcal{D}}_N$ be the empirical distribution of $\mathcal{D}$ using $N = d^{\Theta(\log(1/\epsilon))}\log(1/\delta)$ samples. Let $\rho = C\epsilon^3$, for some small enough constant $C > 0$, and from Lemma 2.4.6, we get that there exists a set $B = \{t_1 \leq \mathbf{v}\cdot\mathbf{x} \leq t_2\}$, with $|t_1 - t_2| = \rho$, for which we

can apply the algorithm of Proposition 2.4.9 to the distribution $\mathcal{D}_B^{\perp \mathbf{w}}$ ($\mathcal{D}_B^{\perp \mathbf{w}}$ is the distribution $\widehat{\mathcal{D}}_N$ conditioned on $B$ and projected into $\mathbf{w}^\perp$), and get a vector space $V$ with the following properties: (i) $V \subseteq \mathbf{w}^\perp$, (ii) the dimension of $V$ is $\mathrm{poly}(1/\epsilon)$, and (iii) $\|\mathrm{proj}_V(\mathbf{w}^*)\|_2 = \mathrm{poly}(\epsilon\rho) = \mathrm{poly}(\epsilon)$ with probability $1-\delta$, if $N \geq d^{O(\log(1/\epsilon))}\log(1/\delta)$. Moreover, the algorithm runs in $\mathrm{poly}(N, d)$ runtime.

The problem, we need to overcome, is that we do not know the set $B$, so we are going to apply the algorithm of Proposition 2.4.9 to any possible set and take the union of the outputs. Let $l = \Theta(\log(1/\epsilon)/\rho)$ and let $t_i = i\rho$ and $t_{-i} = -i\rho$, for $0 \leq i \leq l$. Furthermore, we define $B_i = \{t_i \leq \mathbf{v}\cdot\mathbf{x} \leq t_{i+1}\}$ and $B_{-i} = \{-t_{i+1} \leq \mathbf{v}\cdot\mathbf{x} \leq -t_i\}$ (similarly to Claim 2.4.8). We then apply the algorithm of Proposition 2.4.9 to each of the distributions $\mathcal{D}_{B_i}^{\perp \mathbf{w}}$ and get vector spaces $V_i$. Moreover, the algorithm of Proposition 2.4.9 returns vector spaces $V_i$ such that $V_i \subseteq \mathbf{w}^\perp$ and $\dim(V_i) = \mathrm{poly}(1/\epsilon)$. Finally, from Lemma 2.4.6, we know that there exists an index $j \in \{-l, \ldots, l\}$ such that applying the algorithm Proposition 2.4.9 on the distribution $\mathcal{D}_{B_j}^{\perp \mathbf{w}}$ gives a vector space $V_j$, with the additional property that $\|\mathrm{proj}_{V_j}((\mathbf{w}^*)^{\perp \mathbf{w}})\|_2 = \mathrm{poly}(\epsilon)$, with probability $1-\delta$. Thus, we let $V$ to be the union of all subspaces $V_i$, i.e., $V = \bigcup_{i=-l}^{l} V_i$. It holds that $\dim(V) \leq \sum_{i=1}^{l}\dim(V_i) = \mathrm{poly}(1/\epsilon)$. Moreover, for any $\mathbf{v} \in V$, it holds $\mathbf{v}\cdot\mathbf{w} = 0$, this is because $V_i \subseteq \mathbf{w}^\perp$ and hence $V \subseteq \mathbf{w}^\perp$. Finally, we have that $\|\mathrm{proj}_{V_j}(\mathbf{w}^*)\|_2 = \mathrm{poly}(\epsilon)$ and because $V_j \subseteq V$ it holds that $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp \mathbf{w}})\|_2 \geq \mathrm{poly}(\epsilon)$ with probability $1-\delta$. □

We now show that by finding a vector $\mathbf{u}$ that correlates well with $(\mathbf{v}^*)^{\perp \mathbf{v}}$, we can update our current guess vector $\mathbf{v}$ and get one with increased correlation with $\mathbf{v}^*$.

**Lemma 2.4.14** (Correlation Improvement)**.** *Fix unit vectors $\mathbf{v}^*, \mathbf{v} \in \mathbb{R}^d$. Let $\mathbf{u} \in \mathbb{R}^d$ such that $\mathbf{u}\cdot\mathbf{v}^* \geq c$, $\mathbf{u}\cdot\mathbf{v} = 0$, and $\|\mathbf{u}\|_2 \leq 1$, with $c > 0$. Then, for $\mathbf{v}' = \frac{\mathbf{v}+\lambda\mathbf{u}}{\|\mathbf{v}+\lambda\mathbf{u}\|_2}$, with $\lambda = c/2$, we have that $\mathbf{v}'\cdot\mathbf{v}^* \geq \mathbf{v}\cdot\mathbf{v}^* + \lambda^2/2$.*

*Proof.* We will show that $\mathbf{v}' \cdot \mathbf{v}^* = \cos\theta' \geq \cos\theta + \lambda^2/2$, where $\cos\theta = \mathbf{v} \cdot \mathbf{v}^*$. We have that

$$\|\mathbf{v} + \lambda\mathbf{u}\|_2 = \sqrt{1 + \lambda^2 \|\mathbf{u}\|_2^2 + 2\lambda \mathbf{u} \cdot \mathbf{v}} \leq 1 + \lambda^2 \|\mathbf{u}\|_2^2 \ , \tag{2.44}$$

where we used that $\sqrt{1+a} \leq 1 + a/2$. Using the update rule, we have

$$\mathbf{v}' \cdot \mathbf{v}^* = \mathbf{v}' \cdot (\mathbf{v}^*)^{\perp_\mathbf{v}} \sin\theta + \mathbf{v}' \cdot \mathbf{v} \cos\theta = \frac{\lambda \mathbf{u} \cdot (\mathbf{v}^*)^{\perp_\mathbf{v}}}{\|\mathbf{v} + \lambda\mathbf{u}\|_2} \sin\theta + \frac{\mathbf{v} + \lambda \mathbf{u} \cdot \mathbf{v}}{\|\mathbf{v} + \lambda\mathbf{u}\|_2} \cos\theta \ .$$

Now using Equation (2.44), we get

$$\mathbf{v}' \cdot \mathbf{v}^* \geq \frac{\lambda \mathbf{u} \cdot (\mathbf{v}^*)^{\perp_\mathbf{v}}}{1 + \lambda^2 \|\mathbf{u}\|_2^2} \sin\theta + \frac{\cos\theta}{1 + \lambda^2 \|\mathbf{u}\|_2^2} = \cos\theta + \frac{\lambda \mathbf{u} \cdot (\mathbf{v}^*)^{\perp_\mathbf{v}}}{1 + \lambda^2 \|\mathbf{u}\|_2^2} \sin\theta + \frac{-\lambda^2 \|\mathbf{u}\|_2^2 \cos\theta}{1 + \lambda^2 \|\mathbf{u}\|_2^2} \ .$$

Then, using that $\mathbf{u} \cdot \mathbf{v}^* = \mathbf{u} \cdot (\mathbf{v}^*)^{\perp_\mathbf{v}} \sin\theta$, we have that $\mathbf{u} \cdot (\mathbf{v}^*)^{\perp_\mathbf{v}} \geq \frac{c}{\sin\theta}$, thus

$$\mathbf{v}' \cdot \mathbf{v}^* \geq \cos\theta + \frac{\lambda c - \lambda^2 \|\mathbf{u}\|_2^2}{1 + \lambda^2 \|\mathbf{u}\|_2^2} \geq \cos\theta + \frac{\lambda c - \lambda^2}{1 + \lambda^2 \|\mathbf{u}\|_2^2} = \cos\theta + \frac{1}{4} \frac{c^2}{1 + \lambda^2 \|\mathbf{u}\|_2^2} \ ,$$

where in the first inequality we used that $\|\mathbf{u}\|_2 \leq 1$ and in the second that for $\lambda = c/2$ it holds $c - \lambda \geq c/2$. Finally, we have that

$$\cos\theta' = \mathbf{v}' \cdot \mathbf{v}^* \geq \cos\theta + \frac{1}{4} \frac{c^2}{1 + \lambda^2 \|\mathbf{u}\|_2^2} \geq \cos\theta + \frac{1}{8} c^2 = \cos\theta + \frac{1}{2} \lambda^2 \ .$$

This completes the proof. □

We need the following standard fact that bounds from below the correlation of any vector with a random one.

**Fact 2.4.15** (see, e.g., Remark 3.2.5 of [Ver18])**.** *Let* $\mathbf{v}$ *be a unit vector in* $\mathbb{R}^d$*. For a random unit vector* $\mathbf{u} \in \mathbb{R}^d$*, with at least* $1/3$ *probability, it holds* $\mathbf{v} \cdot \mathbf{u} \gtrsim 1/\sqrt{d}$*.*

**Algorithm 3** The Biased Random Walk for Learning Halfspaces with General Massart Noise, see Theorem 2.4.1.

**Input:**

1. $\epsilon, \delta > 0$.
2. Sample access to $\mathcal{D}$ with standard normal $\mathbf{x}$-marginal which satisfies the general Massart noise condition with respect to the hypothesis $f(\mathbf{x})$.

**Output:** A hypothesis $h(\mathbf{x})$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, with probability $1 - \delta$.

**Define:** $L = \{\}$, $N = d^{C\log(1/\epsilon)}\log(1/\delta)$, $\lambda = \mathrm{poly}(\epsilon)/C$, $T = 2^{C\mathrm{poly}(1/\epsilon)}$, for $C > 0$ a sufficiently large constant.

1. Initialize $\mathbf{w} = \mathbf{e}_1$.
2. Draw $N$ samples and compute to be the empirical distribution $\widehat{D}$.
3. Repeat $T$ times
   (a) Using $\mathbf{w}$ on Algorithm 2, generate vector space $V$.
   (b) Pick a random unit vector $\mathbf{v} \in V$.
   (c) Update current hypothesis $\mathbf{w}$ by $\mathbf{w} \leftarrow \frac{\mathbf{w}+\lambda\mathbf{v}}{\|\mathbf{w}+\lambda\mathbf{v}\|_2}$ .
   (d) Update list of vectors $L \leftarrow L \cup \{\mathbf{w}\}$.
4. **Return:** $h(\mathbf{x}) = \mathrm{sign}(\hat{\mathbf{w}} \cdot \mathbf{x})$, where $\hat{\mathbf{w}} \in S$ and minimizes the error with respect $y$, i.e.,
$$\hat{\mathbf{w}} \leftarrow \operatorname*{argmin}_{\mathbf{w}\in L} \mathbf{Pr}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y] \ .$$

We are now ready to prove the Theorem 2.4.1.

*Proof of Theorem 2.4.1.* First, let $\widehat{\mathcal{D}}_N$ be the empirical distribution of $\mathcal{D}$ using $N \in \mathbb{Z}_+$ samples. Let $\mathbf{w}^{(i)}$ be the current guess. If $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^{(i)}\cdot\mathbf{x}) \neq y] \geq \mathrm{OPT} + \epsilon$ and $\theta(\mathbf{w}^{(i)}, \mathbf{w}^* \leq \pi - \epsilon)$, then from Proposition 2.4.2 with $N_1 = d^{O(\log(1/\epsilon))}\log(1/\delta_1)$ samples from $\widehat{\mathcal{D}}$ and $\mathrm{poly}(N_1, d)$ time, we compute a subspace $V$ such that $\|\mathrm{proj}_V(\mathbf{w}^*)\|_2 \geq$

$\text{poly}(\epsilon)$ and from Fact 2.4.15, we get a random a unit vector $\mathbf{v} \in V$ such that $\mathbf{v} \cdot \mathbf{w}^* = \text{poly}(\epsilon)$ and $\mathbf{v} \cdot \mathbf{w}^{(i)} = 0$, with probability $(1-\delta_1)/3$. We call this event $\mathcal{E}_i$. By conditioning on the event $\mathcal{E}_i$; from Lemma 2.4.14, after we update our current hypothesis vector $\mathbf{w}^{(i)}$ with $\mathbf{v}$, we get a unit vector $\mathbf{w}^{(i+1)}$ such that $\mathbf{w}^{(i+1)} \cdot \mathbf{w}^* \geq \mathbf{w}^{(i+1)} \cdot \mathbf{w}^* + \text{poly}(\epsilon)$.

After running the update step $k$ times and conditioning on the events $\mathcal{E}_1, \ldots, \mathcal{E}_k$, then $\mathbf{w}^{(k)} \cdot \mathbf{w}^* \geq \mathbf{w}^{(1)} \cdot \mathbf{w}^* + k\,\text{poly}(\epsilon)$; therefore, for $k = \text{poly}(1/\epsilon)$, we get that the vector $\mathbf{w}^{(k)}$ that is competitive with the optimal hypothesis, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\text{sign}(\mathbf{w}^{(k)} \cdot \mathbf{x}) \neq y] \leq \text{OPT} + \epsilon$ or $\theta(\mathbf{w}^{(i)}, \mathbf{w}^* \in (\pi - \epsilon, \pi)$ which means that $-\mathbf{w}^{(k)}$ is competitive with the optimal hypothesis, see Fact 2.2.12. The probability that all the events $\mathcal{E}_1, \ldots, \mathcal{E}_k$ hold simultaneously is at least $(1 - k\delta_1) + (1/3)^k$, and thus by choosing $\delta_1 \leq 1/(3k)$, the probability of success is at least $\delta_2 = (1/3)^k$. By running the algorithm above $M = \log(1/\delta)/\delta_2$ times, we get a list of $2M$ vectors, that list contains all the $\mathbf{w}^{(i)}$ that generated in every step of the algorithm and the $-\mathbf{w}^{(i)}$. By applying Hoeffding's inequality, we get that the list $L$ of $2M$ vectors contains a unit vector $\mathbf{w}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\text{sign}(\mathbf{w}^{(k)} \cdot \mathbf{x}) \neq y] \leq \text{OPT} + \epsilon$, with probability $1 - \delta/2$. Finally, to evaluate all the vectors from the list, we need a few samples, from the distribution $\mathcal{D}$ to obtain the best among them, i.e., the one that minimizes the zero-one loss.

The size of the list of candidates is at most $M \leq 2^{\text{poly}(1/\epsilon)} \log(1/\delta)$. Therefore, from Hoeffding's inequality, it follows that $O(\text{poly}(1/\epsilon) \log(1/\delta))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most $\epsilon$ with probability at least $1 - \delta/2$. Thus, with $N = d^{\log(1/\epsilon)} \log(1/\delta)$ samples and $\text{poly}(N, d, 2^{\text{poly}(1/\epsilon)})$ runtime, we get a hypothesis $\hat{\mathbf{w}}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\text{sign}(\hat{\mathbf{w}} \cdot \mathbf{x}) \neq y] \leq \text{OPT} + \epsilon$ with probability $1 - \delta$. This completes the proof. □

# Chapter 3

# Learning Margin Halfspaces with Massart Noise

## 3.1 Introduction

This chapter addresses the fine-grained complexity of learning halfspaces with a margin under the Massart noise model [MN06]. We develop a new, computationally efficient algorithm whose sample complexity nearly matches the information-theoretic lower bound for this task. Our approach circumvents the limitations of prior methods by employing a novel sequence of reweighted convex loss functions.

Our main result (Theorem 3.2.1) is an online SGD based algorithm that achieves an error of $\eta + \epsilon$ with a sample complexity of $\tilde{O}(1/(\gamma^2\epsilon^2))$. This guarantee nearly matches the known computational lower bounds for this problem and significantly improves upon the $\tilde{\Omega}(1/(\gamma^4\epsilon^3))$ complexity of previous efficient algorithms.

Formally concept class of halfspaces with a margin is defined as follows.

**Definition 3.1.1** ($\gamma$-Margin Halfspaces)**.** *Let $\mathcal{D}$ be a distribution over $\mathbb{S}^{d-1} \times \{\pm 1\}$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. Let $\mathbf{w}^* \in \mathbb{S}^{d-1}$ and $\gamma \in (0,1)$. We say that the distribution $\mathcal{D}$ satisfies the $\gamma$-margin condition with respect the halfspace* $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$[1]*, if (i) for $(\mathbf{x}, y) \sim D$, we have that $y = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$, and (ii) $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{w}^* \cdot \mathbf{x}| < \gamma] = 0$. The parameter $\gamma$ is called the margin of the halfspace* $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$.

Our main result is the following:

**Theorem 3.1.2** (Main Result, Informal)**.** *Let $D$ be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ that satisfies the $\eta$-Massart noise condition with respect to an unknown $\gamma$-margin halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. There is algorithm that draws $n = \tilde{O}(1/(\epsilon^2\gamma^2))$ samples from $D$, runs in time $\tilde{O}(dn/\epsilon)$, and with probability at least $9/10$ returns a vector $\hat{\mathbf{w}}$ such that $\mathrm{err}_{\mathcal{D}}(\hat{\mathbf{w}}) \leq \eta + \epsilon$.*

The sample upper bound of Theorem 3.1.2 nearly matches the computational sample complexity of the problem (for SQ algorithms and low-degree polynomial tests), which was shown to be $\Omega(1/(\epsilon^2\gamma) + 1/(\epsilon\gamma^2))$ [MN06, DDK+23a, DDK+23b]. That is, Theorem 3.1.2 comes close to resolving the fine-grained complexity of this basic task. Moreover, it matches known algorithmic guarantees for the easier case of Random Classification Noise [DDK+23a, KIT+23].

**Related and Prior Work** The computational problem of learning halfspaces with Massart noise has been extensively studied, both in the distribution-specific and the distribution-free settings.

In the distribution-specific setting, the first efficient algorithm for homogeneous Massart halfspaces was given in [ABHU15]. Subsequent work generalized this result in

[1]We will henceforth assume that the threshold is $\theta = 0$, which is well-known to be no loss of generality.

various directions [ABHZ16, ZLC17, YZ17, DKTZ20a, DKTZ20b, DKK+20, DKK+21b, DKK+22].

The first algorithmic progress in the distribution-free setting was made by [DGT19], answering a longstanding open problem [Slo88, Slo92, Blu03]. Subsequent work gave an algorithm with improved sample complexity [CKMY20] and provided strong evidence that an error of $\eta+\epsilon$ is the best to hope for in polynomial time [DK20a, NT22, DKMR22a] (in both the Statistical Query model and under plausible cryptographic assumptions). In a related direction, [DIK+21] gave the first efficient boosting algorithm in the presence of Massart noise, which can boost a weak learner to one with error $\eta + \epsilon$. Finally, we note that natural generalizations of the Massart model to learning real-valued functions (in an essentially distribution-free setting) have also been studied [CKMY21, DPT21, DKRS22].

Very recent work [DDK+23a] gave SQ (and low-degree polynomial testing) lower bounds for learning $\gamma$-margin halfspaces with RCN [AL88], which is a special case of Massart noise. Specifically, [DDK+23a] showed that any efficient SQ algorithm for the problem requires sample complexity $\Omega(1/(\gamma^{1/2}\epsilon^2))$. Subsequently, [DDK+23b] showed a related SQ lower bound under the Gaussian distribution, which can be adapted to obtain a lower bound of $\Omega(1/(\gamma\epsilon^2))$ for the margin setting.

Independent work [CKST24] obtained a learning algorithm for $\gamma$-margin halfspaces with essentially the same sample and computational complexity as ours.

**Notation** We use $\mathrm{err}_D(\mathbf{w})$ to denote the 0-1 error of the halfspace defined by the weight vector $\mathbf{w}$ with respect to the distribution $D$, i.e., $\mathrm{err}_D(\mathbf{w}) \stackrel{\mathrm{def}}{=} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y]$. We will use $\mathrm{err}(\mathbf{w},\mathbf{x})$ for the 0-1 error of $\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})$ conditioned on $\mathbf{x}$, i.e., $\mathrm{err}(\mathbf{w},\mathbf{x}) := \mathbf{Pr}_{y\sim D_y(\mathbf{x})}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) \neq y]$. Note that $\mathrm{err}_D(\mathbf{w}) = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathrm{err}(\mathbf{w},\mathbf{x})]$. If $D$

satisfies the $\eta$-Massart noise condition with respect to the halfspace $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})$, then $\mathrm{err}(\mathbf{w}, \mathbf{x}) = \eta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})\} + (1 - \eta(\mathbf{x}))\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})\}$ .

## 3.2 Our Algorithm and its Analysis: Proof of Theorem 3.1.2

In this section, we prove our main result. Algorithm 4 efficiently learns the class of margin halfspaces on the unit ball, in the presence of Massart noise, with sample complexity nearly matching the information-computation limit. Additionally, its runtime is linear in the sample size, excluding a final testing step to select the best hypothesis.

At a high-level, our algorithm leverages a carefully selected convex loss (or, more precisely, a sequence of convex losses) — serving as a proxy to the 0-1 error. A common loss function, introduced in this context by [DGT19] and leveraged in [DGT19, CKMY20], is the LeakyReLU function. This is the univariate function $\mathrm{LeakyReLU}_\lambda(t) = (1 - \lambda)\mathbb{1}\{t \geq 0\}t + \lambda\mathbb{1}\{t < 0\}t$, where $\lambda \in (0, 1)$ is the leakage parameter (that needs to be selected carefully). Roughly speaking, the convex function $\ell_\lambda(\mathbf{w}, \mathbf{x}, y) = \mathrm{LeakyReLU}_\lambda(-y(\mathbf{w} \cdot \mathbf{x}))$ can be viewed as a reasonable proxy to the 0-1 loss of the halfspace $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})$ on the point $(\mathbf{x}, y)$. To see this, note that

$$\ell_\lambda(\mathbf{w}, \mathbf{x}, y) = (\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y\} - \lambda)|\mathbf{w} \cdot \mathbf{x}| \; . \tag{3.1}$$

Observe that a point $\mathbf{x}$ that is classified correctly by the halfspace $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})$ will satisfy

$$\Big(\mathbf{E}_{y \sim D_y(\mathbf{x})}[\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y\}] - \lambda\Big)|\mathbf{w} \cdot \mathbf{x}| = (\eta(\mathbf{x}) - \lambda)|\mathbf{w} \cdot \mathbf{x}|$$

which is non-positive for $\lambda \geq \eta(\mathbf{x})$. Since the only guarantee we have is that $\eta(\mathbf{x}) \leq \eta$, this suggests that we need to select $\lambda \geq \eta$. It turns out that $\lambda := \eta$ is the optimal choice. We fix the choice of $\lambda := \eta$ throughout. On the other hand, if (the halfspace defined by) $\mathbf{w}$ misclassifies the point $\mathbf{x}$, this term becomes non-negative.

The factor $|\mathbf{w} \cdot \mathbf{x}|$ in Equation (3.1) reweights the 0-1 error so that points $\mathbf{x}$ for which $|\mathbf{w} \cdot \mathbf{x}|$ is sufficiently large (i.e., close to 1) have to be classified correctly by a minimizer of $\mathbf{E}_{(\mathbf{x},y)\sim D}[\ell_\lambda(\mathbf{w}, \mathbf{x}, y)]$. On the other hand, points closer to the separating hyperplane defined by $\mathbf{w}$, or points where $\eta(\mathbf{x})$ is close to $\lambda = \eta$, are not guaranteed to be classified correctly by the minimizer of this loss. We leverage this insight to construct a sequence of loss functions that reweight the points so that, to minimize the regret, we need to classify a large fraction of points; this leads to the desired error of $\eta + \epsilon$ with near-optimal sample complexity.

We now provide some intuition justifying our choice of surrogate loss functions. Observe that if we instead could minimize the function

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell_\lambda(\mathbf{w}, \mathbf{x}, y)/|\mathbf{w} \cdot \mathbf{x}|] = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathbb{1}\{\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) \neq y\} - \lambda)] \,, \tag{3.2}$$

with respect to $\mathbf{w}$, we would obtain a halfspace with minimum 0-1 error; unfortunately, this reweighted loss is just a shift of the 0-1 loss, hence non-convex. To fix this issue, instead of reweighting by $1/|\mathbf{w} \cdot \mathbf{x}|$, we will reweight by $W(\mathbf{v} \cdot \mathbf{x}, \gamma) \stackrel{\mathrm{def}}{=} 1/\max(|\mathbf{v} \cdot \mathbf{x}|, \gamma)$, where $\gamma$ is the margin parameter and $\mathbf{v}$ is an appropriately chosen vector that is independent of $\mathbf{w}$. The new loss is defined as follows:

$$\mathcal{L}_{\lambda,\mathbf{v}}(\mathbf{w}) \stackrel{\mathrm{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell_\lambda(\mathbf{w}, \mathbf{x}, y) W(\mathbf{v} \cdot \mathbf{x}, \gamma/2)] \,, \tag{3.3}$$

where for technical reasons we use $\gamma/2$ instead of $\gamma$ in the maximum.

Since the parameter $\mathbf{v}$ is independent of $\mathbf{w}$, the loss $\mathcal{L}_{\lambda,\mathbf{v}}(\mathbf{w})$ remains convex in $\mathbf{w}$. At the same time, by carefully choosing $\mathbf{v}$, we can accurately simulate the non-convex 0-1 loss. Note that our reweighting term is a maximum over two terms. The reason for this choice is that, for some points $\mathbf{x}$, the quantity $|\mathbf{v}\cdot\mathbf{x}|$ can be arbitrarily small; taking the maximum avoids the loss becoming very large. In particular, the loss $\mathcal{L}_{\lambda,\mathbf{v}}(\mathbf{w})$ will be guaranteed to remain in a bounded length interval.

Our algorithm proceeds in a sequence of iterations. In the $(t+1)$-st iteration, it sets $\mathbf{v}$ to be $\mathbf{w}^t$, where $\mathbf{w}^t$ is the weight vector of step $t$. This choice attempts to simulate the 0-1 error at $\mathbf{w}^t$, as is suggested by Equation (3.2). Assume for simplicity that our current hypothesis is the halfspace defined by $\mathbf{w}$ and is such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}|\le\gamma/2\}]=0$. Note this implies that $W(\mathbf{w}\cdot\mathbf{x},\gamma/2)=1/|\mathbf{w}\cdot\mathbf{x}|$. By combining Equations (3.2) and (3.3), we get that $\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w})=\mathrm{err}_\mathcal{D}(\mathbf{w})-\lambda$; note that as long as $\mathrm{err}_\mathcal{D}(\mathbf{w})\ge\lambda+\epsilon$, we have that $\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w})\ge\epsilon$. On the other hand, the optimal halfspace $\mathbf{w}^*$ achieves a non-positive loss; from Equations (3.1) and (3.2), we have that

$$\begin{aligned}\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w}^*) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathbb{1}\{\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})\neq y\}-\lambda)|\mathbf{w}^*\cdot\mathbf{x}|W(\mathbf{w}\cdot\mathbf{x},\gamma/2)]\\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\eta(\mathbf{x})-\lambda)|\mathbf{w}^*\cdot\mathbf{x}|W(\mathbf{w}\cdot\mathbf{x},\gamma/2)]\le 0\,,\end{aligned}$$

where the inequality follows from the fact that $\eta(\mathbf{x})\le\eta$. Recalling that $\mathcal{L}_{\lambda,\mathbf{v}}(\mathbf{w})$ is convex, if we run an Online Convex Optimization (OCO) algorithm, after $T$ steps we are guaranteed to find a vector $\mathbf{w}$ such that $\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w})-\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w}^*)\le O(1/\sqrt{T})$. For $T=O(1/\epsilon^2)$, this gives that $\mathcal{L}_{\lambda,\mathbf{w}}(\mathbf{w})<\epsilon/2$; and therefore we would have $\mathrm{err}_D(\mathbf{w})<\lambda+\epsilon$. We provide an approach using this idea and the cutting planes algorithm in Section B.1

that achieves sample complexity $\widetilde{O}(1/(\epsilon^2\gamma^4))$.

Our algorithm and its analysis work only with the gradient of $\mathcal{L}_{\lambda,\mathbf{v}}(\mathbf{w})$. The key novelty is the analysis of the sample complexity. The gradient of $\ell_\lambda(\mathbf{w},\mathbf{x},y)W(\mathbf{v}\cdot\mathbf{x},\gamma)$ with respect to $\mathbf{w}$ has the following explicit form:

$$\mathbf{g}_{\lambda,\gamma}(\mathbf{w},\mathbf{v},\mathbf{x},y) \stackrel{\mathrm{def}}{=} ((1-2\lambda)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})-y)W(\mathbf{v}\cdot\mathbf{x},\gamma)\mathbf{x} = \frac{((1-2\lambda)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})-y)}{\max(|\mathbf{v}\cdot\mathbf{x}|,\gamma)}\mathbf{x}\,.$$

Furthermore, we denote by $\mathbf{G}_{\mathcal{D}}(\mathbf{w},\mathbf{v},\eta,\gamma) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}_{\eta,\gamma}(\mathbf{w},\mathbf{v},\mathbf{x},y)]$.

Before describing our algorithm and proving Theorem 3.2.1, we simplify our notation. We will omit the parameters $\eta,\gamma$ from the function input (as they are fixed throughout). Therefore, we use $\mathbf{G}_{\widehat{D}_N^t}(\mathbf{w},\mathbf{v}) \equiv \mathbf{G}_{\widehat{D}_N^t}(\mathbf{w},\mathbf{v},\eta,\gamma)$ and $\mathbf{g}(\mathbf{w},\mathbf{v},\mathbf{x},y) \equiv \mathbf{g}_{\eta,\gamma/2}(\mathbf{w},\mathbf{v},\mathbf{x},y)$.

Our algorithm is described in pseudocode below.

**Algorithm 4** Learning Margin Halfspaces with Massart Noise

**Input:** Sample access to distribution $D$ on $\mathbb{S}^{d-1}\times\{\pm1\}$ corrupted with $\eta$-Massart noise w.r.t. $\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})$ satisfying the $\gamma$-margin condition; parameters $\varepsilon,\delta\in(0,1)$; integers $N,T$.
Let $c>0$ be a sufficiently small universal constant.
Initialize $t\leftarrow 0$, $\mathbf{w}^0\leftarrow\mathbf{e}_1=(1,0,\ldots,0)$,

$$T \leftarrow \frac{1}{c}\frac{\log(1/\delta)}{\varepsilon^2\,\gamma^2}.$$

**while** $t\le T$ **do**
  Draw $(\mathbf{x}^{(t)},y^{(t)})\sim D$.
  Set $\lambda_t\leftarrow c\,\gamma^2\,\varepsilon$.
  $\mathbf{v}^{t+1} \leftarrow \mathbf{w}^t - \lambda_t\,\mathbf{g}\big(\mathbf{w}^t,\mathbf{w}^t,\mathbf{x}^{(t)},y^{(t)}\big)$, $\mathbf{w}^{t+1}\leftarrow\dfrac{\mathbf{v}^{t+1}}{\max\big(\|\mathbf{v}^{t+1}\|_2,1\big)}$
  $t\leftarrow t+1$.
**end while**
Draw $N$ samples from $D$ to form empirical distribution $\widehat{D}_N$.
$\hat{\mathbf{w}} \leftarrow \arg\min_{t\in\{0,\ldots,T\}} \mathrm{err}_{\widehat{D}_N}\big(\mathbf{w}^t\big)$
**Return:** $\hat{\mathbf{w}}$

Algorithm 4 employs online SGD applied to a sequence of convex loss functions. We show that, after a certain number of iterations, the algorithm will find a weight vector achieving 0-1 error at most $\eta + \epsilon$. Since the desired vector may not be the last iterate, in the end, our algorithm returns the halfspace that achieves the smallest empirical 0-1 error.

We establish the following result, which implies Theorem 3.1.2.

**Theorem 3.2.1** (Main Result)**.** *Let $D$ be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ satisfying the $\eta$-Massart noise condition with respect to the $\gamma$-margin halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Given $N = \Theta(\log(1/(\gamma\delta))/\epsilon(1-2\eta))$ and $T = \Theta(\log(1/\delta)/(\epsilon^2\gamma^2))$, Algorithm 4 returns a vector $\hat{\mathbf{w}}$ such that $\mathrm{err}_{\mathcal{D}}(\hat{\mathbf{w}}) \leq \eta + \epsilon$ with probability at least $1-\delta$. The algorithm draws $n = O(N+T)$ samples from $D$ and runs in $O(dNT)$ time.*

The rest of this section is devoted to the proof of Theorem 3.2.1.

Our algorithm sets $\mathbf{v} = \mathbf{w}^t$ in each round, therefore for the rest of the section we proceed by setting $\mathbf{v} = \mathbf{w}$ as arguments of $\mathbf{g}$ and $\mathbf{G}$.

We decompose the stochastic gradient $\mathbf{g}(\mathbf{w}, \mathbf{w}, \mathbf{x}, y)$ into two parts: $\mathbf{g}(\mathbf{w}, \mathbf{w}, \mathbf{x}, y) = \mathbf{g}^1(\mathbf{w}, \mathbf{x}) + \mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)$, where

$$\mathbf{g}^1(\mathbf{w}, \mathbf{x}) = \left((1-2\eta)\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) - \mathop{\mathbf{E}}_{y\sim D_y(\mathbf{x})}[y]\right) W(\mathbf{w}\cdot\mathbf{x}, \gamma/2)\mathbf{x}$$

and

$$\mathbf{g}^2(\mathbf{w}, \mathbf{x}, y) = \left(\mathop{\mathbf{E}}_{y\sim D_y(\mathbf{x})}[y] - y\right) W(\mathbf{w}\cdot\mathbf{x}, \gamma/2)\mathbf{x}\,.$$

We also use $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w})$ and $\mathbf{G}^2_{\widehat{D}_N}(\mathbf{w})$ for the same decomposition after taking the empirical expectation, i.e., $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) = \mathbf{E}_{\mathbf{x}\sim(\widehat{D}_{\mathbf{x}})_N}[\mathbf{g}^1(\mathbf{w}, \mathbf{x})]$ and $\mathbf{G}^2_{\widehat{D}_N}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\widehat{D}_N}[\mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)]$.

This serves to decompose the gradient into two parts: one containing the population

expectation over the random variable $y$, and the other containing the error between the empirical estimation of $y$ and the population version of $y$. The vector $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w})$ contains the direction that will decrease the distance between $\mathbf{w}$ and $\mathbf{w}^*$, while $\mathbf{G}^2_{\widehat{D}_N}(\mathbf{w})$ contains the estimation error. To see this, observe that if we take the population expectation of $\mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)$, we will have:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\left((1-2\eta(\mathbf{x}))\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}) - \mathop{\mathbf{E}}_{y\sim D_y(\mathbf{x})}[y]\right)W(\mathbf{w}\cdot\mathbf{x}, \gamma/2)\mathbf{x}\right] = 0 \ ,$$

where we used that $\mathbf{E}_{y\sim D_y(\mathbf{x})}[y] = (1-2\eta(\mathbf{x}))\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})$.

We start by bounding the contribution of $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w})$ in the direction $\mathbf{w}-\mathbf{w}^*$. We show that if instead of the corrupted label $y$ at the point $\mathbf{x}$, we had access to $\mathbf{E}_{y\sim D_y(\mathbf{x})}[y] = (1-2\eta(\mathbf{x}))\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})$, then the gradient has a large component in the direction of $\mathbf{w}-\mathbf{w}^*$. This effectively implies that $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w})$ can be used as a separation oracle, separating all the halfspaces with 0-1 error more than $\eta+\epsilon$ from the ones with smaller error.

**Lemma 3.2.2** (Structural Lemma)**.** *Let $N \in \mathbb{Z}_+$ and let $D$ be a distribution on $\mathbb{S}^{d-1}\times\{\pm 1\}$ satisfying the $\eta$-Massart condition with respect to the optimal classifier $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x})$. Let $\mathbf{w}\in\mathbb{R}^d$ be such that $\|\mathbf{w}\|_2 \le 1$ and let $\{\mathbf{x}^{(i)}\}_{i=1}^N$ be a multiset of $N$ i.i.d. samples from $\mathcal{D}_\mathbf{x}$. Then, it holds $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \ge 2(\mathrm{err}_{\widehat{D}_N}(\mathbf{w}) - \eta)$ , where $\widehat{D}_N$ is the corresponding empirical distribution.*

*Proof.* We partition $\mathbb{R}^d$ into two subsets $R_1, R_2$ as follows: $R_1$ contains the points that lie sufficiently far away from the separating hyperplane $\mathbf{w}\cdot\mathbf{x}=0$, i.e., $R_1 \stackrel{\mathrm{def}}{=} \{\mathbf{x}\in\mathbb{R}^d : |\mathbf{w}\cdot\mathbf{x}| \ge \gamma/2\}$. $R_2$ contains the remaining points, i.e., $R_2 \stackrel{\mathrm{def}}{=} \{\mathbf{x}\in\mathbb{R}^d : |\mathbf{w}\cdot\mathbf{x}| < \gamma/2\}$.

We first show that for any $\mathbf{x}\in R_1$, the vector $\mathbf{g}^1(\mathbf{w}, \mathbf{x})$ has a large component parallel

to the direction $\mathbf{w} - \mathbf{w}^*$.

**Claim 3.2.3.** *For any* $\mathbf{x}^{(i)} \in R_1$*, we have that* $\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq 2(\mathrm{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta)$ .

*Proof of Claim 3.2.3.* For any $\mathbf{x}^{(i)} \in R_1$, we have that

$$\begin{aligned}
&\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot \mathbf{w} \\
&= ((1 - 2\eta)\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}^{(i)}) - (1 - 2\eta(\mathbf{x}^{(i)}))\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x}^{(i)}))\mathbf{w} \cdot \mathbf{x}^{(i)} W(\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= ((1 - 2\eta)\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}^{(i)}) - (1 - 2\eta(\mathbf{x}^{(i)}))\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x}^{(i)}))\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= 2(\mathrm{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta) \ , \qquad (3.4)
\end{aligned}$$

where we used that for any $\mathbf{x}^{(i)} \in R_1$, $W(\mathbf{w} \cdot \mathbf{x}^{(i)}) = 1/|\mathbf{w} \cdot \mathbf{x}^{(i)}|$, and hence $W(\mathbf{w} \cdot \mathbf{x}^{(i)}, \gamma/2)\mathbf{w} \cdot \mathbf{x}^{(i)} = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x}^{(i)})$; and that $\mathrm{err}(\mathbf{w}, \mathbf{x}^{(i)}) = \eta(\mathbf{x}^{(i)})$ if $\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}^{(i)}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$ and $1 - \eta(\mathbf{x}^{(i)})$ otherwise.

We now bound the contribution of $\mathbf{w}^*$. Since $\eta(\mathbf{x}) \leq \eta$, we have

$$(1 - 2\eta(\mathbf{x})) - (1 - 2\eta)\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x}) \geq 0 \ .$$

Therefore, we have that

$$\begin{aligned}
&\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot \mathbf{w}^* \\
&= \Big((1 - 2\eta)\mathrm{sign}(\mathbf{w} \cdot \mathbf{x}) - (1 - 2\eta(\mathbf{x}))\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})\Big)\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})|\mathbf{w}^* \cdot \mathbf{x}| W(\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= -\Big((1 - 2\eta(\mathbf{x})) - (1 - 2\eta)\mathrm{sign}(\mathbf{w} \cdot \mathbf{x})\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})\Big)|\mathbf{w}^* \cdot \mathbf{x}| W(\mathbf{w} \cdot \mathbf{x}^{(i)}) \leq 0 \ ,
\end{aligned}$$

which gives that $-\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot \mathbf{w}^* \geq 0$. This completes the proof of Claim 3.2.3. □

It remains to show that the same holds for all the points in $R_2$.

**Claim 3.2.4.** *For any* $\mathbf{x}^{(i)} \in R_2$*, we have that* $\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq 2(\text{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta)$ .

*Proof of Claim 3.2.4.* We have that

$$\begin{aligned}
&\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) \\
&= \left((1-2\eta)\text{sign}(\mathbf{w} \cdot \mathbf{x}^{(i)}) - (1 - 2\eta(\mathbf{x}^{(i)}))\text{sign}(\mathbf{w}^* \cdot \mathbf{x}^{(i)})\right) \left(\frac{\mathbf{w} \cdot \mathbf{x}^{(i)} - \mathbf{w}^* \cdot \mathbf{x}^{(i)}}{\max(\gamma/2, |\mathbf{w} \cdot \mathbf{x}^{(i)}|)}\right) \\
&= \left((1-2\eta)\text{sign}(\mathbf{w} \cdot \mathbf{x}^{(i)}) - (1 - 2\eta(\mathbf{x}^{(i)}))\text{sign}(\mathbf{w}^* \cdot \mathbf{x}^{(i)})\right) \left(\frac{\mathbf{w} \cdot \mathbf{x}^{(i)} - \mathbf{w}^* \cdot \mathbf{x}^{(i)}}{\gamma/2}\right) ,
\end{aligned}$$

where we used that $\max(\gamma/2, |\mathbf{w} \cdot \mathbf{x}^{(i)}|) = \gamma/2$ for any $\mathbf{x}^{(i)} \in R_2$. Since $\text{sign}(\mathbf{w}^* \cdot \mathbf{x})$ has $\gamma$-margin, we have that $|\mathbf{w}^* \cdot \mathbf{x}^{(i)}| \geq \gamma$. Since $\mathbf{x}^{(i)} \in R_2$, it holds $|\mathbf{w} \cdot \mathbf{x}^{(i)}| < \gamma/2$. Therefore, $-\text{sign}(\mathbf{w}^* \cdot \mathbf{x}^{(i)})(\mathbf{w} \cdot \mathbf{x}^{(i)} - \mathbf{w}^* \cdot \mathbf{x}^{(i)}) = \left(|\mathbf{w}^* \cdot \mathbf{x}^{(i)}| - \text{sign}(\mathbf{w}^* \cdot \mathbf{x}^{(i)})\mathbf{w} \cdot \mathbf{x}^{(i)}\right) \geq \gamma/2$. This in turn implies that

$$\begin{aligned}
\mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) &\geq (1 - 2\eta(\mathbf{x}^{(i)}) - (1-2\eta)\text{sign}(\mathbf{w} \cdot \mathbf{x}^{(i)})\text{sign}(\mathbf{w}^* \cdot \mathbf{x}^{(i)})) \\
&= 2(\text{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta) ,
\end{aligned}$$

completing the proof of Claim 3.2.4. □

Applying Claim 3.2.3 and Claim 3.2.4 for each sample in the set $\{\mathbf{x}^{(i)}\}_{i=1}^N$, we get that

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{g}^1(\mathbf{w}, \mathbf{x}^{(i)}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq \frac{2}{N}\sum_{i=1}^{N} (\text{err}(\mathbf{w}, \mathbf{x}^{(i)}) - \eta) .$$

This completes the proof of Lemma 3.2.2. □

By Lemma 3.2.2, the gradient points towards the direction $\mathbf{w}^t - \mathbf{w}^*$, in the $t$-th iteration. This means that, in fact, the gradient is a subgradient of the potential loss

$\Phi(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^*\|_2^2$. This allows us to show convergence, even though it is generally not possible in a sequence of loss functions in the stochastic setting. We are now ready to prove our main result.

*Proof of Theorem 3.2.1.* Let $T$ be the maximum number of iterations of Algorithm 4. Denote by $\mathcal{Z}^t := \{(\mathbf{x}^{(t)}, y^{(t)})\}$ the i.i.d. sample drawn from $D$ in the $t$-th iteration, $t \in [T]$. Furthermore, let $\mathcal{F}_1, \ldots, \mathcal{F}_T$ be the filtration with respect to the $\sigma$-algebra generated by $\mathcal{Z}^1, \ldots, \mathcal{Z}^T$. We denote by $H_t$ the event that $\mathrm{err}_D(\mathbf{w}^t) \geq \eta + \epsilon$.

Recall that Algorithm 4 uses the following update rule (see Step (7)):

$$\mathbf{w}^{t+1} = \mathrm{proj}_{\{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq 1\}}(\mathbf{w}^t - \lambda_t \mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)})) \ ,$$

with $\lambda_t = c\gamma^2\epsilon$ , for some sufficiently small absolute constant $c > 0$.

We begin by bounding from above the distance between $\mathbf{w}^{t+1}$ and $\mathbf{w}^*$ from the previous distance between $\mathbf{w}^t$ and $\mathbf{w}^*$. We have that

$$\begin{aligned}
&\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 \\
&= \|\mathrm{proj}_{\{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq 1\}}(\mathbf{w}^t - \lambda_t \mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) - \mathbf{w}^*\|_2^2 \\
&\leq \|\mathbf{w}^t - \lambda_t \mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) - \mathbf{w}^*\|_2^2 \\
&= \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - 2\lambda_t \mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) \cdot (\mathbf{w}^t - \mathbf{w}^*) + \lambda_t^2 \|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)})\|_2^2 \ , \quad (3.5)
\end{aligned}$$

where in the first inequality we used the projection inequality, i.e., $\|\mathrm{proj}_B(\mathbf{v}) - \mathrm{proj}_B(\mathbf{u})\|_2 \leq \|\mathbf{v} - \mathbf{u}\|_2$ for any set $B$. We will decouple the mean of the random variable $\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}, y)$ and make it zero-mean.

To simplify the notation, we denote by $\xi_t := \left(\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) - \mathbf{G}^1_{\mathcal{D}}(\mathbf{w}^t)\right) \cdot (\mathbf{w}^t - \mathbf{w}^*)$

and note that $\xi_t$ is a zero-mean random variable over the sample $(\mathbf{x}^{(t)}, y^{(t)})$. Adding and subtracting $\mathbf{G}^1_{\mathcal{D}}(\mathbf{w}^t)$ onto Inequality (3.5) a we get that

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 \underbrace{-2\lambda_t \mathbf{G}^1_D(\mathbf{w}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) + \lambda_t^2 \|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)})\|_2^2}_{I} \underbrace{-2\lambda_t \xi_t}_{\widehat{V}_t} \tag{3.6}$$

We now outline the main steps of our analysis. Instead of accurately estimating the gradients in each round, we denote by $\widehat{V}_t$ the estimation error from which we bound above their sum. We first add and subtract the population gradient to obtain the $I$ term, which is the decreasing direction. In this way, we decouple the expected decrease and the error of the approximation (see Claim 3.2.5). After that, we bound the contribution of the estimation error in Lemma 3.2.8. Observe that $\widehat{V}_t$ is a random variable that corresponds to the estimation error of the gradient. We will argue that with high probability the contribution of $\sum_{t=1}^{T} \widehat{V}_t$ is bounded; therefore, our algorithm will converge to an accurate solution.

Lemma 3.2.2 shows that $\mathbf{G}^1_{\widehat{D}^t_N}(\mathbf{w}^t)$ (and therefore the same holds for $\mathbf{G}^1_D(\mathbf{w}^t)$) contains substantial contribution towards to the direction $\mathbf{w}^t - \mathbf{w}^*$, depending of the current error. We show that our choice of step size guarantees a decreasing direction. To this end, we prove the following:

**Claim 3.2.5.** *Assume that the event $H_t$ happens, i.e., $\mathrm{err}_D(\mathbf{w}^t) \geq \eta + \epsilon$. If $\lambda_t \leq \gamma^2\epsilon/8$, then $I \leq -\lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta)$.*

*Proof of Claim 3.2.5.* Recall that $I = -2\lambda_t \mathbf{G}^1_D(\mathbf{w}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) + \lambda_t^2 \|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)})\|_2^2$. By Lemma 3.2.2, we get that $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) \geq 2(\mathrm{err}_{\widehat{D}_N}(\mathbf{w}^t) - \eta)$; hence, by taking expectations over the samples, we also have $\mathbf{G}^1_{\mathcal{D}}(\mathbf{w}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) \geq 2(\mathrm{err}_D(\mathbf{w}^t) - \eta)$.

Furthermore, we have that $\|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)})\|_2^2 \leq 8/\gamma^2$. Hence,

$$I \leq -2\lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) + 8(\lambda_t^2/\gamma^2) \ .$$

The claim follows by noting that if $\lambda_t \leq \gamma^2\epsilon/8$, then $-\lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) + 8(\lambda_t^2/\gamma^2) \leq 0$. Therefore, we obtain

$$I \leq -\lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) \ .$$

This completes the proof of Claim 3.2.5. □

Therefore, our choice of parameters guarantees that $\lambda_t \leq \gamma^2\epsilon/8$. Using Claim 3.2.5 onto Inequality (3.6), we have that

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - \lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) + \widehat{V}_t \ . \tag{3.7}$$

Using Claim 3.2.5 and Inequality (3.7), we have that

$$\begin{aligned} \|\mathbf{w}^{T+1} - \mathbf{w}^*\|_2^2 &\leq \|\mathbf{w}^T - \mathbf{w}^*\|_2^2 - \lambda_T(\mathrm{err}_D(\mathbf{w}^T) - \eta) + \widehat{V}_T \\ &\leq \|\mathbf{w}^0 - \mathbf{w}^*\|_2^2 - \sum_{t=0}^{T} \lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) + \sum_{t=0}^{T} \widehat{V}_t \ . \end{aligned} \tag{3.8}$$

To complete the proof of Theorem 3.2.1, we need to bound the estimation error that corresponds to the random variable $\widehat{V}_t$. We show that $\widehat{V}_t$ does not increase the error by a lot. Recall that $\widehat{V}_t = -2\lambda_t\xi_t$ .

Before proceeding, we provide some basic background on subgaussian random variables.

**Definition 3.2.6** (Subgaussian Random Variable)**.** *For $\sigma > 0$, a zero-mean random*

*variable* $X \in \mathbb{R}$ *is called* $\sigma$*-subgaussian, if for any* $\lambda \in \mathbb{R}$ *it holds* $\log(\mathbf{E}[\exp(\lambda X)]) \leq \lambda^2\sigma^2$ .

Note that any zero-mean bounded random variable is subgaussian. Specifically, we have the following:

**Fact 3.2.7** (Hoeffding's lemma, see, e.g., [Ver18])**.** *Let* $X \in \mathbb{R}$ *be a zero-mean random variable such that* $|X| \leq \sigma$ *for some* $\sigma > 0$*. Then* $X$ *is* $C\sigma$*-subgaussian, where* $C > 0$ *is a universal constant.*

Equipped with the above context, we show the following:

**Lemma 3.2.8.** *With probability at least* $1 - \delta$ *over the random samples, it holds that* $\sum_{t=0}^{T} \widehat{V}_t \leq C\gamma^2\epsilon^2 T + \log(1/\delta)$*, where* $C > 0$ *is an absolute constant.*

*Proof.* We first show that $\xi_t$ is a subgaussian random variable.

**Claim 3.2.9.** *The random vector* $\xi_t$ *is* $(16/\gamma)$*-subgaussian.*

*Proof of Claim 3.2.9.* Note that $\xi_t = (\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}, y)]) \cdot (\mathbf{w}^t - \mathbf{w}^*)$ and that by construction $\|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}, y)\|_2 \leq 4/\gamma$. Therefore, it holds that

$$|\mathbf{g}(\mathbf{w}^t, \mathbf{w}^t, \mathbf{x}^{(t)}, y^{(t)}) \cdot (\mathbf{w}^t - \mathbf{w}^*)| \leq 8/\gamma \ ,$$

where we used that $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq 2$ as both of these vectors lie in the unit ball. Hence, by Fact 3.2.7, we have that $\xi_t$ is $(16/\gamma)$-subgaussian. □

Using Claim 3.2.9 and Definition 3.2.6 with parameter $\lambda = -2\lambda_t$ and $X = \xi_t$, we have that

$$\log \mathbf{E}[\exp(\widehat{V}_t)] = \log \mathbf{E}[\exp(-2\lambda_t\xi_t)] \leq C(\lambda_t^2/\gamma^2) \ ,$$

where $C > 0$ is a universal constant. To bound the contribution of $\sum_{t=0}^{T} \widehat{V}_t$, we use Markov's inequality with respect to the filtration $\mathcal{F}_1, \ldots, \mathcal{F}_T$. We have that for any $Z \in \mathbb{R}$, it holds that

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\mathcal{Z}^1,\ldots,\mathcal{Z}^T \sim D}\left[\sum_{t=0}^{T} \widehat{V}_t \geq Z\right] &= \mathop{\mathbf{Pr}}_{\mathcal{Z}^1,\ldots,\mathcal{Z}^T \sim D}\left[\exp\left(\sum_{t=0}^{T} \widehat{V}_t\right) \geq \exp(Z)\right] \\
&\leq \mathop{\mathbf{E}}_{\mathcal{Z}^1,\ldots,\mathcal{Z}^T \sim D}\left[\exp\left(\sum_{t=0}^{T} \widehat{V}_t\right)\right] \exp(-Z) \\
&= \prod_{t=1}^{T} \mathop{\mathbf{E}}_{\mathcal{Z}^t \sim D}\left[\exp \widehat{V}_t \mid \mathcal{F}_t\right] \exp(-Z) \leq \exp\left(C \sum_{t=0}^{T} \frac{\lambda_t^2}{\gamma^2} - Z\right) ,
\end{aligned}$$

where in the second inequality we use the independence of $\widehat{V}_t$ with $\{\widehat{V}_k\}_{k=1}^{t-1}$ with respect to the filtration $\mathcal{F}_t$. Recalling that $\lambda_t = c\gamma^2\epsilon$, where $c > 0$ is a sufficiently small universal constant, we have that

$$\mathop{\mathbf{Pr}}_{\mathcal{Z}^1,\ldots,\mathcal{Z}^T \sim D}\left[\sum_{t=0}^{T} \widehat{V}_t \geq Z\right] \leq \exp\left(Cc^2\gamma^2\epsilon^2 T - Z\right) \leq \exp\left(Cc^2\gamma^2\epsilon^2 T - Z\right) .$$

Setting $Z = Cc^2\gamma^2\epsilon^2 T + \log(1/\delta)$ and taking $c$ to be a sufficiently small absolute constant (as is done in our algorithm), we get that $\mathbf{Pr}_{\mathcal{Z}^1,\ldots,\mathcal{Z}^T \sim D}\left[\sum_{t=0}^{T} \widehat{V}_t \geq Z\right] \leq \delta$. This completes the proof of Lemma 3.2.8. □

Assume that until the round $T$ the event $H_T$ holds, i.e., for all $i \in [T]$ we have that $\mathrm{err}_{\mathcal{D}}(\mathbf{w}^i) \geq \eta + \epsilon$. Using Lemma 3.2.8 onto Inequality (3.8), with probability at least $1 - \delta$, we have that:

$$\begin{aligned}
\|\mathbf{w}^{T+1} - \mathbf{w}^*\|_2^2 &\leq \|\mathbf{w}^0 - \mathbf{w}^*\|_2^2 - \sum_{t=0}^{T} \lambda_t(\mathrm{err}_D(\mathbf{w}^t) - \eta) + \sum_{t=0}^{T} \widehat{V}_t \\
&\leq \|\mathbf{w}^0 - \mathbf{w}^*\|_2^2 - cT\epsilon^2\gamma^2 + \log(1/\delta) .
\end{aligned}$$

Running the algorithm for $T = \Theta(\log(1/\delta)/(\epsilon^2\gamma^2))$ iterations guarantees that with probability at least $1-\delta$, we will have that $\|\mathbf{w}^{T+1} - \mathbf{w}^*\|_2^2 \leq 0$, which means $\mathbf{w}^{T+1} = \mathbf{w}^*$. In that case, i.e., in the case where all the events $H_i$ for $i \in [T]$ hold, $\mathbf{w}^{T+1}$ achieves the same error as the optimal halfspace, thus it has 0-1 error of at most $\eta + \epsilon$. Therefore, at least one vector $\mathbf{w}^{t'}$ with $t' \in [T+1]$ achieves 0-1 error of at most $\eta+\epsilon$. The algorithm, in Step (11), returns a vector $\widehat{\mathbf{w}}$ that has 0-1 error at most $\mathrm{err}_D(\widehat{\mathbf{w}}) \leq \min_{t\in[T+1]} \mathrm{err}_D(\mathbf{w}^t) + \epsilon \leq \eta + 2\epsilon$. The algorithm requires $N = O(\log(T/\delta)/(\epsilon(1-2\eta)))$ samples for Step (11), due to [MN06]. The algorithm draws a sample in each round and runs for at most $T$ rounds. Therefore, Algorithm 4 draws $n = N + T = \widetilde{O}(\log(1/\delta)/(\epsilon^2\gamma^2))$ samples. The algorithm needs to test each of the $T$ hypotheses with $N$ samples to find the closest one. Therefore, the total runtime is $O(dTN)$ (as in the other subroutines the algorithm uses the samples only to estimate the gradients $\mathbf{g}$, which requires $O(1)$ additions of $d$-dimenional vectors). This completes the proof of Theorem 3.2.1. □

## 3.3 Conclusions and Open Problems

In this work, we give the first sample near-optimal and computationally efficient algorithm for learning margin halfspaces in the presence of Massart noise. Specifically, the sample complexity of our algorithm nearly matches the computational sample complexity of the problem and its computational complexity is polynomial in the sample size. An interesting direction for future work is to develop a sample near-optimal and computationally efficient learner for general halfspaces (i.e., without the margin assumption). While our approach can likely be leveraged to obtain an efficient algorithm with sample complexity $\mathrm{poly}(d)/\epsilon^2$, the sample dependence on the dimension $d$ would be suboptimal. Obtaining the right dependence on the dimension seems to require novel

ideas, as prior works rely on fairly sophisticated methods [DV04, DKT21, DTK22] to effectively reduce to the large margin case.

# Chapter 4

# Learning Agnostically Proper Halfspaces

## 4.1 Introduction

In this chapter, we develop algorithms for proper learning in the agnostic distribution-specific PAC model. Our main result is the first proper agnostic learner for the class of halfspaces under the Gaussian distribution.

**Theorem 4.1.1** (Proper Agnostic Learning of Halfspaces)**.** *Let $\mathcal{D}$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. There exists an algorithm that, given $\epsilon, \delta > 0$, and $N = d^{O(1/\epsilon^4)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ i.i.d. samples from $\mathcal{D}$, the algorithm runs in time $\mathrm{poly}(N) + (1/\epsilon)^{O(1/\epsilon^6)}\log(1/\delta)$, and computes a halfspace hypothesis $h$ such that, with probability at least $1 - \delta$, it holds $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \leq \mathrm{OPT} + \epsilon$.*

Our second main contribution is the first proper polynomial-time approximation scheme (PTAS) for the agnostic learning problem. In our context, a PTAS is an algorithm

that, for any $\gamma, \epsilon > 0$, runs in time $d^{\mathrm{poly}(1/\gamma)}/\mathrm{poly}(\epsilon)$ and outputs a hypothesis $h$ satisfying $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \leq (1+\gamma)\mathrm{OPT} + \epsilon$. The parameter $\gamma > 0$ quantifies the approximation ratio of the algorithm. Prior work [Dan15] gave an improper PTAS for agnostically learning *homogeneous* halfspaces, i.e., halfspaces whose separating hyperplane goes through the origin. We give a proper algorithm for this problem.

**Theorem 4.1.2** (Proper PTAS for Agnostically Learning Halfspaces)**.** *Let $\mathcal{D}$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. There exists an algorithm that, given $\gamma, \epsilon, \delta > 0$ and $N = d^{\mathrm{poly}(1/\gamma)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ i.i.d. samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and computes a halfspace $h$ such that, with probability $1-\delta$, it holds $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \leq (1+\gamma)\mathrm{OPT} + \epsilon$, where OPT is the optimal misclassification error of any homogeneous halfspace.*

Theorem 4.1.2 gives the first proper PTAS for agnostically learning homogeneous halfspaces under any natural distributional assumptions and qualitatively matches the complexity of the improper PTAS by [Dan15]. We note that the homogeneity assumption is needed for technical reasons and is also required in the known improper learning algorithm. Obtaining a PTAS for agnostically learning arbitrary halfspaces remains an open problem (even for improper learners).

**Proper Agnostic Learning Algorithm Technical Overview** The main idea behind our proper learning algorithm is to start with a good improper hypothesis and compress it down to a halfspace, while maintaining the same error guarantees. Our algorithm starts by computing the low-degree polynomial $P$ that best approximates the labels in $L_2$-norm (Lemma 4.2.3). We then take a two-step approach to identify a near-optimal halfspace. First, by identifying the high-influence directions of $P$, we

construct a low-dimensional subspace of $\mathbb{R}^d$ and show that it contains the normal vector to a near-optimal halfspace (Proposition 4.2.1). Then, we exhaustively search over vectors in this subspace (through an appropriately fine cover) and output the one with minimum error.

The main technical challenge comes in identifying such a subspace that is large enough to contain a good proper hypothesis, but also small enough so that exhaustive searching is efficient. To identify this subspace, we consider an appropriate matrix (defined by the high-influence directions of the polynomial $P$) and take the subspace defined by its large eigenvectors (Proposition 4.2.1). Exploiting the concentration guarantees of polynomials under the Gaussian distribution, we show that the resulting subspace is small enough to enumerate over (Lemma 4.2.2).

In more detail, we first find the polynomial $P(\mathbf{x})$ of degree $k = O(1/\epsilon^4)$ that approximates the labels $y$ in the $L_2$ sense, that is, minimizes $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2]$. We then consider the influence of the polynomial $P$ along a direction $\mathbf{u}$, $\mathrm{Inf}_{\mathbf{u}}(P) = \mathbf{u}^T\mathbf{M}\mathbf{u}$, for the matrix $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\intercal]$, as a measure for how much the polynomial $P$ changes along the direction $\mathbf{u}$. The key observation is that, along low-influence directions, the polynomial remains essentially constant and, as we show, the optimal halfspace must also be essentially constant as well. This allows us to prune down these directions and focus on a subspace of lower-dimension. Our main structural result (Proposition 4.2.1) formalizes this intuition showing that the subspace $V$ of eigenvectors whose eigenvalues are larger than $\Theta(\epsilon^2)$ contains a normal vector $\mathbf{w}_V$ that (together with an appropriate threshold) achieves error $\mathrm{OPT} + \epsilon$.

Finally, while Proposition 4.2.1 establishes that we can remove directions of low-influence, we need to argue that the number of relevant eigenvectors is sufficiently small

to simplify the problem. As we show in Lemma 4.2.2, the dimension of the resulting subspace $V$ is $O(1/\epsilon^6)$, and thus finding a good hypothesis in this subspace takes time independent of the original dimension $d$. The key ingredient in bounding the dimension of $V$ is to use concentration of polynomials under the Gaussian distribution to argue that the Frobenius norm of $\mathbf{M}$ is bounded, and thus the number of eigenvectors with large eigenvalues is bounded.

**Proper PTAS for Agnostic Learning** Our algorithm for obtaining a proper PTAS works in the same framework as [Dan15], who gave a non-proper PTAS for homogeneous halfspaces by combining the algorithm of [ABL17] with the $L_1$-polynomial regression algorithm of [KKMS08].

Similarly to the algorithm of [Dan15], we start by learning a halfspace (with normal vector) $\mathbf{w}_0$ with error $O(\mathrm{OPT})$, using any of the known constant factor approximations as a black-box [ABL17, DKS18, DKTZ20c], and then partition the space according to the distance to the halfspace $\mathbf{w}_0$. Daniely's algorithm [Dan15] is based on the observation that points far from the true halfspace are accurately classified by the halfspace $\mathbf{w}_0$. Thus, one can use the improper learner of [KKMS08] to classify nearby points.

A simple adaptation of this idea would be to replace the improper algorithm of [KKMS08] with our new proper algorithm for agnostically learning halfspaces. There are two main complications however. First, the guarantees of our proper algorithm crucially rely on having Gaussian marginals, and therefore we cannot readily apply it once we restrict our attention only to points around $\mathbf{w}_0$. We deal with this issue by using a "soft" localization technique introduced in [DKS18] to randomly partition points in two groups. In particular, we perform rejection sampling according to a judiciously chosen weight function such that the distribution conditional on acceptance is still a

Gaussian, albeit with very small variance along the direction of $\mathbf{w}_0$, see Lemma 4.3.2. By running our proper algorithm, we can obtain a halfspace $\mathbf{w}_1$ that is near-optimal under the conditional distribution.

The second obstacle is that while we can obtain two halfspaces ($\mathbf{w}_0$ and $\mathbf{w}_1$) that each are near-optimal for their corresponding groups, combining them into a *single* halfspace that works well for the entire distribution is not immediate. We remark that this is not an issue for the improper approximation scheme of [Dan15], since an improper learner is allowed to output a different classifier for different subsets of $\mathbb{R}^d$. To handle this issue, we additionally show that the halfspace $\mathbf{w}_1$ we obtain after localization will in fact perform well overall. In more detail, we show that the halfspace $\mathbf{w}_1$ cannot have very large angle with $\mathbf{w}_0$ and also its bias is small, see Proposition 4.3.3. Given these closeness properties, we can then show that the halfspace $\mathbf{w}_1$ achieves the desired error guarantees over the entire distribution, see Lemma 4.3.4.

**Notation** We use $\mathcal{C}_V$ for the set of Linear Threshold Functions (LTFs) with normal vector contained in $V \subseteq \mathbb{R}^d$, i.e., $\mathcal{C}_V = \{\mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t) : \mathbf{v} \in V, \|\mathbf{v}\|_2 = 1, t \in \mathbb{R}\}$; when $V = \mathbb{R}^d$, we simply write $\mathcal{C}$. Moreover, we define $\mathcal{C}_0$ to be the set of unbiased LTFs, i.e., $\mathcal{C}_0 = \{\mathrm{sign}(\mathbf{v} \cdot \mathbf{x}) : \mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\|_2 = 1\}$. We denote by $\mathcal{P}_k$ the space of polynomials on $\mathbb{R}^d$ of degree at most $k$.

## 4.2 Proper Agnostic Learning Algorithm

In this section, we present our proper agnostic learning algorithm for halfspaces, establishing Theorem 4.1.1. The pseudocode of our algorithm is given in Algorithm 5.

**Algorithm 5** Agnostic Proper Learning Halfspaces

**procedure** AGNOSTIC-PROPER-LEARNER($\epsilon, \delta, \mathcal{D}$)
**Input:** $\epsilon > 0$, $\delta > 0$ and sample access to distribution $\mathcal{D}$
**Output:** A hypothesis $h \in \mathcal{C}$ such as $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \leq \min_{f \in \mathcal{C}} \mathrm{err}^{\mathcal{D}}_{0-1}(f) + \epsilon$ with probability $1 - \delta$.
    $k \leftarrow C/\epsilon^4$, $\eta \leftarrow \epsilon^2/C$. ▷ $C$ is a sufficiently large constant
    Find $P(\mathbf{x})$ such $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + O(\epsilon^3)$.
    Let $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\nabla P(\mathbf{x}) \nabla P(\mathbf{x})^\intercal]$.
    Let $V$ be the subspace spanned by the eigenvectors of $\mathbf{M}$ whose eigenvalues are at least $\eta$.
    Construct an $\epsilon$-cover $\mathcal{H}$ of LTF hypotheses with normal vectors in $V$ ▷ see Fact 4.2.4.
    Draw $\Theta(\frac{1}{\epsilon^2}\log(|\mathcal{H}|/\delta))$ i.i.d. samples from $\mathcal{D}$ and construct the empirical distribution $\widehat{\mathcal{D}}$.
    $h \leftarrow \mathrm{argmin}_{h' \in \mathcal{H}} \mathrm{err}^{\widehat{\mathcal{D}}}_{0-1}(h')$
    **return** $h$.
**end procedure**

### 4.2.1 Analysis of Algorithm 5: Proof of Theorem 4.1.1

The main structural result that allows us to prove Theorem 4.1.1 is the following proposition, establishing the following: Given a multivariate polynomial $P$ of degree $\Theta(1/\epsilon^4)$ that correlates well with the labels, we can use its *high-influence directions* to construct a subspace that contains a near-optimal halfspace. Specifically, we show:

**Proposition 4.2.1.** *Let $C > 0$ be a sufficiently large universal constant. Fix any*

*$\epsilon \in (0,1]$ and set $k = C/\epsilon^4$. Let $P(\mathbf{x}) \in \mathcal{P}_k$ be a degree-$k$ polynomial such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + O(\epsilon^3)$. Moreover, let $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x}) \nabla P(\mathbf{x})^\intercal]$ and $V$ be the subspace spanned by the eigenvectors of $\mathbf{M}$ with eigenvalues larger than $\eta$, where $\eta = \epsilon^2/C$. Then, for any $f \in \mathcal{C}$, it holds $\min_{\mathbf{v}\in V, t\in\mathbb{R}} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - \mathrm{sign}(\mathbf{v}\cdot\mathbf{x} + t))y] \leq \epsilon$.*

The proof of Proposition 4.2.1 is the bulk of the technical work of this section and is deferred to Section (4.2.2). In the body of this subsection, we show how to use Proposition 4.2.1 to establish Theorem 4.1.1.

The next lemma bounds from above the dimension of the subspace spanned by the high-influence directions of a degree-$k$ polynomial that minimizes the $L_2$-error with the labels $y$.

**Lemma 4.2.2.** *Fix $\epsilon > 0$ and let $P(\mathbf{x})$ be a degree-$k$ polynomial, with $k = O(1/\epsilon^4)$, such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + O(\epsilon^3)$. Let $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x}) \nabla P(\mathbf{x})^\intercal]$ and $V$ be the subspace spanned by the eigenvectors of $\mathbf{M}$ with eigenvalues larger than $\eta$. Then the dimension of the subspace $V$ is $\dim(V) = O(k/\eta)$.*

*Proof.* Let $P$ be a polynomial such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + O(\epsilon^3)$ and let $P^* = \mathrm{argmin}_{P' \in \mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2]$. First, we note that $\mathbf{E}_{(x,y)\sim\mathcal{D}}[(y - P^*(\mathbf{x}))^2] \leq \mathbf{E}_{(x,y)\sim\mathcal{D}}[(y-0)^2] = 1$. Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[P(\mathbf{x})^2] &\leq 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y^2] + 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(x))^2] \\ &\leq 2 + 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P^*(x))^2] + O(\epsilon^3) \leq 5\,.\end{aligned}$$

Let $V$ denote the subspace spanned by the eigenvectors of $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x}) \nabla P(\mathbf{x})^\intercal]$

with eigenvalues at least $\eta$. We will show that $m = \dim(V) = O(k/\eta)$. We can write

$$m\,\eta \leq \mathrm{tr}\left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^{\intercal}]\right) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathrm{tr}(\nabla P(\mathbf{x})\nabla P(\mathbf{x})^{\intercal})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\|\nabla P(\mathbf{x})\|_2^2\right]\,. \tag{4.1}$$

It is sufficient to show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\|\nabla P(\mathbf{x})\|_2^2] = O(k)$. By writing $P(\mathbf{x})$ in the Hermite basis, from Fact 1.4.2, it holds

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\|\nabla P(\mathbf{x})\|_2^2\right] = \sum_{\alpha\in\mathbb{N}^d} |\alpha| c_\alpha^2 \leq k \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[P(\mathbf{x})^2] \leq 5k\,. \tag{4.2}$$

Combining Equations (4.1) and (4.2), we obtain that $m = O(k/\eta)$, and the proof is complete. □

For the proof of Theorem 4.1.1, we require a standard result on $L_2$-polynomial regression required to compute the polynomial of Proposition 4.2.1.

**Lemma 4.2.3** ($L_2$-Polynomial Regression)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Let $k \in \mathbb{Z}_+$ and $\epsilon, \delta > 0$. There is an algorithm that draws $N = (dk)^{O(k)} \log(1/\delta)/\epsilon^2$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and outputs a polynomial $P(\mathbf{x})$ of degree at most $k$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[(f(\mathbf{x}) - P(\mathbf{x}))^2] \leq \min_{P'\in\mathcal{P}_k} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[(f(\mathbf{x}) - P'(\mathbf{x}))^2] + \epsilon$, with probability $1 - \delta$.*

*Proof.* Let $S$ denote the empirical distribution of $\mathcal{D}$ with $N = (d/\epsilon)^{O(k)}$ samples. Recall that for any such $P(\mathbf{x})$, it holds that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[P^2(\mathbf{x})] \leq 5$ (see Lemma 4.2.2). Writing $P(\mathbf{x})$ in the Hermite basis, $P(\mathbf{x}) = \sum_{\alpha\in\mathbb{N}^d} c_\alpha \mathrm{He}_\alpha(\mathbf{x})$, it holds that $\sum_{\alpha\in\mathbb{N}^d} c_\alpha^2 = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[P^2(\mathbf{x})]$. The one-dimensional Hermite polynomials of $k$-degree are $\mathrm{He}_k(z) = \sum_{m=0}^{\lfloor k/2\rfloor} \frac{(-1)^m z^{k-2m}}{m!(n-2m)!2^m}$. Thus, each monomial has coefficient absolute bounded by $2^k$. Therefore, the maximum

coefficient of a multidimensional Hermite polynomial $\mathrm{He}_a(\mathbf{x})$ is $2^{|a|}$, thus the maximum coefficient of the polynomial $P(\mathbf{x})$ is $O(2^k)$. Let us now prove that for any degree-$k$ polynomial $P(\mathbf{x})$ with coefficients bounded by $C = 2^{O(k)}$, we have

$$\left| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[P(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[P(\mathbf{x})y] \right| \leq \epsilon \ ,$$

with high constant probability. Write $P(\mathbf{x}) = \sum a_i m_i(\mathbf{x})$, where the summation ranges over all monomials $m_i$ with degree less than $k$ along with their coefficients $a_i$. We have

$$\left| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[P(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[P(\mathbf{x})y] \right| \leq \sum |a_i| \left| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[m_i(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[m_i(\mathbf{x})y] \right| \ . \quad (4.3)$$

Using Markov's inequality, we have

$$\begin{aligned} \mathbf{Pr}\left[ \left| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[m_i(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[m_i(\mathbf{x})y] \right| \geq \epsilon/(d^k C) \right] &\leq \frac{C^2 d^{2k}}{N\epsilon^2} \mathbf{Var}[m_i(\mathbf{x})y] \\ &\leq \frac{C^2 d^{2k}}{N\epsilon^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[m_i^2(\mathbf{x})y^2] \\ &\leq \frac{C^2 d^{2k}}{N\epsilon^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[\|\mathbf{x}\|_2^{2i}] = O\left( \frac{C^2 i^i d^{2k}}{N\epsilon^2} \right) \ . \end{aligned}$$

By using the fact that $N = (d\ k)^{O(k)}/\epsilon^2$ and applying above to the Equation (4.3), we have

$$\left| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[P(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[P(\mathbf{x})y] \right| \leq C \sum \left| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[m_i(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[m_i(\mathbf{x})y] \right| \leq \epsilon \ ,$$

with high probability. Next, we need to bound the difference $\left| \mathbf{E}_{(\mathbf{x},y)\sim S}[P^2(\mathbf{x})] - \mathbf{E}_{(\mathbf{x},y)\sim \mathcal{D}}[P^2(\mathbf{x})] \right|$. This can be done by applying the same procedure as before and noting that the highest coefficient is at most $C^2$ and the degree is $2k$. Thus, for any $k$-degree polynomial $P$

with high probability, we have

$$\left| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[(P(\mathbf{x})-y)^2] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[(P(\mathbf{x})-y)^2] \right| \leq \epsilon \,, \tag{4.4}$$

where we used the fact that $\mathbf{E}_{(\mathbf{x},y)\sim S}[y^2] = \mathbf{E}_{(\mathbf{x},y)\sim \mathcal{D}}[y^2]$. By solving a convex program, we can find a polynomial $P$ such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[(y-P(\mathbf{x}))^2] \leq \min_{P'\in\mathcal{P}_k} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[(y-P'(\mathbf{x}))^2] + \epsilon \,,$$

Note that if $P''(\mathbf{x}) = \mathrm{argmin}_{P'\in\mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim \mathcal{D}}[(y-P'(\mathbf{x}))^2]$, then

$$\min_{P'\in\mathcal{P}_k} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[(y-P'(\mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S}[(y-P''(\mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[(y-P''(\mathbf{x}))^2] \,,$$

where we used Equation (4.4). Thus, we have proved that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[(y-P(\mathbf{x}))^2] \leq \min_{P'\in\mathcal{P}_k} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim \mathcal{D}}[(y-P'(\mathbf{x}))^2] + \epsilon \,,$$

with high constant probability. Using basic boosting procedures (see, e.g., exercise 1, chapter 13 of [SSBD14]), we can boost the probability of success to $1-\delta$, with $N' = N\log(1/\delta) = (dk)^{O(k)}\log(1/\delta)/\epsilon^2$. □

By running the $L_2$-regression algorithm of the above lemma, we obtain a polynomial $P$ matching the requirements of our dimension-reduction result (Proposition 4.2.1). To complete the proof of Theorem 4.1.1, we perform SVD on the influence matrix $\mathbf{M}$ (see Proposition 4.2.1), and then create a sufficiently fine cover of the low-dimensional subspace $V$.

We require the following standard fact showing the existence of a small $\epsilon$-cover $\widetilde{V}$ of the set $V$, i.e., a set $\widetilde{V}$ such that for any $\mathbf{v} \in V$ there exists $\tilde{\mathbf{v}} \in \widetilde{V}$ such that $\|\mathbf{v} - \tilde{\mathbf{v}}\|_2 \leq \epsilon$.

**Fact 4.2.4** (see, e.g., Corollary 4.2.13 of [Ver18])**.** *For any $\epsilon > 0$, there exists an explicit $\epsilon$-cover of the unit ball in $\mathbb{R}^k$, with respect to the $\ell_2$-norm, of size $O(1/\epsilon)^k$.*

In order to create an effective discretization of the hypotheses, we need the following fact.

**Fact 4.2.5.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ be unit vectors and $t_1, t_2 \in \mathbb{R}$. Then the following holds:*

1. $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t_1) - \mathrm{sign}(\mathbf{v}\cdot\mathbf{x} + t_1)|] = O(\|\mathbf{u} - \mathbf{v}\|_2)$,
2. $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t_1) - \mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t_2)|] = O(|t_1 - t_2|)$ *and,*
3. *if $|t| > \log(1/\epsilon)$ then, for any unit vector $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t) - \mathrm{sign}(t)|] = O(\epsilon)$.*

*Proof.* The first statement is proved in [DKS18] (Lemma 4.2). For the second statement, assuming without loss of generality that $t_2 \geq t_1 > 0$, we note that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t_1) - \mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t_2)|] = \frac{2}{\sqrt{2\pi}}\int_{t_1}^{t_2} e^{-t^2/2}\mathrm{d}t = \mathop{\mathbf{Pr}}_{t\sim\mathcal{N}(0,1)}[t_1 \leq t \leq t_2] \,.$$

Using the anti-concentration property of the one-dimensional Gaussian distribution, we have that $\mathbf{Pr}_{t\sim\mathcal{N}(0,1)}[t_1 \leq t \leq t_2] \leq O(t_2 - t_1)$, proving the claim.

For the third statement, note that if $t > \Omega(\sqrt{\log(1/\epsilon)})$, then by the concentration

properties of the Gaussian, we have that:

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t)-\mathrm{sign}(t)|] \le \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|\mathbf{v}\cdot\mathbf{x}|\ge|t|] = O(\epsilon)\ .$$

This completes the proof of Fact 4.2.5. □

We are now ready to prove the main theorem of this section.

*Proof of Theorem 4.1.1.* We first show that there is a set $\mathcal{H}$ of size $(1/\epsilon)^{O(1/\epsilon^6)}$ which contains tuples $(\mathbf{u},t)$ with $\mathbf{u}\in\mathbb{R}^d$ and $t\in\mathbb{R}$, such that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{u}\cdot\mathbf{x}+t)\neq y] \le \inf_{f\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})\neq y]+\epsilon\ .$$

First note we can assume that $1/\epsilon^6\le d$, since otherwise one can directly do a brute-force search over an $\epsilon$-cover of the $d$-dimensional unit ball: we do not need to perform our dimension-reduction process. The runtime to perform this brute-force search will be $(1/\epsilon)^{O(d)}\log(1/\delta)$ which, by the assumption that $1/\epsilon^6 > d$, is smaller than $(1/\epsilon)^{O(1/\epsilon^6)}\log(1/\delta)$.

Let $f\in\mathcal{C}$ be such that the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})y]$ is maximized and let $k=O(1/\epsilon^4)$. By an application of Lemma 4.2.3 for $N=(d/\epsilon)^{O(1/\epsilon^4)}\mathrm{poly}(1/\epsilon)\log(1/\delta)=d^{O(1/\epsilon^4)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$, it follows that there exists a degree $O(1/\epsilon^4)$ polynomial $P(\mathbf{x})$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \le \min_{P'\in\mathcal{P}_k} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2]+O(\epsilon^3)\ ,$$

with probability $1-\delta/2$. Applying Proposition 4.2.1 to the polynomial $P(\mathbf{x})$, we get that the subspace $V$ spanned by the eigenvectors of the matrix $\mathbf{M}=\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^{\intercal}]$

with eigenvalues larger than $\eta = \Theta(1/\epsilon^2)$ contains a vector $\mathbf{v} \in V$, such that

$$\min_{t\in\mathbb{R}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - \mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t))y] \leq \epsilon \ . \tag{4.5}$$

Moreover, by Lemma 4.2.2, the dimension of $V$ is $O(1/\epsilon^6)$. Applying Fact 4.2.4, we get that there exists an $\epsilon$-cover $\widetilde{V}$ of the set $V$ with respect the $\ell_2$-norm of size $(1/\epsilon)^{O(1/\epsilon^6)}$. We show that there is an effective way to discretize the set of biases. From Fact 4.2.5, it is clear that the set $\mathcal{T} = \{\pm\epsilon, \pm 2\epsilon, \ldots, \pm O(\sqrt{\log(1/\epsilon)})\}$ is an effective cover of the parameter $t$.

It remains to show that the set $\mathcal{H}$ is an effective cover, where $\mathcal{H} = \widetilde{V} \times \mathcal{T}$. We show that there exists a set of parameters $(\tilde{\mathbf{v}}, \tilde{t}) \in \mathcal{H}$ which define a halfspace that correlates with the labels as well as the function $f$. Fix the parameters $(\mathbf{v}, t)$ which minimize the Equation (4.5). Indeed, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - \mathrm{sign}(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))y] \\
&\qquad = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - \mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t))y] + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t) - \mathrm{sign}(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))y] \ .
\end{aligned} \tag{4.6}$$

We claim that there exists $(\tilde{\mathbf{v}}, \tilde{t}) \in \mathcal{H}$ such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t) - \mathrm{sign}(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))y] = O(\epsilon) \ .$$

If $|t| > \sqrt{\log(1/\epsilon)}$, then from Fact 4.2.5, the constant hypothesis gets $O(\epsilon)$ error, so we

need to check the case $|t| \leq \sqrt{\log(1/\epsilon)}$. Applying Fact 4.2.5, we get that

$$\min_{(\tilde{\mathbf{v}},\tilde{t})\in\mathcal{H}} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t)-\mathrm{sign}(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t})|] = \min_{(\tilde{\mathbf{v}},\tilde{t})\in\mathcal{H}} O(\|\mathbf{v}-\tilde{\mathbf{v}}\|_2+|t-\tilde{t}|) = O(\epsilon)\ . \quad (4.7)$$

Thus, substituting Equation (4.7) to Equation (4.6), we get

$$\min_{(\tilde{\mathbf{v}},\tilde{t})\in\mathcal{H}} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-\mathrm{sign}(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))y] \leq \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-\mathrm{sign}(\mathbf{v}\cdot\mathbf{x}+t))y] + O(\epsilon) = O(\epsilon)\ , \quad (4.8)$$

where in the last equality we used Equation (4.5). Using the fact that for a boolean function $g(\mathbf{x})$ it holds $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[g(\mathbf{x})y] = 1 - 2\,\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[g(\mathbf{x}) \neq y]$, we get

$$\min_{(\tilde{\mathbf{v}},\tilde{t})\in\mathcal{H}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\tilde{\mathbf{u}}\cdot\mathbf{x}+\tilde{t}) \neq y] \leq \inf_{f\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y] + O(\epsilon)\ .$$

To complete the proof, we show that Step 10 of Algorithm 5 outputs a hypothesis close to the minimizer inside $\mathcal{H}$. From Hoeffding's inequality, it follows that $O(\frac{1}{\epsilon^2}\log(\mathcal{H}/\delta))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most $\epsilon$ with probability at least $1-\delta/2$. To bound the runtime of the algorithm, we note that $L_2$-regression has runtime $d^{O(1/\epsilon^4)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ and exhaustive search over an $\epsilon$-cover takes time $(1/\epsilon)^{O(1/\epsilon^6)}\log(1/\delta)$. Thus, the total runtime of our algorithm in the case where $1/\epsilon^6 \leq d$ is

$$\left(d^{O(1/\epsilon^4)} + (1/\epsilon)^{O(1/\epsilon^6)}\right)\log(1/\delta)\,.$$

This completes the proof of Theorem 4.1.1. □

### 4.2.2 Proof of Proposition 4.2.1

Suppose for the sake of contradiction that there exists a halfspace $f \in \mathcal{C}$ such that for every halfspace $f' \in \mathcal{C}_V$, it holds

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f'(\mathbf{x}))y] \geq \epsilon \,. \tag{4.9}$$

Our plan is to use the above fact in order to contradict the (approximate) optimality of the polynomial $P(\mathbf{x})$. To achieve this, we need to construct a polynomial $P''(\mathbf{x})$ with error *strictly less than* $\min_{P'\in\mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2]$. In the following simple claim, we show that in order to construct such a polynomial $P''(\mathbf{x})$, one needs to find a polynomial $Q(\mathbf{x})$ of degree at most $k$ that correlates well with the difference $y - P(\mathbf{x})$.

**Claim 4.2.6.** *It suffices to show that there exists a polynomial $Q(\mathbf{x})$ of degree at most $k$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q^2(\mathbf{x})] \leq 9$ that $(\epsilon/4)$-correlates with $(y - P(\mathbf{x}))$, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[Q(\mathbf{x})(y - P(\mathbf{x}))] \geq \epsilon/4$.*

*Proof.* Given such a polynomial $Q(\mathbf{x})$, we consider the polynomial $P''(\mathbf{x}) = P(\mathbf{x}) + \zeta Q(\mathbf{x})$, for $\zeta = c\,\epsilon$ and $c$ a sufficiently small constant. Observe that $P''(\mathbf{x})$ has degree at most $k$ and decreases the value of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2]$ by at least $\Omega(\epsilon^2)$, which contradicts the optimality of $P(\mathbf{x})$, i.e., that $P(\mathbf{x})$ is $O(\epsilon^3)$-close to the polynomial that minimizes the $L_2$-error with $y$. □

We now construct such a polynomial $Q(\mathbf{x})$. We can write that $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} + t) = \mathrm{sign}(\mathbf{w}_V \cdot \mathbf{x} + \mathbf{w}_{V^\perp} \cdot \mathbf{x} + t)$. Note that $\mathbf{w}_{V^\perp} \neq \mathbf{0}$, since otherwise we would have $f \in \mathcal{C}_V$. For simplicity, we denote $\xi = \mathbf{w}_{V^\perp}/\|\mathbf{w}_{V^\perp}\|_2$. Notice that the direction $\xi$ has low influence, since $\xi \in V^\perp$. Recall that by $\mathcal{D}_\xi$ we denote the projection of $\mathcal{D}$ onto the (one-dimensional) subspace spanned by $\xi$. We define $f_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z} + \mathbf{x}_V)]$ to be a

convex combination of halfspaces in $\mathcal{C}_V$. In particular, $f_V(\mathbf{x})$ is a smoothed version of the halfspace $\mathrm{sign}(\mathbf{w}_V \cdot \mathbf{x} + t)$ whose normal vector belongs in $V$. Our argument consists of two main claims. In Lemma 4.2.7, we show that the function $f(\mathbf{x}) - f_V(\mathbf{x})$ correlates non-trivially with $y - P(\mathbf{x})$. Then we show that we can approximate $f(\mathbf{x}) - f_V(\mathbf{x})$ with a low-degree polynomial $Q(\mathbf{x})$ that maintains non-trivial correlation with $y - P(\mathbf{x})$; see Lemma 4.2.9. We start with the first lemma.

**Lemma 4.2.7.** *It holds* $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(y - P(\mathbf{x}))] \geq \epsilon - 2\sqrt{\eta}$.

*Proof.* We have that $f_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z} + \mathbf{x}_{\xi^\perp})] = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[\mathrm{sign}(\mathbf{w}_V \cdot \mathbf{x}_V + \mathbf{w} \cdot \mathbf{z} + t)]$ and, since $f_V$ is a convex combination of halfspaces in $\mathcal{C}_V$, from Equation (4.9), we see that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \geq \epsilon$. Thus, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(y - P(\mathbf{x}))] &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))P(\mathbf{x})] \\ &\geq \epsilon - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))P(\mathbf{x})]\,. \end{aligned} \tag{4.10}$$

To deal with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))P(\mathbf{x})]$, we first observe that for any function $g(\mathbf{x})$ depending only on the projection of $\mathbf{x}$ onto the subspace $\xi^\perp$, i.e., such that $g(\mathbf{x}) = g(\mathbf{x}_{\xi^\perp})$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))g(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}_{\xi^\perp}}\left[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{v}+\mathbf{z}) - f_V(\mathbf{v})]\; g(\mathbf{v})\right] = 0\,,$$

since for every $\mathbf{x} \in \mathbb{R}^d$ it holds $f_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{x}_{\xi^\perp} + \mathbf{z})] = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{x}_V + \mathbf{z})]$. Unfortunately, we cannot directly do the above trick because $P(\mathbf{x})$ does not depend only on $\mathbf{x}_{\xi^\perp}$. However, since $V$ contains the high-influence eigenvectors, it holds that $P$ is almost a function of $\mathbf{x}_{\xi^\perp}$. In fact, we show that we can replace the polynomial

$P$ by a different polynomial of degree at most $k$ that only depends on the projection of $\mathbf{x}$ on $\xi^{\perp}$. Similarly to the definition of the "smoothed" halfspace $f_V$, we define $R(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[P(\mathbf{x}_{\xi^\perp} + \mathbf{z})]$. We first prove that $R(\mathbf{x})$ is close to $P(\mathbf{x})$ in the $L_2$-sense.

**Claim 4.2.8.** *Let* $R(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[P(\mathbf{x}_{\xi^\perp} + \mathbf{z})]$. *It holds* $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(P(\mathbf{x}) - R(\mathbf{x}))^2] \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\nabla P(\mathbf{x}) \cdot \xi)^2]$.

*Proof.* We start by showing that without loss of generality we may assume that $\xi = \mathbf{e}_1$. Let $\mathbf{U}$ be an orthogonal matrix such that $\mathbf{U}\xi = \mathbf{e}_1$. Since $P(\mathbf{x})$ is a polynomial, we can apply the orthogonal transformation $\mathbf{U}$ to $\mathbf{x}$ and then use the Hermite basis to represent it, that is $P(\mathbf{x}) = \sum_{\alpha\in\mathbb{N}^d} c_\alpha \mathrm{He}_\alpha(\mathbf{U}^\intercal\mathbf{x})$. Our objective is equivalent to

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\nabla P(\mathbf{x}) \cdot \xi)^2 - (P(\mathbf{x}) - R(\mathbf{x}))^2] \geq 0 \,.$$

By the change of variables $\mathbf{x} \mapsto \mathbf{U}\mathbf{x}$, we have

$$\mathop{\mathbf{E}}_{(\mathbf{U}\mathbf{x})\sim\mathcal{D}_\mathbf{x}}\left[(\nabla_{\mathbf{U}\mathbf{x}} P(\mathbf{U}\mathbf{x}) \cdot (\mathbf{U}\xi))^2 - (P(\mathbf{U}\mathbf{x}) - R(\mathbf{U}\mathbf{x}))^2\right] \geq 0 \,,$$

where we used the chain rule for the gradient. Observe that

$$R(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[P((\mathbf{I} - \xi\xi^\intercal)\mathbf{x} + \mathbf{z})] = \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}\left[\sum_{\alpha\in\mathbb{N}^d} c_\alpha \mathrm{He}_\alpha(\mathbf{U}^\intercal(\mathbf{I} - \xi\xi^\intercal)\mathbf{x} + \mathbf{U}^\intercal\mathbf{z})\right] \,,$$

thus $R(\mathbf{U}\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[\sum_{\alpha\in\mathbb{N}^d} c_\alpha \mathrm{He}_\alpha(\mathbf{U}^\intercal(\mathbf{I} - \xi\xi^\intercal)\mathbf{U}\mathbf{x} + \mathbf{U}^\intercal\mathbf{z})]$. Moreover, $P(\mathbf{U}\mathbf{x}) = \sum_{\alpha\in\mathbb{N}^d} c_\alpha \mathrm{He}_\alpha(\mathbf{U}^\intercal\mathbf{U}\mathbf{x})$. Using the fact that $\mathbf{U}^\intercal\mathbf{U} = \mathbf{I}$ and $\mathbf{U}^\intercal\xi\xi^\intercal\mathbf{U} = \mathbf{e}_1\mathbf{e}_1^\intercal$, it follows that without loss of generality, we may assume that $\xi = \mathbf{e}_1$.

To keep notation simple, we write $P(\mathbf{x}) = \sum_{\alpha \in \mathbb{N}^d} c_\alpha \mathrm{He}_\alpha(\mathbf{x})$. Note that

$$P(\mathbf{x}) - \mathop{\mathbf{E}}_{\mathbf{x}_1 \sim \mathcal{D}_{\mathbf{e}_1}} [P(\mathbf{x}_{\xi^\perp} + \mathbf{x}_1)] = \sum_{\alpha \in \mathbb{N}^d} c_\alpha \mathrm{He}_\alpha(\mathbf{x}) - \sum_{\alpha \in \mathbb{N}^d} c_\alpha \mathop{\mathbf{E}}_{\mathbf{x}_1 \sim \mathcal{D}_{\mathbf{x}_1}} [\mathrm{He}_\alpha(\mathbf{x})] = \sum_{\alpha \in \mathcal{S}} c_\alpha \mathrm{He}_\alpha(\mathbf{x}) \,, \tag{4.11}$$

where $\mathcal{S}$ contains all the tuples for which the first index is non-zero, this follows from the fact that $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}}[\mathrm{He}_\alpha(\mathbf{x})] = 0$. Applying Parseval's identity yields

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(P(\mathbf{x}) - \mathop{\mathbf{E}}_{\mathbf{x}_1 \sim \mathcal{D}_{\mathbf{e}_1}} [P(\mathbf{x}_{\xi^\perp} + \mathbf{x}_1)])^2] = \sum_{\alpha \in \mathcal{S}} c_\alpha^2 \,. \tag{4.12}$$

From Fact 1.4.2, we have

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(\nabla P(\mathbf{x}) \cdot \mathbf{e}_1)^2] = \sum_{\alpha \in \mathbb{N}^d} \alpha_1 c_\alpha^2 \geq \sum_{\alpha \in \mathcal{S}} c_\alpha^2 \,, \tag{4.13}$$

where we used that $\alpha_1 \geq 1$ on the set $\mathcal{S}$. Combining (4.12) and (4.13) completes the proof. □

Adding and subtracting $R(\mathbf{x}) = \mathbf{E}_{\mathbf{z} \sim \mathcal{D}_\xi}[P(\mathbf{x}_{\xi^\perp} + \mathbf{z})]$, we get

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(f(\mathbf{x}) - f_V(\mathbf{x}))P(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(f(\mathbf{x}) - f_V(\mathbf{x}))(P(\mathbf{x}) - R(\mathbf{x}_{\xi^\perp}))] + \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(f(\mathbf{x}) - f_V(\mathbf{x}))R(\mathbf{x}_{\xi^\perp})] \,.$$

The second term is equal to zero, from the fact that $\mathbf{E}_{\mathbf{z} \sim \mathcal{D}_\xi}[f(\mathbf{z} + \mathbf{x}_{\xi^\perp}) - f_V(\mathbf{x}_{\xi^\perp})] = 0$. From the Cauchy-Schwartz inequality, we get

$$\begin{aligned} \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(f(\mathbf{x}) - f_V(\mathbf{x}))(P(\mathbf{x}) - R(\mathbf{x}_{\xi^\perp}))] &\leq \sqrt{\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(f(\mathbf{x}) - f_V(\mathbf{x}))^2] \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(P(\mathbf{x}) - R(\mathbf{x}_{\xi^\perp}))^2]} \\ &\leq 2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(P(\mathbf{x}) - R(\mathbf{x}_{\xi^\perp}))^2]} \leq 2\sqrt{\eta} \,, \end{aligned} \tag{4.14}$$

where we used Claim 4.2.8. Using Equation (4.10), we get that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(y - P(\mathbf{x}))] \geq \epsilon - 2\sqrt{\eta}$. which completes the proof of Lemma 4.2.7. □

Our final claim replaces $f - f_V$ by its polynomial approximation. By Hermite concentration arguments, we can show that we can use a polynomial $Q(\mathbf{x})$ of degree $O(1/\epsilon^4)$. More specifically, we show:

**Lemma 4.2.9.** *There exists a polynomial $Q(\mathbf{x})$ of degree $O(1/\epsilon^4)$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q(\mathbf{x})(y - P(\mathbf{x}))] \geq \epsilon/2 - 2\sqrt{\eta}$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q^2(\mathbf{x})] \leq 9$.*

*Proof.* We will require the following result from [KOS08] which bounds the Hermite concentration of LTFs.

**Fact 4.2.10** (Theorem 15 of [KOS08])**.** *Let $f \in \mathcal{C}$, and let $S$ be the Hermite expansion up to degree $k$ of $f$, i.e., $S(\mathbf{x}) = \sum_{|\alpha|\leq k} \hat{f}(\alpha)\mathrm{He}_\alpha(\mathbf{x})$. Then $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(S(\mathbf{x}) - f(\mathbf{x}))^2] = O(1/\sqrt{k})$.*

For any polynomial $Q(\mathbf{x})$, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q(\mathbf{x})(y - P(\mathbf{x}))] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x}) + (f(\mathbf{x}) - f_V(\mathbf{x})) - (f(\mathbf{x}) - f_V(\mathbf{x})))(y - P(\mathbf{x}))] \\ &\geq \epsilon - 2\sqrt{\eta} + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x}) - (f(\mathbf{x}) - f_V(\mathbf{x})))(y - P(\mathbf{x}))] \ , \quad (4.15)\end{aligned}$$

where we used Lemma 4.2.7. By choosing $Q(\mathbf{x}) = S(\mathbf{x}) - \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp} + \mathbf{z})]$, where we denote by $S(\mathbf{x})$ the Hermite expansion of $f$ truncated up to degree $k$, $S(\mathbf{x}) = \sum_{|\alpha|\leq k} \hat{f}(\alpha)\mathrm{He}_\alpha(\mathbf{x})$, we will show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}) - Q(\mathbf{x}))^2] = O(1/\sqrt{k})$ . Using the elementary inequality $(a + b)^2 \leq 2a^2 + 2b^2$, we get that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}) - Q(\mathbf{x}))^2] \leq 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - S(\mathbf{x}))^2] + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f_V(\mathbf{x}) - \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp})])^2] \ .$$

Moreover, by Jensen's inequality, it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f_V(\mathbf{x}) - \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp} + \mathbf{z})])^2] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x}) - S(\mathbf{x}))^2] = O(1/\sqrt{k}) \,,$$

where in the last equality we used Fact 4.2.10. Note that from the reverse triangle inequality, it holds that

$$\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[Q^2(\mathbf{x})]} \leq \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(f(\mathbf{x}) - f_V(\mathbf{x}))^2]} + O(1/k^{1/4}) \leq 2 + O(1/k^{1/4}) \,. \tag{4.16}$$

Choosing $k = O(1/\epsilon^4)$ and applying Cauchy-Schwartz to the Equation (4.15), we get

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[Q(\mathbf{x})(y - P(\mathbf{x}))] &\geq \epsilon - 2\sqrt{\eta} - \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(Q(\mathbf{x}) - (f(\mathbf{x}) - f_V(\mathbf{x})))^2] \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2]} \\ &\geq \epsilon/2 - 2\sqrt{\eta} \,,\end{aligned}$$

where we used the fact that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \leq \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-0)^2] \leq 1$; the polynomial $P(\mathbf{x})$ is closer to $y$ than the trivial polynomial 0. For this choice of $k$, Equation (4.16) gives $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[Q^2(\mathbf{x})] \leq 9$. This completes the proof of Lemma 4.2.9. □

By choosing $\eta = \epsilon^2/64$, Lemma 4.2.9 contradicts our assumption that $P(\mathbf{x})$ is $O(\epsilon^3)$-close to the polynomial $P'(\mathbf{x})$ that minimizes the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2]$, see Claim 4.2.6. This completes the proof of Proposition 4.2.1.

## 4.3 Agnostic Proper PTAS for Homogeneous Halfspaces

In this section, we provide a proper PTAS for agnostically learning homogeneous halfspaces, thereby establishing Theorem 4.1.2. Concretely, let $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$ be an optimal halfspace, i.e., $\mathrm{OPT} = \min_{h \in \mathcal{C}_0} \mathrm{err}^{\mathcal{D}}_{0-1}(h) = \mathrm{err}^{\mathcal{D}}_{0-1}(f)$. For any $\gamma, \epsilon \in (0,1)$ our algorithm computes a halfspace $h(\mathbf{x})$ such that $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \leq (1+\gamma)\mathrm{OPT} + \epsilon$.

The pseudocode of our algorithm is given in Algorithm 6.

**Algorithm 6** Agnostic Proper PTAS for Homogeneous Halfspaces

**procedure** AGNOSTIC-PROPER-PTAS($\gamma, \epsilon, \delta, \mathcal{D}$) ▷ $C$, $C'$ are absolute constants
**Input:** $\gamma > 0$, $\epsilon, \delta > 0$, and distribution $\mathcal{D}$
**Output:** A hypothesis $h \in \mathcal{C}$ such that $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \leq (1+\gamma)\mathrm{OPT} + \epsilon$ with probability $1 - \delta$
    $\sigma \leftarrow C' \mathrm{OPT}/\gamma$
    Let $\mathbf{w}_0$ be the normal vector of the homogeneous halfspace computed using Lemma 4.3.1
    **If** $\epsilon > C\,\mathrm{OPT}$:
        **return** $h_0(\mathbf{x}) = \mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x})$
    Let $\mathcal{D}_A$ be the distribution $\mathcal{D}$ after applying rejection sampling with $\mathbf{w}_0$ and $\sigma$, see Lemma 4.3.2
    Run Algorithm 5 on $\mathcal{D}_A$ with accuracy $\Theta(\gamma^2)$ and confidence $\delta$ to get $(\mathbf{w}, t)$
    **return** $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t)$
**end procedure**

### 4.3.1 Analysis of Algorithm 6: Proof of Theorem 4.1.2

The following lemma provides us with an efficient algorithm that learns a halfspace within error $O(\mathrm{OPT})$ in polynomial time. This halfspace serves as the initialization of Algorithm 6: we will use it to perform localization around it.

**Lemma 4.3.1** ([ABL17, DKTZ20c]). *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. There is an algorithm that draws $N = O((d/\epsilon^4)\log(1/\delta))$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N, d)$, and outputs a hypothesis $h \in \mathcal{C}$ such that, with probability at least $1 - \delta$, we have $\mathrm{err}_{0-1}^{\mathcal{D}}(h) \leq O(\mathrm{OPT}) + \epsilon$.*

The following lemma provides a "soft" localization procedure. Instead of performing rejection sampling inside a band around $\mathbf{w}_0$, i.e., $|\mathbf{x} \cdot \mathbf{w}_0| < \sigma$, we perform rejection sampling with weight $e^{-\mathbf{w}_0 \cdot \mathbf{x}(\sigma^{-2}-1)}$: samples that are far from the halfspace $\mathbf{w}_0$ are accepted with very small probability. Using this rejection sampling process, we get that the distribution conditional on acceptance is a normal distribution. This allows us to use our proper learning algorithm of Section 4.2.

**Lemma 4.3.2** (Lemma 4.7 of [DKS18]). *Let $\mathbf{w}_0 \in \mathbb{R}^d$ be a unit vector and let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Fix $\sigma \in (0, 1)$ and define the distribution $\mathcal{D}_A$ as follows: draw a sample $(\mathbf{x}, y)$ from $\mathcal{D}$ and accept it with probability $e^{-(\mathbf{w}_0 \cdot \mathbf{x})^2(\sigma^{-2}-1)/2}$. $\mathcal{D}_A$ is the distribution of $(\mathbf{x}, y)$ conditional on acceptance. The $\mathbf{x}$-marginal of $\mathcal{D}_A$ is $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{I} - (1 - \sigma^2)\mathbf{w}_0\mathbf{w}_0^{\mathsf{T}}$, and the probability that some point will be accepted is $\sigma$.*

The main technical tool of this section is the following proposition. It shows that, if a halfspace performs reasonably well with respect to the "localized" distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ of Lemma 4.3.2, then it cannot be very biased or have very large angle with the initial

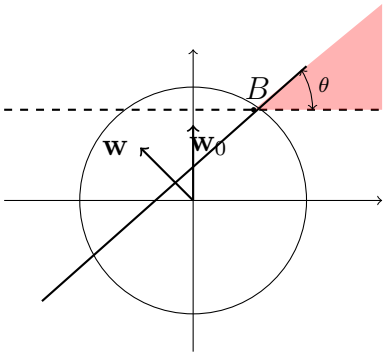



Figure 4.1: Localization technique, Lemma 4.3.4

guess $\mathbf{w}_0$. This allows us to prove that the halfspace that we find using the "localized" distribution will perform well over the initial Gaussian, $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

**Proposition 4.3.3.** *Let $\mathbf{w}_0 \in \mathbb{R}^d$ be a unit vector and let $\alpha, \gamma \in (0, 1/4]$. Let $\mathcal{D}_A$ be defined as in Lemma 4.3.2, i.e., its $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{I} - (1-\sigma^2)\mathbf{w}_0\mathbf{w}_0^T$ for some $\sigma \in (0, \cos(\pi\alpha))$. Moreover, assume that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x}) \neq y] \leq \alpha/3$. There exists an algorithm that runs in time $d^{\mathrm{poly}(1/(\gamma\alpha))}\log(1/\delta)$ and with probability at least $1-\delta$ returns a halfspace $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t)$ such that $|t| = O(\sigma\alpha)$, $\theta(\mathbf{w}, \mathbf{w}_0) = O(\sigma\alpha)$. Moreover, it holds*

$$\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq y] \leq \min_{\bar{h}\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\bar{h}(\mathbf{x}) \neq y] + \alpha\gamma\,.$$

The proof of Proposition is quite technical and is deferred to Section (4.3.2).

The following lemma is similar to the localization lemma (Lemma 2.1) given in [Dan15]. We need to adapt it to work in our setting, where we use a soft localization procedure (see Lemma 4.3.2), as opposed to a hard one.

**Lemma 4.3.4** (Gaussian Localization)**.** *Let $R(\mathbf{x})$ be the event that the sample $\mathbf{x}$ is*

*rejected from the rejection sampling procedure of Lemma 4.3.2 with vector* $\mathbf{w}_0$ *and* $\sigma = \Theta\left(\frac{\mathrm{OPT}}{\alpha}\right)$. *Let* $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x})$, $h'(\mathbf{x}) = \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t)$ *be halfspaces with* $t = O(\sigma\alpha)$ *and* $\theta(\mathbf{w}_0, \mathbf{w}) = O(\sigma\alpha)$. *Then,* $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), R(\mathbf{x})] = O(\alpha\,\mathrm{OPT})$.

*Proof.* Let $\theta = \theta(\mathbf{w}_0, \mathbf{w})$ and fix $r = \Theta(1/\alpha^{1/3})\max(1, t/\sin\theta)$. We define $B = r\sin\theta(\sqrt{1 - t^2/r^2} - t/r\frac{\cos\theta}{\sin\theta})$. Observe that in order for $B > 0$ we need $r \geq t/\sin\theta$, which is true by our assumptions. Also note that $B = \Theta(r\sin\theta)$. We have that

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), R(\mathbf{x})] = \\ \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| \geq B, R(\mathbf{x})] + \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| < B, R(\mathbf{x})]\,.\end{aligned}$$

We first bound from above the term $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| < B, R(\mathbf{x})]$. It holds that

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| < B, R(\mathbf{x})] &\leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}_0 \cdot \mathbf{x}| < B, R(\mathbf{x})] \\ &= \mathrm{erf}(B/\sqrt{2}) - \sigma\,\mathrm{erf}(B/(\sigma\sqrt{2})) \\ &= O(B^3/\sigma^2) = O(\alpha\,\mathrm{OPT})\,, \qquad (4.17)\end{aligned}$$

where erf is the error function and in the last equality we used the error of the Taylor expansion of degree-2. In order to bound the second term, we define $V$ to be the subspace spanned by the vectors $\mathbf{w}_0, \mathbf{w}$. Let $\mathbf{x}', x''$ be the solutions of the system of equations $\{\mathbf{w} \cdot \mathbf{x} + t = 0, \|\mathbf{x}_V\|_2 = r^2\}$; observe that $\min(|\mathbf{x}' \cdot \mathbf{w}_0|, |\mathbf{x}'' \cdot \mathbf{w}_0|) = B$, thus,

if $\|\mathbf{x}_V\|_2 \leq r$ and $|\mathbf{w}_0 \cdot \mathbf{x}| \geq B$, then $h(\mathbf{x}) = h'(\mathbf{x})$ (see Figure 4.1), thus

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| \geq B, R(\mathbf{x})] &\leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), |\mathbf{w}_0 \cdot \mathbf{x}| \geq B, R(\mathbf{x})] \\ &\leq C\theta e^{-r} = O(\alpha\,\mathrm{OPT})\,. \end{aligned} \tag{4.18}$$

Combining Equations (4.17), (4.18), we get $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[h(\mathbf{x}) \neq h'(\mathbf{x}), R(\mathbf{x})] = O(\alpha\,\mathrm{OPT})$. □

We are now ready to prove Theorem 4.1.2.

*Proof of Theorem 4.1.2.* Our analysis follows the cases of Algorithm 6. Initially, Algorithm 6 computes $h_0 = \mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x})$, using the algorithm of Lemma 4.3.1. From Lemma 4.3.1, we have that for this halfspace it holds, with probability at least $1 - \delta/2$, that $\mathrm{err}^{\mathcal{D}}_{0-1}(h_0) = C\,\mathrm{OPT} + \epsilon$ for some absolute constant $C > 1$. The runtime of this step is $\mathrm{poly}(d, 1/\epsilon)\log(1/\delta)$. Therefore, when $\epsilon > 2C\,\mathrm{OPT}$, we directly get that $\mathrm{err}^{\mathcal{D}}_{0-1}(h_0) \leq \mathrm{OPT} + \epsilon$.

For the case when $\epsilon \leq 2C\,\mathrm{OPT}$, Algorithm 6 returns a halfspace $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)$, where $\mathbf{w} \in \mathbb{R}^d$ and $t \in \mathbb{R}$. We show that for this hypothesis $h$ it holds $\mathrm{err}^{\mathcal{D}}_{0-1}(h) \leq (1+\gamma)\mathrm{OPT}+\epsilon$. Let $\mathcal{D}_A$ be the distribution conditional on acceptance (see Lemma 4.3.2) with parameters $\mathbf{w}_0$ (the normal vector of the halfspace $h_0$) and $\sigma = \Theta(\mathrm{OPT}/\alpha)$, for some parameter $\alpha = \Theta(\gamma)$ and $R(\mathbf{x})$ (resp. $A(\mathbf{x})$) be the event that the sample $\mathbf{x}$ is rejected (resp. accepted). The error of the hypothesis $h$ is

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y, A(\mathbf{x})] + \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y, R(\mathbf{x})]\,. \tag{4.19}$$

Let us first bound the $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y, A(\mathbf{x})] = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq y]\,\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[A(\mathbf{x})]$.

From Proposition 4.3.3, we have that with sample complexity and runtime $d^{\mathrm{poly}(1/\gamma)}\log(1/\delta)$ we get that, with probability at least $1-\delta/2$, it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x})\neq y] \leq \min_{\bar{h}\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\bar{h}(\mathbf{x})\neq y] + \alpha\gamma \leq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[f(\mathbf{x})\neq y] + \alpha\gamma\,,$$

and $|t| = O(\sigma\alpha)$, $\theta(\mathbf{w},\mathbf{w}_0) = O(\sigma\alpha)$. From Lemma 4.3.2, it holds $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[A(\mathbf{x})] = \sigma$, thus

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq y, A(\mathbf{x})] \leq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[f(\mathbf{x})\neq y, A(\mathbf{x})] + \alpha\gamma\sigma\,. \tag{4.20}$$

To bound $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq y, R(\mathbf{x})]$, observe that $\theta(\mathbf{w}_0,\mathbf{w}^*) = O(\mathrm{OPT})$, because $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h_0(\mathbf{x})\neq f(\mathbf{x})] = \theta(\mathbf{w}_0,\mathbf{w}^*)/\pi = O(\mathrm{OPT})$. Thus, with two applications of Lemma 4.3.4, we get $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})\neq h_0(\mathbf{x}), R(\mathbf{x})] = O(\alpha\mathrm{OPT})$ and $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq h_0(\mathbf{x}), R(\mathbf{x})] = O(\alpha\mathrm{OPT})$. Using the triangle inequality, we get

$$\begin{aligned}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq y, R(\mathbf{x})] &\leq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})\neq y, R(\mathbf{x})] + \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq h_0(\mathbf{x}), R(\mathbf{x})] \\ &+ \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})\neq h_0(\mathbf{x}), R(\mathbf{x})] = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})\neq y, R(\mathbf{x})] + O(\alpha\mathrm{OPT})\,. \end{aligned} \tag{4.21}$$

Substituting Equations (4.20) and (4.21) into Equation (4.19), we get

$$\begin{aligned}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq y] &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[f(\mathbf{x})\neq y, A(\mathbf{x})] + \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})\neq y, R(\mathbf{x})] + O(\alpha\mathrm{OPT} + \alpha\gamma\sigma) \\ &= \mathrm{OPT} + O((\alpha+\gamma))\mathrm{OPT} = (1+O(\gamma))\mathrm{OPT}\,,\end{aligned}$$

where we used the fact that $\alpha = \Theta(\gamma)$. Combining the above cases, we obtain that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq y] \leq (1+O(\gamma))\mathrm{OPT}+\epsilon$. Combining the runtime of the above two steps, we obtain that the total runtime of our algorithm is $d^{\mathrm{poly}(1/\gamma)}\mathrm{poly}(1/\epsilon)\log(1/\delta)$. □

### 4.3.2 Proof of Proposition 4.3.3

We first make the $\mathbf{x}$-marginal isotropic by multiplying samples with $\mathbf{\Sigma}^{-1/2}$ and then use the proper learning algorithm of Theorem 4.1.1 with target accuracy $\epsilon = \alpha\gamma$. The sample complexity and runtime are thus $d^{\mathrm{poly}(1/(\alpha\gamma))}$. From the guarantee of Theorem 4.1.1, we immediately obtain that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq y] \leq \min_{\bar{h}\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\bar{h}(\mathbf{x}) \neq y] + \alpha\gamma/3$. It now remains to bound the bias $t$ and the angle $\theta(\mathbf{w}, \mathbf{w}_0)$ of the returned halfspace $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} + t)$. Using our assumption for the misclassification error of $\mathbf{w}_0$ with respect to $\mathcal{D}_A$, we obtain that $\min_{\bar{h}\in\mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\mathrm{sign}(\mathbf{w}_0\cdot\mathbf{x}) \neq y] \leq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\mathrm{sign}(\mathbf{w}_0\cdot\mathbf{x}) \neq y] \leq \alpha$. From the triangle inequality, we obtain

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}_0\cdot\mathbf{x})] \leq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq y] + \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[\mathrm{sign}(\mathbf{w}_0\cdot\mathbf{x}) \neq y] \leq \alpha\,.$$

Therefore, we have

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}_0\cdot\mathbf{x})] = \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h(\mathbf{\Sigma}^{1/2}\mathbf{x}) \neq \mathrm{sign}((\mathbf{\Sigma}^{1/2}\mathbf{w}_0)\cdot\mathbf{x})]\,.$$

If the halfspace $h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} + t)$ has zero bias, i.e., $t = 0$, we have that by the spherical symmetry of the Gaussian distribution it holds

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h(\mathbf{\Sigma}^{1/2}\mathbf{x}) \neq \mathrm{sign}((\mathbf{\Sigma}^{1/2}\mathbf{w}_0)\cdot\mathbf{x})] = \theta(\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{\Sigma}^{1/2}\mathbf{w}_0)/\pi\,.$$

Unfortunately, the same is not true when one of the halfspaces has non-zero bias. However, we can prove that the angle $\theta(\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{\Sigma}^{1/2}\mathbf{w}_0)/\pi$ is still a *lower bound* on the probability of disagreement, i.e., $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h(\mathbf{\Sigma}^{1/2}\mathbf{x}) \neq \mathrm{sign}((\mathbf{\Sigma}^{1/2}\mathbf{w}_0)\cdot\mathbf{x})]$.

Formally, we prove the following claim showing that when one of the halfspaces is

homogeneous, the probability mass of the disagreement region is at least a constant multiple of the angle between the normal vectors. The proof follows from the observation that we can always minimize the disagreement probability between a homogeneous and an arbitrary halfspace by centering the Gaussian exactly at their intersection point.

**Claim 4.3.5.** *For* $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d, t \in \mathbb{R}$ *define the halfspaces* $h_0(\mathbf{x}) = \mathrm{sign}(\mathbf{u} \cdot \mathbf{x})$, $h_1(\mathbf{x}) = \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t)$. *It holds* $\mathbf{Pr}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[h_1(\mathbf{x}) \neq h_0(\mathbf{x})] \geq \theta(\mathbf{u}, \mathbf{v})/\pi$.

*Proof.* Denote $\theta = \theta(\mathbf{u}, \mathbf{v})$ and first assume that $\theta \in [0, \pi/2)$. Since the Gaussian distribution is invariant under rotations, for simplicity we may assume that $\mathbf{u}, \mathbf{v}$ span $\mathbb{R}^2$. Morover, assume that two halfspaces intersect at the origin $(0, 0)$ (if they do not intersect then the claimed lower bound on the disagreement $\mathbf{Pr}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[h_1(\mathbf{x}) \neq h_0(\mathbf{x})]$ is trivially true as their angle is 0). Moreover, assume that $\mathbf{u} = \mathbf{e}_1$ and that the Gaussian is centered at some point $(z, 0)$, i.e., a point that lies on the $\mathbf{x}_1$-axis. This follows from the fact that $h_0(\mathbf{x}) = \mathrm{sign}(\mathbf{u} \cdot \mathbf{x})$ is homogeneous. After we change coordinates, the halfspace $h_1$ is also homogeneous, with normal vector $\mathbf{v} = (-\sin\theta, \cos\theta)$. We will show that the disagreement between the two halfspaces is *minimized* where $z = 0$, i.e., when the Gaussian is centered on the intersection point of the two halfspaces. Using the above, we obtain that

$$\begin{aligned}
&\mathbf{Pr}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[h_1(\mathbf{x}) \neq h_0(\mathbf{x})] \\
&\quad = \int_{-\infty}^{0} \int_{\mathbf{x}_1 \tan\theta}^{0} e^{-((\mathbf{x}_1 - z)^2/2 - \mathbf{x}_2^2/2} \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{x}_1 + \int_{0}^{\infty} \int_{0}^{\mathbf{x}_1 \tan\theta} e^{-((\mathbf{x}_1 - z)^2/2 - \mathbf{x}_2^2/2} \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{x}_1 := q(z)
\end{aligned}$$

We will show that the function $q$ is minimized for $z = 0$. Taking the derivative with

respect to $z$, we obtain

$$q'(z) = \int_{-\infty}^{0}\int_{\mathbf{x}_1\tan\theta}^{0}(\mathbf{x}_1-z)e^{-((\mathbf{x}_1-z)^2/2-\mathbf{x}_2^2/2}\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{x}_1+\int_{0}^{\infty}\int_{0}^{\mathbf{x}_1\tan\theta}(\mathbf{x}_1-z)e^{-((\mathbf{x}_1-z)^2/2-\mathbf{x}_2^2/2}\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{x}_1\,.$$

Observe that $q'(-z) = -q'(z)$, i.e, $q'(z)$ is an odd function with $q'(0) = 0$. Thus, it can only be minimized at 0. We have that $q''(0) > 0$ and therefore $z = 0$ is the global minimizer of $q(z)$. The case $\theta \in [\pi/2, \pi]$ can be shown similarly. □

*Proof.* We first observe that

$$\begin{aligned}\Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] &= \Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma x}) \neq h_2(\mathbf{\Sigma x})] \\ &\geq \left|\Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma x}) \neq 0] - \Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_2(\mathbf{\Sigma x}) \neq 0]\right|\,,\end{aligned}$$

where in the last step we used triangle inequality. Moreover, using that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma x}) \neq 0] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma x})]$, we have

$$\Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_2(\mathbf{\Sigma x})]\right|\,.$$

Note that $h_1(\mathbf{\Sigma x}) = \mathrm{sign}(\mathbf{u}\cdot\mathbf{\Sigma x}/\left\|\mathbf{u}\mathbf{\Sigma}^{1/2}\right\|_2+r_1)$ and $h_2(\mathbf{\Sigma x}) = \mathrm{sign}(\mathbf{v}\cdot\mathbf{\Sigma x}/\left\|\mathbf{v}\mathbf{\Sigma}^{1/2}\right\|_2+r_2)$, thus

$$\begin{aligned}\Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] &\geq \left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_1(\mathbf{\Sigma x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h_2(\mathbf{\Sigma x})]\right| \\ &= \left|\Pr_{r\sim\mathcal{N}(0,1)}[r \leq r_1] - \Pr_{r\sim\mathcal{N}(0,1)}[r \leq r_2]\right|\,,\end{aligned}$$

which completes the proof. □

Using Claim 4.3.5 and the fact that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq \text{sign}(\mathbf{w}_0 \cdot \mathbf{x})] \leq \alpha$ that we showed above, we have that $\theta(\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{\Sigma}^{1/2}\mathbf{w}_0) \leq \pi\alpha$. We have

$$\cos(\theta(\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{\Sigma}^{1/2}\mathbf{w}_0)) = \frac{\mathbf{w} \cdot (\mathbf{\Sigma}\mathbf{w}_0)}{\sqrt{\mathbf{w} \cdot (\mathbf{\Sigma}\mathbf{w})}\sqrt{\mathbf{w}_0 \cdot (\mathbf{\Sigma}\mathbf{w}_0)}} = \frac{\sigma\ \mathbf{w} \cdot \mathbf{w}_0}{\sqrt{1 - (1-\sigma^2)(\mathbf{w} \cdot \mathbf{w}_0)^2}}\,.$$

Since $\theta(\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{\Sigma}^{1/2}\mathbf{w}_0) \leq \pi\alpha$ and cosine is a decreasing function in $[0, \pi]$, we obtain that

$$\sigma\ \mathbf{w} \cdot \mathbf{w}_0 \geq \cos(\pi\alpha)\sqrt{1 - (1-\sigma^2)(\mathbf{w} \cdot \mathbf{w}_0)^2}\,.$$

Solving this quadratic inequality with respect to $\mathbf{w} \cdot \mathbf{w}_0$, we obtain

$$\mathbf{w} \cdot \mathbf{w}_0 \geq \sqrt{\frac{1}{1 + \sigma^2(\frac{1}{\cos^2(\pi\alpha)} - 1)}} = \sqrt{\frac{1}{1 + \sigma^2 \tan^2(\pi\alpha)}}\,. \tag{4.22}$$

Using the inequality $\cos^{-1}(\sqrt{1/(1+x)}) \leq \sqrt{x}$ that holds for every $x \geq 0$, we obtain that the angle $\theta(\mathbf{w}, \mathbf{w}_0) \leq \sigma \tan(\pi\alpha)$. Using the elementary inequality $\tan(\pi x) \leq 4x$ that holds for all $x \in [0, 1/4]$, we can further simplify the bound for the angle to $\theta(\mathbf{w}, \mathbf{w}_0) \leq 4\sigma\alpha = O(\sigma\alpha)$.

We next bound the bias of the returned halfspace $h$. Now that we know that the angle between the vectors $\mathbf{w}_0, \mathbf{w}$ is small, we can use the following lower bound on the disagreement between two halfspaces to get that the bias cannot be too large.

**Claim 4.3.6.** *Let $h_1(\mathbf{x}) = \text{sign}(\mathbf{u} \cdot \mathbf{x} + t_1)$, $h_2(\mathbf{x}) = \text{sign}(\mathbf{v} \cdot \mathbf{x} + t_2)$ be two halfspaces. Let $r_1 = t_1 / \left\|\mathbf{u}\mathbf{\Sigma}^{1/2}\right\|_2$, $r_2 = t_2 / \left\|\mathbf{v}\mathbf{\Sigma}^{1/2}\right\|_2$. It holds*

$$\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \mathbf{Pr}_{r\sim\mathcal{N}(0,1)}[\min(r_1, r_2) \leq r \leq \max(r_1, r_2)]\,.$$

From Claim 4.3.6 and the fact that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_A}[h(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w}_0 \cdot \mathbf{x})] \leq \alpha$, we obtain that $\mathbf{Pr}_{r\sim\mathcal{N}(0,1)}[0 \leq r \leq |t_1|/\left\|\mathbf{w}\mathbf{\Sigma}^{1/2}\right\|_2] \leq \alpha$. Using the anti-anti-concentration property of the univariate Gaussian distribution, i.e., that $\mathbf{Pr}_{r\sim\mathcal{N}(0,1)}[0 \leq r \leq t] \geq \min(t/2, 2/3)$ and the fact that $\alpha \leq 1/4$, we obtain that $|t_1|/\left\|\mathbf{w}\mathbf{\Sigma}^{1/2}\right\|_2 \leq \alpha$. From Equation (4.22), we obtain that

$$\left\|\mathbf{w}\mathbf{\Sigma}^{1/2}\right\|_2 = \sqrt{1-(1-\sigma^2)(\mathbf{w}\cdot\mathbf{w}_0)^2} \leq \sqrt{\sigma^2\frac{1+\tan^2(\pi a)}{1+\sigma^2\tan^2(\pi\alpha)}} \leq 2\sigma\,,$$

using the fact that $\sigma < 1$ and $\alpha < 1/4$. Therefore, we conclude that $|t_1| \leq 2\sigma\alpha = O(\sigma\alpha)$.

This concludes the proof of Proposition 4.3.3.

## 4.4 Agnostic Proper Learning of ReLUs

In this section, we use our techniques to develop a proper agnostic learning algorithm that handles ReLU activations. We work in the standard $L_2$-regression setting, i.e., we want to find a weight vector $\mathbf{w}$ such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-\max(0,\mathbf{w}\cdot\mathbf{x}))^2] \leq \min_{f\in\mathcal{C}^\rho}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-f(\mathbf{x}))^2] + \epsilon\,,$$

where by $\mathcal{C}^\rho$ we denote the class of ReLU activations, i.e., $\mathcal{C}^\rho = \{\mathbf{x}\mapsto\max(0,\mathbf{w}\cdot\mathbf{x}+t) : \|\mathbf{w}\|_2 \leq 1, \mathbf{w}\in\mathbb{R}^d, t\in\mathbb{R}^d\}$. Moreover, we are going to use $\mathcal{C}_0^\rho = \{\mathbf{x}\mapsto\max(0,\mathbf{w}\cdot\mathbf{x}) : \|\mathbf{w}\|_2 \leq 1, \mathbf{w}\in\mathbb{R}^d\}$ and $\rho(\mathbf{x})$ to denote the ReLU activation function. Finally, observe that for $\mathbf{w}\in\mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq 1$ it holds $\max(0,\mathbf{w}\cdot\mathbf{x}) = \|\mathbf{w}\|_2\max(0,\mathbf{w}\cdot\mathbf{x}/\|\mathbf{w}\|_2)$. To keep the presentation simple we are going to assume, similarly to [DGK+20] that the observed labels $y$ are bounded in $[-1,1]$. For the rest of the section, we assume that for

the labels $y$, it holds $|y| < 1$.

The pseudocode of our algorithm is given in Algorithm 7.

**Algorithm 7** Agnostic Proper Algorithm for ReLU Regression

**procedure** AGNOSTIC-LEARNER($\epsilon, \delta, \mathcal{D}$)

**Input:** $\epsilon > 0$, $\delta > 0$ and distribution $\mathcal{D}$

**Output:** A hypothesis $h \in \mathcal{C}$ such as $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(h(\mathbf{x}) - y)^2] \leq \min_{f\in\mathcal{C}^\rho} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - y)^2] + \epsilon$ with probability $1 - \delta$.

$k \leftarrow C/\epsilon^{4/3}$, $\eta \leftarrow \epsilon^2/C$.

Find $P(\mathbf{x})$ such $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P'\in\mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + O(\epsilon^3)$.

Let $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\intercal]$.

Let $V$ be the subspace spanned by the eigenvectors of $\mathbf{M}$ whose eigenvalues are at least $\eta$.

Construct an $\epsilon$-cover $\mathcal{H}$ of hypotheses with normal vectors in $V$ ▷ see Fact 4.2.4.

Draw $\Theta(\frac{1}{\epsilon^2}\log(|\mathcal{H}|/\delta))$ i.i.d. samples from $\mathcal{D}$ and construct the empirical distribution $\widehat{\mathcal{D}}$.

$h \leftarrow \operatorname{argmin}_{h'\in\mathcal{H}} \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[(h'(\mathbf{x}) - y)^2]$

**return** $h$.

**end procedure**

**Theorem 4.4.1.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \mathbb{R}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Algorithm 7 draws $N = d^{O(1/\epsilon^{4/3})}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ samples from $\mathcal{D}$, runs in time $(d^{O(1/\epsilon^{4/3})} + (1/\epsilon)^{O(1/\epsilon^{10/3})})\log(1/\delta)$, and computes a hypothesis $h \in \mathcal{C}^\rho$ such that, with*

*probability at least* $1-\delta$*, we have that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-h(\mathbf{x}))^2] \le \min_{f\in\mathcal{C}_0^\rho} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-f(\mathbf{x}))^2] + \epsilon\,.$$

The main structural result of this section is the following proposition showing that we can perform dimension reduction by looking at high-influence directions of a low-degree polynomial.

**Proposition 4.4.2.** *Let* $C$ *be a sufficiently large constant, fix* $\epsilon > 0$*,* $k = C/\epsilon^{4/3}$*. Let* $P(\mathbf{x}) \in \mathcal{P}_k$ *be a polynomial such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \le \min_{P'\in\mathcal{P}_k} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2] + O(\epsilon^3)\,.$$

*Moreover, let* $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\intercal]$ *and* $V$ *be the subspace spanned by the eigenvectors of* $\mathbf{M}$ *with eigenvalues larger than* $\eta$*, where* $\eta = \epsilon^2/C$*. Then, for any function* $f \in \mathcal{C}_0^\rho$*, it holds*

$$\min_{\mathbf{v}\in V, \|\mathbf{v}\|_2\le 1, t\in\mathbb{R}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\mathbf{v}\cdot\mathbf{x}+t)-y)^2] \le \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-y)^2] + \epsilon\,.$$

The proof of Proposition 4.4.2 is similar to the proof of Proposition 4.2.1. We provide the details below for completeness.

*Proof.* Suppose for the sake of contradiction that there exists a hypothesis $f \in \mathcal{C}_0^\rho$ such that for every hypothesis $f' \in \mathcal{C}_V^\rho$, it holds

$$\min_{\mathbf{v}\in V, \|\mathbf{v}\|_2\le 1, t\in\mathbb{R}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\mathbf{v}\cdot\mathbf{x}+t)-y)^2] > \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-y)^2] + \epsilon\,. \tag{4.23}$$

Equivalently, from the above equation, we have that for every $\mathbf{v} \in V$ with $\|\mathbf{v}\|_2 \leq 1$ and $t \in \mathbb{R}$:

$$2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - \rho(\mathbf{v}\cdot\mathbf{x}+t))y] > \epsilon + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[f^2(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\rho^2(\mathbf{v}\cdot\mathbf{x}+t)]\,. \tag{4.24}$$

**Claim 4.4.3.** *It suffices to show that there exists some polynomial $Q(\mathbf{x})$ of degree at most $k$, with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[Q^2(\mathbf{x})] \leq 4$, that $(\epsilon/4)$-correlates with $(y - P(\mathbf{x}))$, i.e.,*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[Q(\mathbf{x})(y - P(\mathbf{x}))] \geq \epsilon/4\,.$$

*Proof.* We have that the polynomial $P(\mathbf{x}) + \zeta Q(\mathbf{x})$, for $\zeta = \Theta(\epsilon)$, has degree at most $k$ and decreases the value of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2]$ by at least $\Omega(\epsilon^2)$, which contradicts the optimality of $P(\mathbf{x})$, i.e., that $P(\mathbf{x})$, $O(\epsilon^3)$-close to the polynomial that minimizes the $L_2$ error with $y$. □

We now construct such a polynomial $Q(\mathbf{x})$. We have $f(\mathbf{x}) = \rho(\mathbf{w}\cdot\mathbf{x}) = \rho(\mathbf{w}_V\cdot\mathbf{x} + \mathbf{w}_{V^\perp}\cdot\mathbf{x})$, for some $0 < a \leq 1$. It holds that $\mathbf{w}_{V^\perp} \neq \mathbf{0}$ since otherwise we would have that $f \in \mathcal{C}_V^\rho$. For simplicity, we denote $\xi = \mathbf{w}_{V^\perp}/\left\|\mathbf{w}_{V^\perp}\right\|_2$. Notice that the direction $\xi$ is of low influence since $\xi \in V^\perp$. Recall, that by $\mathcal{D}_\xi$ we denote the projection of $\mathcal{D}$ onto the (one-dimensional) subspace spanned by $\xi$. We define $f_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z}+\mathbf{x}_V)]$: a convex combination of hypotheses in $\mathcal{C}_V^\rho$. In particular, $f_V(\mathbf{x})$ is a smoothed version of the hypothesis $\rho(\mathbf{w}_V\cdot\mathbf{x}+t)$, whose normal vector belongs in $V$. We first observe that

by (4.24) $f_V(\mathbf{x})$ cannot correlate too well with $y$:

$$2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] = 2 \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - \rho(\mathbf{w}\cdot\mathbf{x}_V + \mathbf{w}\cdot\mathbf{z}))y]]$$
$$\geq \epsilon + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[f^2(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\rho^2(\mathbf{w}\cdot\mathbf{x}_V + \mathbf{w}\cdot\mathbf{z})]] = \epsilon\,, \tag{4.25}$$

where the last equality follows by the fact that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[f^2(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{D}_{\xi^\perp}}\left[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[f^2(\mathbf{u}+\mathbf{z})]\right] = \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}\left[\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{D}_{\xi^\perp}}[\rho^2(\mathbf{u}\cdot\mathbf{w}_V+\mathbf{w}\cdot\mathbf{z})]\right] = \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}\left[\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{D}_\mathbf{x}}[\rho^2(\mathbf{u}_V\cdot\mathbf{w}+\mathbf{w}\cdot\mathbf{z})]\right].$$

Our argument consists two main claims. We first show that the function $f(\mathbf{x}) - f_V(\mathbf{x})$ correlates non-trivially with $y - P(\mathbf{x})$. Then we show that we can approximate $f(\mathbf{x}) - f_V(\mathbf{x})$ by a low degree polynomial $Q(\mathbf{x})$ that maintains non-trivial correlation with $y - P(\mathbf{x})$, see Claim 4.4.5. We start by proving the first claim.

**Claim 4.4.4.** *It holds*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(y - P(\mathbf{x}))] \geq \epsilon/2 - \sqrt{2\eta}\,.$$

*Proof.* We have $f_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z} + \mathbf{x}_{\xi^\perp})] = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[\rho(\mathbf{w}_V \cdot \mathbf{x}_V + \mathbf{w}\cdot\mathbf{z})]$ and, since $f_V$ is a convex combination of hypothesis in $\mathcal{C}_V^\rho$, from Equation (4.23), we see that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] \geq \epsilon$. Thus, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(y - P(\mathbf{x}))] = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - f_V(\mathbf{x}))y] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))P(\mathbf{x})]$$
$$\geq \epsilon/2 - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))P(\mathbf{x})]\,. \tag{4.26}$$

To deal with $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))P(\mathbf{x})]$, we first observe that for any function $g(\mathbf{x})$ depending only on the projection of $\mathbf{x}$ onto the subspace $\xi^\perp$, i.e., it holds $g(\mathbf{x}) = g(\mathbf{x}_{\xi^\perp})$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))g(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{v}\sim\mathcal{D}_{\xi^\perp}}\left[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{v}+\mathbf{z}) - f_V(\mathbf{v})]\; g(\mathbf{v})\right] = 0\,,$$

since for every $\mathbf{x} \in \mathbb{R}^d$, it holds $f_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{x}_{\xi^\perp} + \mathbf{z})] = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{x}_V + \mathbf{z})]$. Unfortunately, this is not true since $P(\mathbf{x})$ is not only a function of $\mathbf{x}_{\xi^\perp}$. However, since $V$ contains the high influence eigenvectors it holds that $P$ is almost a function of $\mathbf{x}_{\xi^\perp}$. In fact, we show that we can replace the polynomial $P$ by a different polynomial of degree at most $k$ that only depends on the projection of $\mathbf{x}$ on $\xi^\perp$. Similarly to the definition of the "smoothed" hypothesis $f_V$, we define $R(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[P(\mathbf{x}_{\xi^\perp} + \mathbf{z})]$. We first prove that $R(\mathbf{x})$ is close to $P(\mathbf{x})$ in the $L_2$ sense.

Now, adding and subtracting $R(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[P(\mathbf{x}_{\xi^\perp} + \mathbf{z})]$, we get

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))P(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(P(\mathbf{x}) - R(\mathbf{x}_{\xi^\perp}))] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))R(\mathbf{x}_{\xi^\perp})]\,.$$

The second term equals to zero, from the fact that $\mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[f(\mathbf{z} + \mathbf{x}_{\xi^\perp}) - f_V(\mathbf{x}_{\xi^\perp})] = 0$. Using Cauchy-Schwarz inequality, we get

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(P(\mathbf{x}) - R(\mathbf{x}_{\xi^\perp}))] &\le \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))^2]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(P(\mathbf{x}) - R(\mathbf{x}_{\xi^\perp}))^2]} \\ &\le \sqrt{2}\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(P(\mathbf{x}) - R(\mathbf{x}_{\xi^\perp}))^2]} \le \sqrt{2\eta}\,, \quad (4.27)\end{aligned}$$

where we used Claim 4.2.8. Using Equation (4.26), we get that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-f_V(\mathbf{x}))(y-P(\mathbf{x}))] \geq \epsilon/2 - \sqrt{2\eta}\,,$$

which completes the proof of Claim 4.4.4. □

**Claim 4.4.5.** *There exists a polynomial $Q(\mathbf{x})$ of degree $O(1/\epsilon^{4/3})$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q(\mathbf{x})(y-P(\mathbf{x}))] \geq \epsilon/4 - \sqrt{2\eta}$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q^2(\mathbf{x})] \leq 4$.*

*Proof.* For any polynomial $Q(\mathbf{x})$, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Q(\mathbf{x})(y-P(\mathbf{x}))] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x})+(f(\mathbf{x})-f_V(\mathbf{x}))-(f(\mathbf{x})-f_V(\mathbf{x})))(y-P(\mathbf{x}))] \\ &\geq \epsilon/2 - 2\sqrt{\eta} + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x})-(f(\mathbf{x})-f_V(\mathbf{x})))(y-P(\mathbf{x}))]\,,\end{aligned} \tag{4.28}$$

where we used Claim 4.4.4. By choosing $Q(\mathbf{x}) = S(\mathbf{x}) - \mathbf{E}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp}+\xi)]$, where we denote by $S(\mathbf{x})$ the Hermite expansion of $f$ truncated up to degree $k$, $S(\mathbf{x}) = \sum_{|\alpha|\leq k}\hat{f}(\alpha)\mathrm{He}_\alpha(\mathbf{x})$, we show that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-f_V(\mathbf{x})-Q(\mathbf{x}))^2] \leq \epsilon^2\,.$$

We need the following fact:

**Fact 4.4.6** ([GGK20])**.** *Let $f\in\mathcal{C}_0^\rho$, and let $S$ be the Hermite expansion up to $k$-degree of $f$, i.e., $S(\mathbf{x}) = \sum_{|\alpha|\leq k}\hat{f}(\alpha)\mathrm{He}_\alpha(\mathbf{x})$. Then $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(S(\mathbf{x})-f(\mathbf{x}))^2] = O(k^{-3/2})$.*

Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we get that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-f_V(\mathbf{x})-Q(\mathbf{x}))^2] \leq 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x})-S(\mathbf{x}))^2]+2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f_V(\mathbf{x})- \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp})])^2] \,.$$

Moreover, from Jensen's inequality, it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f_V(\mathbf{x}) - \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{D}_\xi}[S(\mathbf{x}_{\xi^\perp}+\xi)])^2] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - S(\mathbf{x}))^2] = O\left(k^{-3/2}\right) \,,$$

where in the last equality we used the Fact 4.4.6. Choose $k = \Theta(1/\epsilon^{4/3})$. Applying Cauchy-Schwartz to the Equation (4.28), we get

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[Q(\mathbf{x})(y-P(\mathbf{x}))] &\geq \epsilon/2 - \sqrt{2\eta} - \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(Q(\mathbf{x}) - (f(\mathbf{x}) - f_V(\mathbf{x})))^2] \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2]} \\ &\geq \epsilon/4 - \sqrt{2\eta} \,,\end{aligned}$$

where we used the fact that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \leq 2$. Note that from the reverse triangle inequality it holds that

$$\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[Q^2(\mathbf{x})]} \leq \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(f(\mathbf{x}) - f_V(\mathbf{x}))^2]} + \epsilon \leq \sqrt{2} + \epsilon \,. \tag{4.29}$$

Equation (4.29) gives $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[Q^2(\mathbf{x})] \leq 4$. This completes the proof of Claim 4.4.5, which completes the proof of Claim 4.4.5.

□

By choosing $\eta = \Theta(\epsilon^2)$, Claim 4.4.5 contradicts our assumption that $P(\mathbf{x})$ is $O(\epsilon^3)$-close to the polynomial $P'(\mathbf{x})$ that minimizes the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2]$. This completes the proof. □

The next lemma bounds the dimension of the subspace spanned by the high-influence directions of a polynomial that minimizes the $L_2$ error with the labels $y$.

Before we proceed to the proof of Theorem 4.4.1, we need an algorithm that calculates an approximate minimal polynomial for the Proposition 4.4.2.

**Lemma 4.4.7** ($L_2$-Polynomial Regression)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \mathbb{R}$ whose $\mathbf{x}$-marginal is the standard normal and whose labels are bounded by 1. Moreover, let $k \in \mathbb{Z}_+$, and $\epsilon, \delta > 0$. There is an algorithm that draws $N = (dk)^{O(k)} \log(1/\delta)/\epsilon^2$ samples, runs in time $\mathrm{poly}(N, d)$, and outputs a polynomial $P(\mathbf{x})$ of degree at most $k$ such that with probability $1-\delta$ it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-P'(\mathbf{x}))^2] + \epsilon$.*

The proof of this lemma is nearly identical to the proof of Lemma 4.2.3.

We need the following simple fact for ReLUs. An essentially identical fact was shown in [DKKZ20] Equation (2) for the zero threshold case. We provide the proof here for completeness.

**Fact 4.4.8.** *Let $f_1(\mathbf{x}) = \rho(\mathbf{v} \cdot \mathbf{x} + T)$ and $f_2(\mathbf{x}) = \rho(\mathbf{u} \cdot \mathbf{x} + T)$, for $T \in \mathbb{R}$ and $\mathbf{v}, \mathbf{u}$ unit vectors in $\mathbb{R}^d$. Then $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f_1(\mathbf{x}) - f_2(\mathbf{x}))^2] = O(\|\mathbf{v} - \mathbf{u}\|_2^2)$.*

*Proof.* The proof relies on the following fact.

**Fact 4.4.9** (Correlated Differences, Lemma 6 of [KTZ19])**.** *Let $r(\mathbf{x}) \in L_2(\mathbb{R}^d, \mathcal{N}^d)$ be differentiable almost everywhere and let*

$$D_\rho = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{I} & \rho\mathbf{I} \\ \rho\mathbf{I} & \mathbf{I} \end{pmatrix}\right).$$

*We call $\rho$-correlated a pair of random variables $(\mathbf{x}, \mathbf{y}) \sim D_\rho$. It holds*

$$\frac{1}{2} \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{z})\sim D_\rho}[(r(\mathbf{x}) - r(\mathbf{z}))^2] \leq (1-\rho) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|\nabla r(\mathbf{x})\|_2^2\right] .$$

Using this fact for $\rho = \mathbf{v}\cdot\mathbf{u}$, and using the approximation $(1 - \mathbf{v}\cdot\mathbf{u}) = \|\mathbf{u}-\mathbf{v}\|_2^2$ the result follows. □

We also need the following fact about the biases of ReLUs.

**Fact 4.4.10.** *Let $f_1(\mathbf{x}) = \rho(\mathbf{v}\cdot\mathbf{x} - T)$ and $f_2(\mathbf{x}) = \rho(\mathbf{v}\cdot\mathbf{x} - T')$ with $T' \geq T$. Then $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f_1(\mathbf{x}) - f_2(\mathbf{x}))^2] = O((T-T')^2 + 2T'(T'-T)e^{-T^2/2})$.*

*Proof.* Without loss of generality, we can assume that $\mathbf{v} = \mathbf{e}_1$. The result follows by noting that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f_1(\mathbf{x}) - f_2(\mathbf{x}))^2] \leq \int_{\mathbf{x}_1\geq T}(f_1(\mathbf{x}) - f_2(\mathbf{x}))^2\phi(\mathbf{x})\mathrm{d}\mathbf{x} + \int_{T'}^{T} f_2(\mathbf{x})\phi(\mathbf{x})\mathrm{d}\mathbf{x}$. □

We can now prove the main theorem of this section.

*Proof of Theorem 4.4.1.* Let $f \in \mathcal{C}_0^\rho$ such that the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x})y]$ is maximized. Using Lemma 4.4.7 on the labels $y$, with $N = d^{O(1/\epsilon^{4/3})}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ samples, we get an $k = O(1/\epsilon^{4/3})$-degree polynomial $P(\mathbf{x})$ and it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P'\in\mathcal{P}_k} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + \epsilon^3 ,$$

with probability $1 - \delta/2$. Applying Proposition 4.4.2 to the polynomial $P(\mathbf{x})$, we get that subspace $V$ spanned by the eigenvectors of the matrix $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x})\nabla P(\mathbf{x})^\intercal]$ with eigenvalues larger than $\eta = \Omega(1/\epsilon^2)$ contains a vector $\mathbf{v} \in V$, so that

$$\min_{\mathbf{v}\in V, \|\mathbf{v}\|_2\leq 1, t\in\mathbb{R}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\mathbf{v}\cdot\mathbf{x} + t) - y)^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - y)^2] + \epsilon . \tag{4.30}$$

Moreover, from Lemma 4.2.2, the dimension of $V$ is $O(1/\epsilon)^{10/3}$. Thus, applying Fact 4.2.4, we get that there exists a set $\tilde{V}$ which is an $\epsilon$-cover of the set $V$ with respect the $\ell_2$-norm of size $(1/\epsilon)^{O(1/\epsilon^{10/3})}$. We will use the set $\mathcal{T} = \{\epsilon/A, 2\epsilon/A, \ldots, 1\}$, where $A$ is a large enough constant, and show that it is a good cover of the parameter $a$ which is used as the guess of the norm of the vector $\mathbf{v}$. Finally, we need an effective cover for the biases $t$. Observe that from Fact 4.4.10, we need step-size $s = \epsilon^2/\sqrt{\log(1/\epsilon)}$ and the maximum negative value is $-\Theta(\sqrt{\log(1/\epsilon)})$. (If the value was larger, then the zero function would correlate as well.) We also need to bound the maximum positive value. We claim that the maximum positive value is some universal constant $C'$. This is because the error scales with the norm of the function that is returned by the algorithm and because we are trying to be competitive against the unbiased ReLU, the norm of the ReLU that is returned cannot be more than $2(\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\rho^2(\mathbf{w}\cdot\mathbf{x})] + 4\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y^2]) \leq C'$, for some large enough constant $C'$. Thus, the set of biases is $\mathcal{T}' = \{-C\sqrt{\log(1/\epsilon)}A/\epsilon, \ldots, 0, s, 2s, \ldots, CA/\epsilon'\}$, where we multiply with the minimal value of the guess of the norm. This is because if $\|\mathbf{v}\|_2 = \alpha$, then we have that $\rho(\mathbf{v}\cdot\mathbf{x} + t) = \alpha\rho(\mathbf{v}\cdot\mathbf{x}/\|\mathbf{v}\|_2 + t/\alpha)$.

We show that the set $\mathcal{H}$ is an effective cover, where $\mathcal{H} = \tilde{V} \times \mathcal{T} \times \mathcal{T}'$. We show that there exist a set of parameters $(\tilde{\mathbf{v}}, \tilde{a}, \tilde{t}) \in \mathcal{H}$ which define a ReLU that correlates with the labels as well as the function $f$. Fix the parameters $\mathbf{v}, \alpha, t$ which minimize the Equation (4.30). Indeed, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[a\rho(\mathbf{v}\cdot\mathbf{x}+t) - \tilde{a}\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))^2]^{1/2} &\leq a \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\mathbf{v}\cdot\mathbf{x}+t) - \rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+t))^2]^{1/2} \\ &+ a \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}) - \rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+t))^2]^{1/2} + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t})^2]^{1/2}|(a-\tilde{a})| \,. \end{aligned} \tag{4.31}$$

Applying Facts 4.4.8 and 4.4.10, we get that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[a\rho(\mathbf{v}\cdot\mathbf{x}+t)-\tilde{a}\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))^2]\leq O(\epsilon)\,.$$

Thus, from the triangle inequality, we get that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\tilde{a}\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t})-y)^2]\\
&\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\rho(\mathbf{v}\cdot\mathbf{x}+t)-y)^2]+\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[a\rho(\mathbf{v}\cdot\mathbf{x}+t)-\tilde{a}\rho(\tilde{\mathbf{v}}\cdot\mathbf{x}+\tilde{t}))^2]\\
&\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x})-y)^2]+O(\epsilon)\,.
\end{aligned}$$

To complete the proof, it remains to show that Step 11 outputs a hypothesis close to the minimizer inside $\mathcal{H}$. We need the following claim:

**Claim 4.4.11.** *Let $h\in\mathcal{C}^\rho$ and let $\widehat{\mathcal{D}}$ be the empirical distribution with $N=O((1/\epsilon^2)\log(1/\delta))$ samples. Then, with probability $1-\delta$, it holds*

$$|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[(y-h(\mathbf{x}))^2]-\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-h(\mathbf{x}))^2]|\leq\epsilon\,.$$

*Proof.* We need first to prove that with probability $1-\delta$ it holds:

$$|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[yh(\mathbf{x})]-\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[yh(\mathbf{x})]|\leq\epsilon\,.$$

Using Markov's inequality, we have

$$\begin{aligned}\mathbf{Pr}[|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[h(\mathbf{x})y] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})y]| \geq \epsilon] &\leq \frac{1}{N\epsilon^2}\mathbf{Var}[h(\mathbf{x})y] \\ &\leq \frac{1}{N\epsilon^2}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[h^2(\mathbf{x})y^2] \\ &\leq O\left(\frac{1}{N\epsilon^2}\right) ,\end{aligned}$$

where we used the fact that our functions are bounded in $L_2$-norm. With the same procedure we bound the difference $|\mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}[h^2(\mathbf{x})] - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[h^2(\mathbf{x})]| \leq \epsilon$. By using the fact that $N = O(1/\epsilon^2)$, we get our result for constant probability. By applying a standard probability amplification technique, we can boost the confidence to $1-\delta$ with $N' = O(N\log(1/\delta))$ samples. □

Therefore, from Claim 4.4.11, it follows that $O(\frac{1}{\epsilon^2}\log(\mathcal{H}/\delta))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most $\epsilon$ with probability at least $1-\delta/2$.

To bound the runtime of the algorithm, we note that $L_2$-regression has runtime $d^{O(1/\epsilon^{4/3})}\mathrm{poly}(1/\epsilon)\log(1/\delta)$ and the exhaustive search over an $\epsilon$-cover takes time $(1/\epsilon)^{O(1/\epsilon^{10/3})}\log(1/\delta)$ time. The total runtime of our algorithm in the case where $1/\epsilon^{10/3} \leq d$ is

$$\left(d^{O(1/\epsilon^{4/3})} + (1/\epsilon)^{O(1/\epsilon^{10/3})}\right)\log(1/\delta)\,.$$

In the case where $1/\epsilon^{10/3} > d$, one can directly do a brute-force search over an $\epsilon$-cover of the $d$-dimensional unit ball: we do not need to perform our dimension-reduction process and the runtime is bounded above by the previous case. □

# Part II

# Robustly Learning Monotone Single-Index-Models

# Chapter 5

# Robustly Learning Monotone GLMs under Well-Behaved-Distributions

## 5.1 Introduction

Single-index models (SIMs) [Ich93, HJS01, HMS+04, DJS08, KS09, KKSK11, DH18] are a classical supervised learning model extensively studied in statistics and machine learning. SIMs capture the common assumption that the target function $f$ depends on an unknown direction $\mathbf{w}$, i.e., $f(\mathbf{x}) = u(\mathbf{w} \cdot \mathbf{x})$ for some link (a.k.a. activation) function $u : \mathbb{R} \mapsto \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$. In most settings, the link function is unknown and is assumed to satisfy certain regularity properties. Classical works [KS09, KKSK11] studied the efficient learnability of SIMs for monotone and Lipschitz link functions and data distributed on the unit ball. These early algorithmic results succeed in the realizable setting (i.e., with clean labels) or in the presence of zero-mean label noise.

The focus of this work is on learning SIMs in the challenging agnostic (or adversarial label noise) model [Hau92, KSS94], where no assumptions are made on the labels

of the examples and the goal is to compute a hypothesis that is competitive with the *best-fit* function in the class. Importantly, as will be formalized below, in this chapter we assume prior knowledge to the link function where as in Chapter 8 we do not. In more detail, let $\mathcal{D}$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ and $\mathcal{L}_2(h) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(h(\mathbf{x}) - y)^2]$ be the squared loss of the hypothesis $h : \mathbb{R}^d \to \mathbb{R}$ with respect to $\mathcal{D}$. Given i.i.d. samples from $\mathcal{D}$, the goal of the learner is to output a hypothesis $h$ with squared error competitive with OPT, where $\mathrm{OPT} = \inf_{f\in\mathcal{C}} \mathcal{L}_2(f)$ is the best attainable error by any function in the target class $\mathcal{C}$ ($\mathcal{C}$ contains one element in the case of known link-function).

In this part of the thesis we answer the following question:

*Is there an efficient constant-factor approximate learner*
*for single neurons under well-behaved distributions?*

In particular, we consider the following class of well-behaved distributions.

**Definition 5.1.1** (Well-behaved Distributions)**.** *Let $L, R > 0$. An isotropic (i.e., zero mean and identity covariance) distribution $\mathcal{D}_{\mathbf{x}}$ on $\mathbb{R}^d$ is called $(L, R)$-well-behaved if for any projection $(\mathcal{D}_{\mathbf{x}})_V$ of $\mathcal{D}_{\mathbf{x}}$ onto a subspace $V$ of dimension at most two, the corresponding pdf $\gamma_V$ on $\mathbb{R}^2$ satisfies the following:*

- *For all $\mathbf{x} \in V$ such that $\|\mathbf{x}\|_\infty \leq R$ it holds $\gamma_V(\mathbf{x}) \geq L$ (anti-anti-concentration).*
- *For all $\mathbf{x} \in V$ it holds that $\gamma_V(\mathbf{x}) \leq (1/L)(e^{-L\|\mathbf{x}\|_2})$ (anti-concentration and concentration).*

*When the parameters $L, R$ are bounded above by universal constants (independent of the dimension), we will simply say that the distribution $\mathcal{D}_{\mathbf{x}}$ is well-behaved.*

The class of well-behaved distributions is fairly broad. Specifically, isotropic log-concave distributions are well-behaved, i.e., they are $(L, R)$-well-behaved for some $L, R = O(1)$, see, e.g., [LV07, KLT09]. Similar assumptions were introduced in [DKTZ20a] and have been used in various classification and regression settings [DKTZ20c, DKTZ20b, DKK+20, DKK+21b, FCG20, FCG21, ZL21, ZFG21].

**Learning Sigmoidal Activations** Our first main result holds for a natural class of activations that roughly have "sigmoidal" shape.

**Definition 5.1.2** (Sigmoidal Activations)**.** *Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be a non-decreasing activation function and $\tau, \mu, \xi > 0$. We say that $\sigma$ is $(\tau, \mu, \xi)$-sigmoidal if it satisfies (a) $\sigma'(t) \geq \tau$, for all $t \in [-1, 1]$, and (b) $\sigma'(t) \leq \xi e^{-\mu|t|}$, for all $t \in \mathbb{R}$. We will simply say that an activation is "sigmoidal" when $\tau, \mu, \xi$ are universal constants.*

Arguably the most popular sigmoidal activation is the logistic activation or sigmoid, corresponding to $\sigma(t) = 1/(1 + e^{-t})$. Other well-studied sigmoidal activations include the hyperbolic tangent, the Gaussian error function, and the ramp activation (see Figure 5.1). We note that all these activations satisfy the requirement of Definition 5.1.2 for some universal constants $\tau, \mu, \xi$. In what follows, we will simply refer to them as sigmoidal.

The most commonly used method to solve our learning problem in practice is to directly attempt to minimize the $L_2^2$-loss via (stochastic) gradient descent. Due to the non-convexity of the objective, this method is of a heuristic nature in general and comes with no theoretical guarantees, even in noiseless settings. In our setting, the situation is even more challenging due to the adversarial noise in the labels. Indeed, we show that the "vanilla" $L_2^2$-objective may contain bad local optima, even under Gaussian marginals.

Specifically, even with an arbitrarily small amount of adversarial noise, the vanilla $L_2^2$ objective will have local-minima whose $L_2^2$-error is larger than $1/2$ (which is essentially trivial, since sigmoidal activations take values in $[-1, 1]$); see Proposition 5.2.1 and Figure 5.2.

Our main structural result for sigmoidal activations is that we can "correct" the optimization landscape of the $L_2^2$-loss by introducing a standard $\ell_2$-regularization term. We prove the following theorem showing that any stationary point of the regularized $L_2^2$-loss is approximately optimal.

**Theorem 5.1.3** (Informal: Landscape of Sigmoidals)**.** *For sigmoidal activations and $\epsilon$-corrupted well-behaved distributions, any (approximate) stationary point $\bar{\mathbf{w}}$ of the $\ell_2$ regularized objective $F_\rho(\mathbf{w}) = F(\mathbf{w}) + (\rho/2)\|\mathbf{w}\|_2^2$ with $\rho = \Theta(\epsilon^3)$, satisfies $F(\bar{\mathbf{w}}) = O(\epsilon)$.*

Standard gradient methods (such as SGD) are known to efficiently converge to stationary points of non-convex objectives under certain assumptions. By running any such method on our regularized objective, we readily obtain an efficient algorithm that outputs a weight vector with $L_2^2$-loss $O(\epsilon)$. This is already a new result in this context. Yet, black-box application of optimization results for finding stationary points of non-convex functions would resulting in a sample complexity with suboptimal dependence on $\epsilon$, up to polynomial factors.

Aiming towards an algorithmic result with near-optimal sample complexity, we perform a "white-box" analysis of gradient descent, leveraging the optimization landscape of sigmoidals. Specifically, we show that "vanilla" gradient descent, with a fixed step size, finds an approximately optimal solution when run on the empirical (regularized) $L_2^2$-loss with a near-optimal number of samples.

**Theorem 5.1.4** (Informal: Learning Sigmoidals via Gradient Descent)**.** *For sigmoidal activations and $\epsilon$-corrupted well-behaved distributions, gradient descent on the empirical regularized loss $\widehat{F}_\rho(\cdot)$ with $N = \widetilde{\Theta}(d/\epsilon)$ samples, converges, in $\mathrm{poly}(1/\epsilon)$ iterations, to a vector $\widetilde{\mathbf{w}}$ satisfying $F(\widetilde{\mathbf{w}}) \leq O(\epsilon)$ with high probability.*

Theorem 5.1.4 gives the first efficient constant-factor approximate learner for sigmoid activations in the presence of adversarial label noise, answering an open problem of [DGK+20]. As an additional bonus, our algorithm is simple and potentially practical (relying on gradient-descent), has near-optimal sample complexity (see Lemma C.1.3), and succeeds for a broad family of bounded activation functions (i.e., the ones satisfying Definition 5.1.2).

For simplicity of the exposition, we have restricted our attention to sigmoidal activations (corresponding to $\tau, \mu, \xi$ being universal constants in Definition 5.1.2) and well-behaved distributions (corresponding to $L, R = O(1)$ in Definition 5.1.1). In the general case, we note that the complexity of our algorithm is polynomial in the parameters $L, R, \tau, \mu, \xi$, see Theorem 5.4.1.

**Learning Unbounded Activations** We next turn our attention to activation functions that are not uniformly bounded. The most popular such activation is the ReLU function $\sigma(t) = \mathrm{Relu}(t) = \max(0, t)$. Our algorithmic results apply to the following class of unbounded activations.

**Definition 5.1.5** (Unbounded Activations)**.** *Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be a non-decreasing activation function and $\alpha, \lambda > 0$. We say that $\sigma$ is $(\alpha, \lambda)$-unbounded if it satisfies (a) $\sigma$ is $\lambda$-Lipschitz and (b) $\sigma'(t) \geq \alpha$, for all $t \in [0, +\infty)$. We will simply say that an activation is unbounded when the parameters $\alpha, \lambda$ are universal constants.*

We use the term "unbounded" for these activations, as they tend to $\infty$ as $t \to +\infty$. Most well-known unbounded activation functions such as the ReLU, Leaky-ReLU, ELU are $(\alpha, \lambda)$-unbounded for some absolute constants $\alpha, \lambda > 0$ (see Figure 5.4). For example, the ReLU activation is $(1, 1)$-unbounded.

Our main structural result for unbounded activations is that all stationary points $\mathbf{w}$ of the $L_2^2$-loss that lie in the halfspace $\mathbf{w}^* \cdot \mathbf{w} \geq 0$, where $\mathbf{w}^*$ is the optimal weight vector, are approximately optimal. In more detail, we establish the following result.

**Theorem 5.1.6** (Informal: Landscape of Unbounded Activations)**.** *For unbounded activations and $\epsilon$-corrupted well-behaved distributions, any stationary point $\bar{\mathbf{w}}$ of $F(\mathbf{w})$ with $\mathbf{w} \cdot \mathbf{w}^* \geq 0$ satisfies $F(\bar{\mathbf{w}}) = O(\epsilon)$, and any $\mathbf{w}$ with $\mathbf{w} \cdot \mathbf{w}^* \geq 0$ and $F(\mathbf{w}) = \Omega(\epsilon)$, satisfies $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) > 0$.*

We remark that the constant in the error guarantee of the above theorem only depends on the (universal) constants of the distribution and the activation and not on the radius $W$ of Definition 1.3.15. Interestingly, Theorem 5.1.6 does not preclude the existence of suboptimal stationary points. On the other hand, similarly to the case of sigmoidal activations, our structural result can be readily combined with "black-box" optimization to efficiently find a constant-factor approximately optimal weight vector $\mathbf{w}$ (see Section 5.2). To obtain near-optimal sample complexity, we again perform an "white-box" analysis of (approximate) gradient descent, showing that simply by initializing at $\mathbf{0}$ we can avoid bad stationary points.

**Theorem 5.1.7** (Informal: Learning Unbounded Activations via Gradient Descent)**.** *For unbounded activations and $(\epsilon, W)$-corrupted well-behaved distributions, (approximate) gradient descent on the $L_2^2$-loss $F(\cdot)$ with sample size $N = \widetilde{\Theta}(dW^2 \max(\mathrm{polylog}(W/\epsilon), 1))$*

*and* $\mathrm{polylog}(1/\epsilon)$ *iterations, converges to a vector* $\widetilde{\mathbf{w}} \in \mathbb{R}^d$*, satisfying* $F(\widetilde{\mathbf{w}}) \leq O(\epsilon)$ *with high probability.*

Theorem 5.1.7 is a broad generalization of the main result of [DGK+20], which gave a constant-factor approximation (using a different approach) for the special case of ReLU activations under isotropic log-concave distributions. While the result of [DGK+20] was tailored to the ReLU activation, Theorem 5.1.7 works for a fairly broad class of unbounded activations (and a more general class of distributions). A key conceptual difference between the two results lies in the overall approach: The prior work [DGK+20] used a *convex* surrogate for optimization. In contrast, we leverage our structural result and directly optimize the natural non-convex objective.

Additionally, Theorem 5.1.7 achieves significantly better (and near-optimal) sample complexity as a function of both $d$ and $1/\epsilon$. In more detail, the algorithm of [DGK+20] had sample complexity $\Omega(d/\epsilon)$ (their results are phrased for the special case that $W = 1$), while our algorithm has sample complexity $\widetilde{O}(d)\mathrm{polylog}(1/\epsilon)$ — i.e., near-linear in $d$ and polylogarithmic in $1/\epsilon$.

## 5.2 The Landscape of the $L_2^2$ Loss

In this section, we present our results on the landscape of the $L_2^2$ loss for the sigmoidal activation functions of Definition 5.1.2 and the unbounded activation functions of Definition 5.1.5. Before we proceed, we remark again that Definition 5.1.2 models a general class of bounded activation functions, including, for example, the logistic activation, $\sigma(t) = 1/(1 + e^{-t})$, the hyperbolic tangent, $\sigma(t) = \tanh(t)$, the Gaussian error function, and the ramp activation, see Figure 5.1.

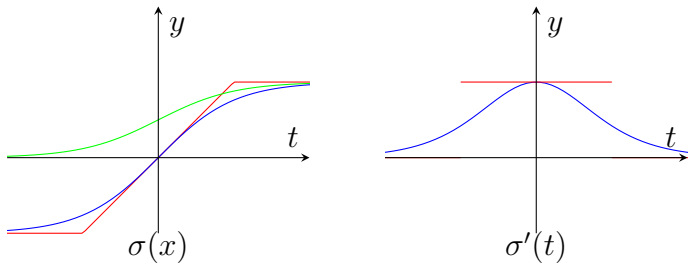


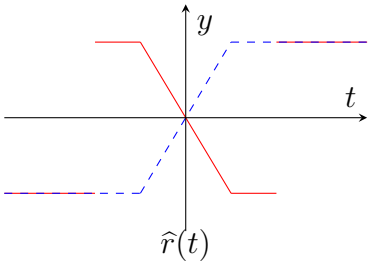

Figure 5.1: The $\tanh(\cdot)$ activation is $(0.4, 1, 1.4)$-sigmoidal. The ramp activation is $(1, 1, 3)$-sigmoidal. The logistic activation is $(0.19, 1, 1)$-sigmoidal. In the right figure, we plot the derivative of the ramp activation that is simply a rectangular function $\mathbb{1}\{|t| \leq 1\}$ (drawn in red) and the derivative of $\tanh(t)$ (drawn in blue). Notice that both decay at least exponentially fast: the derivative of the ramp activation is non-zero only in the interval $[-1, 1]$ and the derivative of $\tanh(t)$ decays exponentially fast, i.e., it is always smaller than $(4/3)e^{-|x|}$.

We first show that we can construct noisy instances that have "bad" stationary points, i.e., local-minima whose $L_2^2$ loss is $\omega(\epsilon)$. Perhaps surprisingly, this is the case even when the underlying $\mathbf{x}$-marginal is the standard normal and the level of corruption $\epsilon$ of the corresponding labeled instance $\mathcal{D}$ is arbitrarily small. Moreover, the sigmoidal activation used in the construction of the counterexample is very simple (in particular, we use the ramp activation). We show that, even though the constructed instance is only $\epsilon$-corrupted, there exists a local minimum whose $L_2^2$ loss is at least $\omega(\epsilon)$ (and in fact larger than some universal constant). An example of the noisy $L_2^2$ landscape is shown in Figure 5.2.

**Proposition 5.2.1** (Vanilla $L_2^2$ has "Bad" Local-Minima)**.** *For any $\epsilon \in (0, 1]$, there exists a well-behaved sigmoidal activation $\sigma(\cdot)$ and an $\epsilon$-corrupted distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{\pm 1\}$ with standard Gaussian $\mathbf{x}$-marginal such that $F^{\mathcal{D},\sigma}(\cdot)$ has a local minimum $\mathbf{u}$ with $F^{D,\sigma}(\mathbf{u}) \geq 1/2$.*

*Proof.* We define the following deterministic noise function

$$\widehat{r}(t) = \begin{cases} -1 & \text{if } t \le -2 \\ +1 & \text{if } -2 \le t \le -1 \\ -t & \text{if } -1 \le t \le 1 \\ -1 & \text{if } -1 \le t \le 2 \\ +1 & \text{if } t \ge 2 \end{cases}$$




In what follows, we denote by $\gamma(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ the density of the standard normal distribution. It suffices to consider noise that only depends on a single direction, i.e., $y(\mathbf{x}) = \widehat{r}(\mathbf{x}_1/\epsilon)$. We first show that there exists a vector with small $L_2^2$ loss. Take $\mathbf{v} = \mathbf{e}_1/\epsilon$, where $\mathbf{e}_1 = (1, 0, \ldots, 0)$ is the first vector of the standard orthogonal basis of $\mathbb{R}^d$. It holds that

$$F(\mathbf{v}) = (1/2) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(r(\mathbf{x}_1/\epsilon) - y(\mathbf{x}_1))^2] = (1/2) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(r(\mathbf{x}_1/\epsilon) - \widehat{r}(\mathbf{x}_1/\epsilon))^2]$$
$$= (1/2) \int_{-\epsilon}^{\epsilon} (t/\epsilon + t/\epsilon)^2 \gamma(t) dt + \int_{-2\epsilon}^{-\epsilon} 4\gamma(t) dt \le \frac{\sqrt{2}}{\sqrt{\pi}\epsilon^2} \int_{-\epsilon}^{\epsilon} t^2 dt + 4\epsilon = O(\epsilon) \,.$$

We next show that there exists a "bad" stationary point $\mathbf{u}$, i.e., a $\mathbf{u}$ with $\nabla F(\mathbf{u}) = 0$ and $F(\mathbf{u}) \ge 1/2$. We have

$$\nabla F(\mathbf{u}) = \mathbf{E}[(r(\mathbf{u}\cdot\mathbf{x}) - y(\mathbf{x}))r'(\mathbf{u}\cdot\mathbf{x})\mathbf{x}] \,.$$

We shall take $\mathbf{u} = -\mathbf{e}_1/\epsilon$, i.e., the opposite direction of the almost optimal vector $\mathbf{v}$ that we used above. Using the coordinate-wise independence of the Gaussian distribution, we have that for every orthogonal direction $\mathbf{e}_i$, $i \ge 2$, it holds that $\nabla F(\mathbf{u}) \cdot \mathbf{e}_i = 0$. For

the direction $\mathbf{e}_1$, we obtain

$$\begin{aligned}\nabla F(\mathbf{u}) \cdot \mathbf{e}_1 &= \mathbf{E}[(r(-\mathbf{x}_1/\epsilon) - \widehat{r}(\mathbf{x}_1/\epsilon))\ \mathbb{1}\{|\mathbf{x}_1| \leq \epsilon\}\mathbf{x}_1] \\ &= \mathbf{E}[(r(-\mathbf{x}_1/\epsilon) - \widehat{r}(-\mathbf{x}_1/\epsilon))\ \mathbb{1}\{|\mathbf{x}_1| \leq \epsilon\}\mathbf{x}_1] = 0\,.\end{aligned}$$

We next proceed to show that $\mathbf{u}$ is a local minimum of the population $L_2^2$-loss $F$. We compute the Hessian of $F$ at $\mathbf{u}$. In what follows, we denote by $\delta(t)$ the standard Dirac delta function. We have that

$$\nabla^2 F(\mathbf{u}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(r'(\mathbf{u}\cdot\mathbf{x}))^2\mathbf{x}\mathbf{x}^T] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(r(\mathbf{u}\cdot\mathbf{x}) - y(\mathbf{x}))(\delta(\mathbf{u}\cdot\mathbf{x}+1) - \delta(\mathbf{u}\cdot\mathbf{x}-1))\mathbf{x}\mathbf{x}^T]\,.$$

For $i \neq j$, using the fact that the Gaussian marginals are independent, we have that $(\nabla^2 F(\mathbf{u}))_{ij} = 0$. We next observe that the second term of the Hessian vanishes. We have

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(r(\mathbf{u}\cdot\mathbf{x}) - y(\mathbf{x}))(\delta(\mathbf{u}\cdot\mathbf{x}+1) - \delta(\mathbf{u}\cdot\mathbf{x}-1))\mathbf{x}_i^2] \\ &\qquad = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(r(-\mathbf{x}_1/\epsilon) - r(\mathbf{x}_1/\epsilon))(\delta(\mathbf{x}_1/\epsilon+1) - \delta(\mathbf{x}_1/\epsilon-1))\mathbf{x}_i^2] = 0\,,\end{aligned}$$

where we used the fact that both $r(-\mathbf{x}_1/\epsilon), \widehat{r}(\mathbf{x}_1/\epsilon)$ are continuous at $\mathbf{x}_1 = \pm\epsilon$, and it holds $r(\pm 1) = \widehat{r}(\pm 1)$.

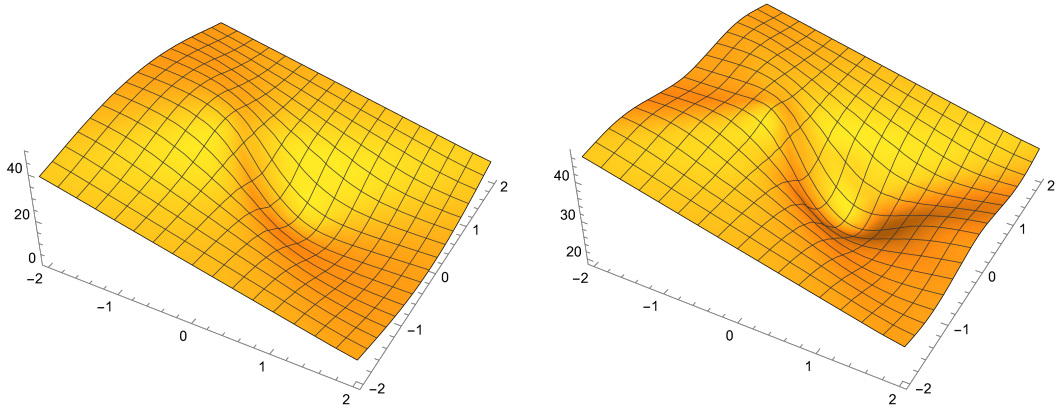



Figure 5.2: The non-convex optimization landscape of the "vanilla" $L_2^2$ loss. The activation is a ramp function $r(t)$, defined as $r(t) \triangleq (-1)\ \mathbb{1}\{t < -1\} + t\ \mathbb{1}\{|t| \leq 1\} + (+1)\ \mathbb{1}\{t > 1\}$. The $\mathbf{x}$-marginal is the uniform distribution on the square $[-2, 2] \times [-2, 2]$. The "true" underlying weight vector is $\mathbf{w}^* = (1, 0)$. In the left figure, we plot the noiseless population objective $F(\mathbf{w})$ as a function of $\mathbf{w} \in [-2, 2] \times [-2, 2]$, where $y = r(\mathbf{w}^* \cdot \mathbf{x})$. We observe that even though the objective is non-convex, in this case, the only stationary point is the true weight vector $\mathbf{w}^*$. In the right figure, we introduce noise and observe that the objective has another stationary point, which in fact is $-\mathbf{w}^*$. Finally, we notice that the landscape becomes more "flat" as we move away from the origin and the "bad" stationary point is a local-minimum that the "noise" was able to create in a particularly flat region.

To complete the proof, we need to show that the $L_2^2$ loss of $\mathbf{u}$ is large. It holds that

$$\begin{aligned} F(\mathbf{u}) &= \frac{1}{2}\,\mathbf{E}[(r(-\mathbf{x}_1/\epsilon) - r'(\mathbf{x}_1/\epsilon))^2] \\ &\geq \frac{1}{2}\left(\int_{-\infty}^{-\epsilon} (r(-t/\epsilon) - r'(t/\epsilon))^2 \gamma(t)dt + \int_{\epsilon}^{+\infty} (r(-t/\epsilon) - r'(t/\epsilon))^2 \gamma(t)dt\right) \\ &\geq \frac{1}{2}\left(\int_{-\infty}^{-\epsilon} 4\gamma(t)dt + \int_{\epsilon}^{+\infty} 4\gamma(t)dt\right) \geq 4\int_{1}^{\infty} \gamma(t)dt \geq 1/2\,. \end{aligned}$$

This completes the proof of Proposition 5.2.1. □

Our positive result shows that by using a regularized version of the $L_2^2$ loss, we can guarantee that all stationary points have error within a constant multiple of $\epsilon$. Before

stating our formal result, we first define the norm that we will use frequently together with its corresponding dual norm. In fact, we show that this norm characterizes the landscape of the (regularized) $L_2^2$ loss, in the sense that minimizing the gradient with respect to its dual norm will give a point with small error.

**Definition 5.2.2** ($\mathbf{w}$-weighted Euclidean Norm)**.** *Given some vector $\mathbf{u} \in \mathbb{R}^d$, we define its weighted Euclidean norm with respect to a non-zero vector $\mathbf{w} \in \mathbb{R}^d$ to be*

$$\|\mathbf{u}\|_{\mathbf{w}} = \frac{\|\text{proj}_{\mathbf{w}}\mathbf{u}\|_2}{\|\mathbf{w}\|_2^{3/2}} + \frac{\|\text{proj}_{\mathbf{w}^\perp}\mathbf{u}\|_2}{\|\mathbf{w}\|_2^{1/2}} \,.$$

*We also define the dual norm of $\|\cdot\|_{\mathbf{w}}$ as follows:*

$$\|\mathbf{v}\|_{*,\mathbf{w}} = \max(\|\text{proj}_{\mathbf{w}}\mathbf{v}\|_2\|\mathbf{w}\|_2^{3/2}, \|\text{proj}_{\mathbf{w}^\perp}\mathbf{v}\|_2\|\mathbf{w}\|_2^{1/2}) \,.$$

The main intuition behind the norm of Definition 5.2.2 is the following: for well-behaved distributions and sigmoidal activations, the "noiseless" $L_2^2$ loss $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2]$ behaves similarly to the non-convex function $\mathbf{w} \mapsto \|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}^2$.

**Remark 5.2.3** (Distribution/Activation parameters)**.** Before we present our main structural result, we would like to revisit the parameters of Definition 5.1.1 and Definition 5.1.2. We observe that an $(L, R)$-well-behaved distribution is also $(L', R')$-well-behaved for any $L' \leq L, R' \leq R$. Therefore, without loss of generality (and to simplify the presentation), we shall assume that $L, R \in (0, 1]$. For the same reason, for sigmoidal activations, we will assume that $\xi \in [1, \infty)$ and $\tau, \mu \in (0, 1]$. Similarly, for unbounded activations we assume $\lambda \in [1, \infty), \alpha \in (0, 1]$.

We now state our main structural result for sigmoidal activations, namely that all stationary points of the $\ell_2$-regularized $L_2^2$ objective are approximately optimal solutions.

**Theorem 5.2.4** (Stationary Points of Sigmoidal Activations)**.** *Let $\mathcal{D}$ be an $\epsilon$-corrupted, $(L, R)$-well-behaved distribution on $\mathbb{R}^d \times \mathbb{R}$ and $\sigma$ be a $(\tau, \mu, \xi)$-sigmoidal activation. Set $\kappa = L^6 R^6 \mu^3 \tau^4 / \xi^2$ and $\rho = C\epsilon^3/\kappa^5$, where $C > 0$ is a universal constant, and define the $\ell_2$-regularized objective as $F_\rho(\mathbf{w}) = (1/2)\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] + (\rho/2)\|\mathbf{w}\|_2^2$. For some sufficiently small universal constant $c > 0$, we have the following*

- *If $\|\mathbf{w}\|_2 \le 2/R$ and $\|\nabla F_\rho(\mathbf{w})\|_2 \le c\sqrt{\epsilon}$, then $F(\mathbf{w}) = \epsilon \cdot \mathrm{poly}(1/\kappa)$.*
- *If $\|\mathbf{w}\|_2 \ge 2/R$ and $\|\nabla F_\rho(\mathbf{w})\|_{*,\mathbf{w}} \le c\sqrt{\epsilon}$, then $F(\mathbf{w}) = \epsilon \cdot \mathrm{poly}(1/\kappa)$.*

Theorem 5.2.4 shows that the $L_2^2$ objective behaves differently when $\mathbf{w}$ lies close vs far from the origin. The landscape becomes more "flat" as we move further away from the origin, and in order to achieve loss $\epsilon$ we have to make the dual of the weighted $\mathbf{w}$-norm of the gradient small; for example, the component orthogonal to $\mathbf{w}$ has to be smaller than $\sqrt{\epsilon}/\|\mathbf{w}\|_2^{1/2}$.

At a high level, we prove Theorem 5.2.4 in two main steps. First, we analyze the population $L_2^2$ loss without a regularizer, and show that all "bad" stationary points are in fact at distance $\Omega(1/\epsilon)$ from the origin. Since the non-regularized (vanilla) gradient field becomes very flat far from the origin, adding a small amount of noise potentially creates bad local minima (see also Figure 5.2). We show that by adding an appropriate $\ell_2$ regularizer, we essentially introduce radial gradients pulling towards the origin whose contribution is large enough to remove bad stationary points, while ensuring that the "good" stationary points do not change by a lot.

Before proceeding to the proof of Theorem 5.2.4, we first highlight an intricacy of the landscape of the $L_2^2$ loss for sigmoidal activations: the $L_2^2$ loss may be "minimized" for some direction of infinite length. Consider, for example, the case where the activation is the logistic loss $\sigma(t) = 1/(1 + e^{-t})$, the label $y$ is equal to $\mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$ for some unit

vector $\mathbf{w}^*$, and the $\mathbf{x}$-marginal is the standard normal distribution. Then, we have that for any fixed vector $\mathbf{w}$ it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2] > 0$ and

$$\lim_{t\to+\infty} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(t\ \mathbf{w}^*\cdot\mathbf{x})-y)^2] = 0\,.$$

We prove the following lemma showing that even though the true "minimizer" may be a vector of infinite length, there exist almost optimal solutions inside a ball of radius $O(1/\epsilon)$.

**Lemma 5.2.5** (Radius of Approximate Optimality of Sigmoidal Activations)**.** *Let $\mathcal{D}$ be an $(L,R)$-well-behaved distribution in $\mathbb{R}^d$ and let $\sigma(t)$ be a $(\tau,\mu,\xi)$-sigmoidal activation function. There exists a vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 \le 1/\epsilon$ and $F(\mathbf{v}) \le (1+O(\frac{\xi}{\mu L}))\epsilon$.*

*Proof.* We first observe that since $F(\mathbf{w})$ is a continuous function of $\mathbf{w}$, there exists a vector $\mathbf{w}^*$ with $\|\mathbf{w}^*\|_2 < +\infty$ such that $F(\mathbf{w}^*) \le 2\epsilon$. If $\epsilon \ge 1$, then the zero vector achieves $4\epsilon$ error since, using the inequality $(a+b)^2 \le 2a^2+2b^2$, we have that $F(\mathbf{0}) = \mathbf{E}[y^2] \le 2\,\mathbf{E}[(y-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] + 2\,\mathbf{E}[\sigma(\mathbf{w}^*\cdot\mathbf{x})^2] \le 2\epsilon+2\epsilon \le 4\epsilon$. Denote $1/\epsilon := \rho$. If $\|\mathbf{w}^*\|_2 \le \rho$, then we are done. If $\|\mathbf{w}^*\|_2 > \rho$, then we consider a scaled down version of $\mathbf{w}^*$, namely, the vector $\mathbf{v} = \rho\ \mathbf{w}^*/\|\mathbf{w}^*\|_2$. Using again the inequality $(a+b)^2 \le 2a^2+2b^2$, we have

$$F(\mathbf{v}) \le 2F(\mathbf{w}^*) + 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\sigma(\mathbf{v}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2]\,.$$

Since $\mathbf{v}$ is parallel to $\mathbf{w}^*$ and its norm is smaller than $\|\mathbf{w}^*\|$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{v}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] &\leq \frac{1}{L}\int_{-\infty}^{+\infty}(\sigma(\|\mathbf{v}\|_2 t)-\sigma(\|\mathbf{w}^*\|_2|t|)^2 dt \\ &\leq \frac{1}{L}\int_{-\infty}^{+\infty}\left(\int_{\|\mathbf{v}\|t}^{+\infty}\sigma'(z)dz\right)^2 dt \\ &\leq \frac{1}{L}\int_{-\infty}^{+\infty}(\xi e^{-\mu\|\mathbf{v}\||t|})^2 dt \\ &= \frac{\xi}{\mu L\|\mathbf{v}\|},\end{aligned}$$

where we used the fact that the density of 1-dimensional marginals of $\mathcal{D}_\mathbf{x}$ is bounded from above by $1/L$, see Definition 5.1.1, and the fact that $\sigma'(t)\leq\xi e^{-\mu|t|}$. We see that by choosing $\|\mathbf{v}\|_2 = 1/\epsilon$ it holds that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{v}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] = O(\xi/(\mu L))\epsilon$. Thus, combining the above we have that there always exists a $\mathbf{v}$ with $\|\mathbf{v}\|_2 = O(1/\epsilon$ that achieves $L_2^2$ loss $F(\mathbf{v})\leq(1+O(\frac{\xi}{\mu L}))\epsilon$. □

To prove Theorem 5.2.4 we will show that for most vectors $\mathbf{w}$ the gradients of vectors that have large error, i.e., $\mathbf{w}$ with $F(\mathbf{w})\geq\Omega(\epsilon)$, will have large contribution towards some optimal vector $\mathbf{w}^*$, i.e., $\nabla F_\rho(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*)>0$. This is a more useful property than simply proving that the norm of the gradient is large, and we will later use it to obtain an efficient algorithm. However, as we already discussed, this is not always the case: we can only show that the gradient field "pulls" towards some optimal solution only inside a ball of radius roughly $1/\epsilon$ around the origin (see Cases 1,2 of Proposition 5.2.6). Outside of this ball, we show that the regularizer will take effect and pull us back into the ball where the gradient field helps to improve the guess; this corresponds to Case 3 of Proposition 5.2.6. In particular, we show that the projection of the gradient of the $L_2^2$ objective on the direction $\mathbf{w}-\mathbf{w}^*$ is proportional to the standard $\ell_2$ distance of $\mathbf{w}-\mathbf{w}^*$, when $\mathbf{w}$ is close to the origin, and proportional to the $\mathbf{w}$-weighted Euclidean

distance of Definition 5.2.2, when $\mathbf{w}$ is far from the origin.

**Proposition 5.2.6** (Gradient of the Regularized $L_2^2$ Loss)**.** *Let $\mathcal{D}$ be an $(L, R)$-well-behaved distribution and define $F_\rho(\mathbf{w}) = (1/2)\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] + (1/2)\rho\|\mathbf{w}\|_2^2$, where $\sigma$ is a $(\tau, \mu, \xi)$-sigmoidal activation and $\rho > 0$. Let $\epsilon \in (0, 1)$ and let $\mathbf{w}^* \in \mathbb{R}^d$ such that $F(\mathbf{w}^*) \leq \epsilon$ and $\|\mathbf{w}^*\|_2 \leq U/\epsilon$, for some $U \geq 0$. Furthermore, set $\kappa = L^6R^6\mu^3\tau^4/\xi^2$ and $\rho = C\epsilon^3/\kappa^5$, where $C \geq 0$ is a sufficiently large universal constant. There exists a universal constant $c' > 0$, such that for any $\mathbf{w} \in \mathbb{R}^d$, we have:*

1. *When $\|\mathbf{w}\|_2 \leq 2/R$ and $\|\mathbf{w}^* - \mathbf{w}\|_2 \geq \sqrt{\epsilon}/(c'\kappa^5)$, then $\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c'\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_2$.*

2. *When $2/R \leq \|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$ and either $\|\mathbf{w} - \mathbf{w}^*\|_\mathbf{w} \geq \sqrt{\epsilon}(U/(c'\kappa^5))$ or $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}^*\|_2$, then $\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c'\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_\mathbf{w}$.*

3. *When $\|\mathbf{w}\|_2 \geq c'\kappa/(2\epsilon)$, then $\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} \geq c'\sqrt{\epsilon}\|\mathbf{w}\|_\mathbf{w}$.*

Before we prove Proposition 5.2.6, we show that without adding the regularizer the gradient field of the "vanilla" $L_2^2$ loss points towards some optimal solution $\mathbf{w}^*$, as long as the guess $\|\mathbf{w}\|_2$ is not very far from the origin.

**Proposition 5.2.7** (Gradient of the "Vanilla" $L_2^2$ Loss (Inside a Ball))**.** *Let $\mathcal{D}$ be an $(L, R)$-well-behaved distribution and let $\sigma$ be a $(\tau, \mu, \xi)$-sigmoidal activation. Let $\epsilon \in (0, 1)$ be smaller than a sufficiently small multiple of $L^2R^6\tau^4/\xi^2$ and let $\mathbf{w}^* \in \mathbb{R}^d$ with $F(\mathbf{w}^*) \leq \epsilon$. Set $\kappa = L^6R^6\mu^3\tau^4/\xi^2$. There exists a universal constant $c' > 0$, such that for any $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$, we have:*

1. *When $\|\mathbf{w}\|_2 \leq 2/R$ and $\|\mathbf{w}^* - \mathbf{w}\|_2 \geq \sqrt{\epsilon}\xi/(c'LR^4\tau^2)$, then*

$$\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c'\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_2\,.$$

*2. When* $\|\mathbf{w}\|_2 \geq 2/R$ *and either* $\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} \geq \sqrt{\epsilon/\kappa}$ *or* $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}^*\|_2$*, then*

$$\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c'\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} .$$

*Proof of Proposition 5.2.7.* We decompose the gradient into a part that corresponds to the contribution of the "true" labels, i.e., $\sigma(\mathbf{w}^* \cdot \mathbf{x})$ (see $I_2$ below), and a part corresponding to the noise (see $I_1$ below). We have that

$$\begin{aligned}\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) = &\underbrace{\mathbf{E}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})]}_{I_1} \\ &+ \underbrace{\mathbf{E}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - \sigma(\mathbf{w}^* \cdot \mathbf{x}))\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})]}_{I_2} . \end{aligned} \tag{5.1}$$

In the following subsections, we bound $I_1$ and $I_2$ from below. We start by bounding from below the effect of the noise, i.e., the contribution of $I_1$.

**Estimating the Effect of the Noise** We start by showing that the noise cannot affect the gradient by a lot, i.e., we bound the contribution of $I_1$. We prove the following lemma:

**Lemma 5.2.8.** *Let* $\mathcal{D}$ *be an* $(L, R)$*-well-behaved distribution and* $\sigma$ *be a* $(\tau, \mu, \xi)$*-sigmoidal activation. For any vector* $\mathbf{w} \in \mathbb{R}^d$*, it holds that*

$$I_1 \geq -\sqrt{8\xi^2\epsilon}\min(\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}}/(\mu^{3/2}L^2), \|\mathbf{w}^* - \mathbf{w}\|_2) .$$

*Proof.* Using the Cauchy-Schwarz inequality, we obtain:

$$\begin{aligned} I_1 &\geq -\mathbf{E}[|(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})|] \\ &\geq -(\mathbf{E}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2])^{1/2} (\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \sigma'(\mathbf{w} \cdot \mathbf{x})^2])^{1/2} \\ &= -\sqrt{F(\mathbf{w}^*)}\, (\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \sigma'(\mathbf{w} \cdot \mathbf{x})^2])^{1/2} \\ &\geq -\sqrt{\epsilon}\, (\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \sigma'(\mathbf{w} \cdot \mathbf{x})^2])^{1/2}\,, \end{aligned} \tag{5.2}$$

where we used that $F(\mathbf{w}^*) \leq \epsilon$, from the assumptions of Proposition 5.2.7. We proceed to bound the term $\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \sigma'(\mathbf{w} \cdot \mathbf{x})^2]$. Note that we can use the upper bound on the derivative of the activation function, i.e., $\sigma'(t) \leq \xi$ for all $t \in \mathbb{R}$. However, this would result in $\mathbf{E}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \sigma'(\mathbf{w} \cdot \mathbf{x})^2] \leq O(\xi^2 \|\mathbf{w} - \mathbf{w}^*\|_2^2)$. While, this was sufficient for the case of unbounded activations of the Definition 5.1.5, for bounded activation functions, we need a tighter estimate that takes into account the fact that the functions have exponential tails outside the interval $[-1/\mu, +1/\mu]$. Recall that we denote by $\mathbf{q}^{\|\mathbf{w}^*}$ the component of $\mathbf{q} \in \mathbb{R}^d$ parallel to $\mathbf{w}^* \in \mathbb{R}^d$, i.e., $\mathbf{q}^{\|\mathbf{w}^*} = \mathrm{proj}_{\mathbf{w}^*} \mathbf{q}$. Similarly, we denote $\mathbf{q}^{\perp \mathbf{w}^*} = \mathrm{proj}_{\mathbf{w}^{*\perp}} \mathbf{q}$. We prove the following.

**Lemma 5.2.9.** *Let $\mathcal{D}$ be an $(L, R)$-well-behaved distribution and $\sigma$ be a $(\tau, \mu, \xi)$-sigmoidal activation. For any vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, it holds that*

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(\mathbf{w} \cdot \mathbf{x} - \mathbf{v} \cdot \mathbf{x})^2 (\sigma'(\mathbf{w} \cdot \mathbf{x}))^2] \leq 8\xi^2 \min\left( \frac{1}{L^4 \mu^3} \|\mathbf{v} - \mathbf{w}\|_\mathbf{w}^2, \|\mathbf{v} - \mathbf{w}\|_2^2 \right) .$$

*Proof.* First note that $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{v} \cdot \mathbf{x})^2 \sigma'(\mathbf{w} \cdot \mathbf{x})^2] \leq \xi^2 \mathbf{E}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}}[(\mathbf{w} \cdot \mathbf{x} - \mathbf{v} \cdot \mathbf{x})^2 \exp(-2|\mathbf{w} \cdot$

$\mathbf{x}|\mu)]$, from the assumption that $|\sigma'(t)| \leq \xi \exp(-\mu|t|)$. It holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x}-\mathbf{v}\cdot\mathbf{x})^2\exp(-2|\mathbf{w}\cdot\mathbf{x}|\mu)] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x}-\mathbf{v}\cdot\mathbf{x})^2] = \|\mathbf{w}-\mathbf{v}\|_2^2 \,,$$

where we used the fact that the distribution $\mathcal{D}_\mathbf{x}$ is isotropic, i.e., for any vector $\mathbf{u}\in\mathbb{R}^d$, it holds that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{u}\cdot\mathbf{x})^2] = \|\mathbf{u}\|_2^2$.

Next show that by using the tails of the distribution $\mathcal{D}_\mathbf{x}$, we can prove a tighter upper bound for some cases. We use that the distribution is $(L,R)$-well-behaved, and we prove the following claim that bounds from above the expectation.

**Claim 5.2.10.** *Let $\mathcal{D}_\mathbf{x}$ be an $(L,R)$-well-behaved distribution. Let $b>0$ and $\mathbf{u},\mathbf{v}$ be unit norm orthogonal vectors. It holds that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{u}\cdot\mathbf{x})^2\exp(-b|\mathbf{v}\cdot\mathbf{x}|)] \leq \frac{8}{L^4 b} \quad \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{v}\cdot\mathbf{x})^2\exp(-b|\mathbf{v}\cdot\mathbf{x}|)] \leq \frac{8}{L^2 b^3} \,.$$

*Proof.* Without loss of generality take $\mathbf{u}=\mathbf{e}_1$ and $\mathbf{v}=\mathbf{e}_2$. Using the fact that the distribution is $(L,R)$-well-behaved, we have that the 2-dimensional projection on the subspace $V$ spanned by $\mathbf{v},\mathbf{u}$ is bounded from above by $(1/L)\exp(-L\|\mathbf{x}_V\|_2)$ everywhere. We have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x}_1^2\exp(-b\mathbf{v}\cdot\mathbf{x})] &\leq (1/L)\int_{\mathbf{x}_2\in\mathbb{R}}\int_{\mathbf{x}_1\in\mathbb{R}} \mathbf{x}_1^2\exp(-b\mathbf{v}\cdot\mathbf{x})\exp(-L\|\mathbf{x}_V\|_2)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2 \\
&\leq (4/L)\int_0^\infty \mathbf{x}_1^2\exp(-L|\mathbf{x}_1|)\int_0^\infty \exp(-b|\mathbf{x}_2|)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2 \\
&= 8/(L^4 b) \,.
\end{aligned}$$

Putting the above estimates together, we proved the first part of Claim 5.2.10. For the

other part, it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{v}\cdot\mathbf{x})^2\exp(-b\mathbf{v}\cdot\mathbf{x})] \le (4/L)\int_0^\infty\int_0^\infty \mathbf{x}_2^2\exp(-L|\mathbf{x}_1|)\exp(-b|\mathbf{x}_2|)\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{x}_1 = 8/(L^2b^3)\ .$$

This completes the proof of Claim 5.2.10. □

Let $\mathbf{q}=\mathbf{v}-\mathbf{w}$ and note that $(\mathbf{q}\cdot\mathbf{x})^2 = (\mathbf{q}^{\parallel\mathbf{w}}\cdot\mathbf{x})^2 + (\mathbf{q}^{\perp\mathbf{w}}\cdot\mathbf{x})^2$. Using Claim 5.2.10, it holds that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{q}\cdot\mathbf{x})^2\exp(-2\mathbf{w}\cdot\mathbf{x}\mu)] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[((\mathbf{q}^{\parallel\mathbf{w}}\cdot\mathbf{x})^2 + (\mathbf{q}^{\perp\mathbf{w}}\cdot\mathbf{x})^2)\exp(-2\mathbf{w}\cdot\mathbf{x}\mu)]\\ &\le \frac{8}{L^4}(\|\mathbf{q}^{\parallel\mathbf{w}}\|_2^2\frac{1}{\mu^3\|\mathbf{w}\|_2^3} + \|\mathbf{q}^{\perp\mathbf{w}}\|_2^2\frac{1}{\mu\|\mathbf{w}\|_2}) \le \frac{8\|\mathbf{q}\|_\mathbf{w}^2}{L^4\mu^3}\ ,\end{aligned}$$

where we used that $a^2+b^2 \le (a+b)^2$ for $a,b\ge 0$. This completes the proof of Lemma 5.2.9. □

Lemma 5.2.8 follows from combining Equation (5.2) along with Lemma 5.2.9. □

**Estimating the Contribution of the "Noiseless" Gradient** In order to compute the contribution of the gradient when there is no noise to the instance, we show that in fact the contribution of $I_2$ is bounded from below by the contribution of a ramp function instead of $\sigma$. To show this, we use the property that $\sigma'(t)\ge\tau$, for all $t\in[-1,1]$, see Definition 5.1.2.

**Claim 5.2.11.** *It holds that* $I_2 \ge \tau^2\,\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(r(\mathbf{w}\cdot\mathbf{x}) - r(\mathbf{v}\cdot\mathbf{x}))r'(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}-\mathbf{v})\cdot\mathbf{x}]$.

*Proof.* We have that

$$\begin{aligned} I_2 &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))\sigma'(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}-\mathbf{v})\cdot\mathbf{x}] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))\sigma'(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}-\mathbf{v})\cdot\mathbf{x}|] \\ &\geq \tau E_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))(\mathbf{w}-\mathbf{v})\cdot\mathbf{x}|\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}|\leq 1\}] \ , \end{aligned}$$

where we used that $\sigma$ is non-decreasing and that $\sigma(t)\geq\tau$ for $t\in[-1,1]$ from the assumptions of $(\tau,\mu,\xi)$-sigmoidal activations. Moreover, note that from the fundamental theorem of calculus, we have that

$$\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}) = \int_{\mathbf{v}\cdot\mathbf{x}}^{\mathbf{w}\cdot\mathbf{x}}\sigma'(t)\mathrm{d}t \geq \tau\int_{\mathbf{v}\cdot\mathbf{x}}^{\mathbf{w}\cdot\mathbf{x}}\mathbb{1}\{|t|\leq 1\}\mathrm{d}t = \tau(r(\mathbf{w}\cdot\mathbf{x})-r(\mathbf{v}\cdot\mathbf{x})) \ ,$$

where we used again that $\sigma(t)\geq\tau$ for $t\in[-1,1]$. Therefore, we have that

$$I_2 \geq \tau^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(r(\mathbf{w}\cdot\mathbf{x})-r(\mathbf{v}\cdot\mathbf{x}))r'(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}-\mathbf{v})\cdot\mathbf{x}] \ .$$

This completes the proof of Claim 5.2.11. □

To bound $I_2$ from below, we consider three cases depending on how far the vector $\mathbf{w}$ is from the target vector $\mathbf{w}^*$. The first two cases correspond to $\theta(\mathbf{w},\mathbf{w}^*)\in(0,\pi/2)$. In the first case, we have that either $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2\geq 2K/R$, for any $K\geq 1$, or $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}^*\|_2\geq 2\|\mathbf{w}\|_2$ and in the second case when both of the aforementioned conditions are false. The last case corresponds to $\theta(\mathbf{w},\mathbf{w}^*)\in(\pi/2,\pi)$. Notice that when $\mathbf{w}$ is close to the target $\mathbf{w}^*$ then its projection onto the orthogonal complement of $\mathbf{w}$, i.e., $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2$, will be small and also its projection on the direction of $\mathbf{w}$, i.e., $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}^*\|_2$, will be close to $\|\mathbf{w}\|_2$.

If either of the conditions of the first case are satisfied or when $\mathbf{w} \cdot \mathbf{w}^* \leq 0$, then there is a large enough region that we can substitute the $r(\mathbf{w}^* \cdot \mathbf{x})$ by the constant function $r(\mathbf{w}^* \cdot \mathbf{x}) = 1$ (left plot in Figure 5.3), these conditions corresponds to the case that the vectors $\mathbf{w}$ and $\mathbf{w}^*$ are "far" apart from each other. Whereas if both of these conditions are not satisfied then there exists a large enough region that $r(\mathbf{w}^* \cdot \mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x}$ (right plot in Figure 5.3), in this case the vectors $\mathbf{w}$ and $\mathbf{w}^*$ are "close" to each other. Without loss of generality, we may assume that $\mathbf{w}/\|\mathbf{w}\|_2 = \mathbf{e}_2$ and $\mathbf{w}^* = \|\mathbf{w}^*\|_2(\cos\theta \mathbf{e}_2 - \sin\theta \mathbf{e}_1)$. For simplicity, we abuse notation and denote by $\mathcal{D}_{\mathbf{x}}$ the marginal distribution on the subspace spanned by the vectors $\mathbf{w}, \mathbf{w}^*$.

**Case 1: $\mathbf{w}$ and $\mathbf{w}^*$ are "close"** We now handle the case where $\theta(\mathbf{w}, \mathbf{w}^*) \in (0, \pi/2)$, $\|\mathrm{proj}_{\mathbf{w}^\perp} \mathbf{w}^*\|_2 \leq 2K/R$ and $\|\mathrm{proj}_{\mathbf{w}} \mathbf{w}^*\|_2 \leq 2\|\mathbf{w}\|_2$, i.e., the case where $\mathbf{w}$ and $\mathbf{w}^*$ are close to each other but still not close enough to guarantee small $L^2$ error, i.e., the $L_2^2$ error of $\mathbf{w}$, $F(\mathbf{w})$ is much larger than $\epsilon$.

**Lemma 5.2.12.** *Let $\mathcal{D}$ be an $(L, R)$-well-behaved distribution. For any vector $\mathbf{w} \in \mathbb{R}^d$, let $\theta = \theta(\mathbf{w}^*, \mathbf{w}) \in (0, \pi/2)$. If $\|\mathrm{proj}_{\mathbf{w}^\perp} \mathbf{w}^*\|_2 \leq 2K/R$, for $K \geq 1$, and $\|\mathrm{proj}_{\mathbf{w}} \mathbf{w}^*\|_2 \leq 2\|\mathbf{w}\|_2$, it holds:*

- *If $\|\mathbf{w}\|_2 \geq 2/R$, then $I_2 \geq \frac{\tau^2 L R^3}{4096 K^3} \|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}^2$.*
- *Otherwise, $I_2 \geq \frac{\tau^2 L R^4}{2048 K^3} \|\mathbf{w} - \mathbf{w}^*\|_2^2$.*

*Proof.* Using Claim 5.2.11, we have that

$$\begin{aligned} I_2 &\geq \tau^2 \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} [(r(\mathbf{w} \cdot \mathbf{x}) - r(\mathbf{w}^* \cdot \mathbf{x})) r'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} - \mathbf{w}^*) \cdot \mathbf{x}] \\ &\geq \tau^2 \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} [(\mathbf{w} \cdot \mathbf{x} - r(\mathbf{w}^* \cdot \mathbf{x}))(\mathbf{w} - \mathbf{w}^*) \cdot \mathbf{x} \mathbb{1}\{0 \leq \mathbf{w} \cdot \mathbf{x} \leq 1\}] \,. \end{aligned}$$

Note that in the last inequality we used that because $r(\cdot)$ is a non-decreasing function, it holds that $(r(\mathbf{w}\cdot\mathbf{x})-r(\mathbf{w}^*\cdot\mathbf{x}))(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}\geq 0$ for any value of $\mathbf{x}\in\mathbb{R}^d$. Moreover, we assume without loss of generality that $\mathbf{w}/\|\mathbf{w}\|_2=\mathbf{e}_2$, and therefore $\mathbf{w}^*=\|\mathbf{w}^*\|_2(\cos\theta\mathbf{e}_2-\sin\theta\mathbf{e}_1)$. Note that in this case the condition $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}^*\|_2\leq 2\|\mathbf{w}\|_2$ is equivalent to $2\|\mathbf{w}\|_2\geq|\cos\theta|\|\mathbf{w}^*\|_2$.

Consider the region $-R/(4K)\leq\mathbf{x}_1\leq -R/(8K)$ and $\mathbf{x}_2\leq 1/(4\|\mathbf{w}\|_2)$ which is chosen to guarantee that $0\leq\mathbf{w}^*\cdot\mathbf{x}\leq 1$, which holds because

$$0\leq\mathbf{w}^*\cdot\mathbf{x}\leq\|\mathbf{w}^*\|_2\cos\theta\mathbf{x}_2+R\|\mathbf{w}^*\|_2\sin\theta/(4K)\leq 1/2+\cos\theta\|\mathbf{w}^*\|_2/(4\|\mathbf{w}\|_2)\leq 1\ .$$

We show the following claim which will be used to bound from below $I_2$ for this case.

**Claim 5.2.13.** *Let $\mathcal{D}_{\mathbf{x}}$ be an $(L,R)$-well-behaved distribution. For any vectors $\mathbf{w},\mathbf{v}\in\mathbb{R}^d$ and $a,b\in[0,R]$, let $\mathbf{q}=\mathbf{w}-\mathbf{v}$, it holds*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{q}\cdot\mathbf{x})^2\mathbb{1}\left\{\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\cdot\mathbf{x}\in(0,a),\frac{\mathbf{v}^{\mathrm{proj}_{\mathbf{w}}}}{\|\mathbf{v}^{\mathrm{proj}_{\mathbf{w}}}\|_2}\cdot\mathbf{x}\in(-b,0)\right\}\right]\geq\frac{Lab}{8}(\|\mathbf{q}^{\|_{\mathbf{w}}}\|_2^2a^2+\|\mathbf{q}^{\perp_{\mathbf{w}}}\|_2^2b^2)\ .$$

*Proof.* We have that $\mathbf{q}=\mathbf{q}^{\|_{\mathbf{w}}}+\mathbf{q}^{\perp_{\mathbf{w}}}$ and using the Pythagorean theorem we have that $((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2=\|\mathbf{q}^{\|_{\mathbf{w}}}\|_2^2(\mathbf{e}_1\cdot\mathbf{x})^2+\|\mathbf{q}^{\perp_{\mathbf{w}}}\|_2^2(\mathbf{e}_2\cdot\mathbf{x})^2$. Therefore, we have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{e}_1\cdot\mathbf{x}\in(0,a),\mathbf{e}_2\cdot\mathbf{x}\in(-b,0)\}]\\
&\quad\geq\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{e}_1\cdot\mathbf{x}\in(\frac{a}{2},a),\mathbf{e}_2\cdot\mathbf{x}\in(-b,-\frac{b}{2})\}]\\
&\quad\geq\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\|\mathbf{q}^{\|_{\mathbf{w}}}\|_2^2(\mathbf{e}_1\cdot\mathbf{x})^2+\|\mathbf{q}^{\perp_{\mathbf{w}}}\|_2^2(\mathbf{e}_2\cdot\mathbf{x})^2)\mathbb{1}\{\mathbf{e}_1\cdot\mathbf{x}\in(\frac{a}{2},a),\mathbf{e}_2\cdot\mathbf{x}\in(-b,-\frac{b}{2})\}]\\
&\quad\geq\frac{1}{4}(\|\mathbf{q}^{\|_{\mathbf{w}}}\|_2^2a^2+\|\mathbf{q}^{\perp_{\mathbf{w}}}\|_2^2b^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{\mathbf{e}_1\cdot\mathbf{x}\in(\frac{a}{2},a),\mathbf{e}_2\cdot\mathbf{x}\in(-b,-\frac{b}{2}\}]\\
&\quad\geq\frac{Lab}{8}(\|\mathbf{q}^{\|_{\mathbf{w}}}\|_2^2a^2+\|\mathbf{q}^{\perp_{\mathbf{w}}}\|_2^2b^2)\ ,
\end{aligned}$$

where in the last inequality we used that the distribution $\mathcal{D}_{\mathbf{x}}$ is $(L, R)$-well-behaved. $\square$

For the case where $\|\mathbf{w}\|_2 \geq 2/R$. Using the Claim 5.2.13, we have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2\mathbb{1}\{0\leq \mathbf{x}_2 \leq 1/(4\|\mathbf{w}\|_2), -R/(4K)\leq \mathbf{x}_1 \leq -R/(8K)\}] \\
&\qquad \geq \frac{LR}{128K\|\mathbf{w}\|_2}\left(\|\mathbf{q}^{\|_{\mathbf{w}}}\|_2^2\frac{1}{16\|\mathbf{w}\|_2^2} + \|\mathbf{q}^{\perp_{\mathbf{w}}}\|_2^2\frac{R^2}{K^2 16}\right) \\
&\qquad \geq \frac{LR^3}{K^3 2048}\left(\frac{\|\mathbf{q}^{\|_{\mathbf{w}}}\|_2^2}{\|\mathbf{w}\|_2^3} + \frac{\|\mathbf{q}^{\perp_{\mathbf{w}}}\|_2^2}{\|\mathbf{w}\|_2}\right) \geq \frac{LR^3}{4096K^3}\|\mathbf{q}\|_{\mathbf{w}}^2 \,.
\end{aligned}$$

For the case where $\|\mathbf{w}\|_2 \leq 2/R$, using Claim 5.2.13, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2\mathbb{1}\{0\leq \mathbf{x}_2 \leq 1/(4\|\mathbf{w}\|_2), -R/(4K)\leq \mathbf{x}_1 \leq -R/(8K)\}] \\
&\quad \geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[((\mathbf{w}-\mathbf{v})\cdot\mathbf{x})^2\mathbb{1}\{R/8\leq \mathbf{x}_2 \leq R/4, -R/(4K)\leq \mathbf{x}_1 \leq -R/(8K)\}] \\
&\quad \geq \frac{LR^2}{K128}\left(\|\mathbf{q}^{\|_{\mathbf{w}}}\|_2^2\frac{R^2}{16} + \|\mathbf{q}^{\perp_{\mathbf{w}}}\|_2^2\frac{R^2}{K^2 16}\right) \geq \frac{LR^4}{K^3 2048}\|\mathbf{q}\|_2^2 \,.
\end{aligned}$$

This completes the proof of Lemma 5.2.12. $\square$

From Lemma 5.2.8, we have that $I_1 \geq -\sqrt{8\epsilon/\mu^3}(\xi/L^2)\|\mathbf{w}^*-\mathbf{w}\|_{\mathbf{w}}$. Using Lemma 5.2.12, for $K=1$, we have that

$$\begin{aligned}
I_1 + I_2 &\geq \frac{\tau^2 LR^3}{4096}\|\mathbf{w}^*-\mathbf{w}\|_{\mathbf{w}}\left(\|\mathbf{w}^*-\mathbf{w}\|_{\mathbf{w}} - \sqrt{\epsilon}\frac{4096\sqrt{8}\xi}{L^3\mu^{3/2}R^3\tau^2}\right) \\
&\geq 2(\xi/\mu^{3/2})\sqrt{8\epsilon}\|\mathbf{w}^*-\mathbf{w}\|_{\mathbf{w}} \geq \sqrt{\epsilon}\|\mathbf{w}^*-\mathbf{w}\|_{\mathbf{w}} \,,
\end{aligned}$$

where in the second inequality we used that $\|\mathbf{w}^*-\mathbf{w}\|_{\mathbf{w}} \geq \sqrt{\epsilon}\frac{8192\sqrt{8}\xi}{L^3\mu^{3/2}R^3\tau^2}$ and that $\xi \geq 1, \mu \leq 1$. Moreover, note that in the special case that $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}^*\|_2$, we have that

$\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} \geq (1/2)\|\mathbf{w}\|_2^{-1/2}$. Therefore, it holds

$$\begin{aligned} I_1 + I_2 &\geq \frac{\tau^2 LR^3}{4096}\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} \left( \|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} - \sqrt{\epsilon}\frac{4096\sqrt{8}\xi}{L^3\mu^{3/2}R^3\tau^2} \right) \\ &\geq \frac{\tau^2 LR^3}{4096}\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} \left( (1/2)\|\mathbf{w}\|_2^{-1/2} - \sqrt{\epsilon}\frac{4096\sqrt{8}\xi}{L^3\mu^{3/2}R^3\tau^2} \right) \\ &\geq \frac{\tau^2 LR^3}{16384}\frac{\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}}}{\|\mathbf{w}\|_2^{1/2}} , \end{aligned}$$

where in the third inequality we used that $\epsilon\|\mathbf{w}\|_2\xi^2/(L^6R^6\mu^3\tau^2)$ is less than a sufficient small constant. Using that $\|\mathbf{w}\|_2 \leq CL^6R^6\mu^3\tau^4/(\epsilon\xi^2)$ for a sufficiently small constant $C > 0$, from the assumptions of Proposition 5.2.7, we get that $I_1 + I_2 \geq c'\frac{\xi}{\mu^{3/2}}\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} \geq c'\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}}$, where we used that $\xi \geq 1, 1 \geq \mu$ by assumption.

Moreover, for the case that $\|\mathbf{w}\|_2 \leq 2/R$, we have $I_1 \geq -\sqrt{\epsilon}\xi\|\mathbf{w}^* - \mathbf{w}\|_2$. Using Lemma 5.2.12, we have that

$$\begin{aligned} I_1 + I_2 &\geq \frac{\tau^2 LR^4}{2048}\|\mathbf{w}^* - \mathbf{w}\|_2 \left( \|\mathbf{w}^* - \mathbf{w}\|_2 - \sqrt{\epsilon}\frac{2048\xi}{LR^4\tau^2} \right) \\ &\geq \sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_2 , \end{aligned}$$

where in the last inequality we used that $\|\mathbf{w}^* - \mathbf{w}\|_2 \geq \sqrt{\epsilon}\frac{4096\xi}{LR^4\tau^2}$ and that $\xi \geq 1$.

**Case 2: $\mathbf{w}$ is far from the target $\mathbf{w}^*$** We now handle the case where our current guess $\mathbf{w}$ is far from the target weight vector $\mathbf{v}$. In particular, we assume that either $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 \geq 2K/R$ for some $K \geq 1$, or $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}^*\|_2 \geq 2\|\mathbf{w}\|_2$. In this case there is a large enough region for which we can substitute the $r(\mathbf{w}^* \cdot \mathbf{x})$ by the constant function $r(\mathbf{w}^* \cdot \mathbf{x}) = 1$ (left plot in Figure 5.3). First, we handle the case where $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 \geq 2K/R$ or $\|\mathrm{proj}_{\mathbf{w}}\mathbf{w}^*\|_2 \geq 2\|\mathbf{w}\|_2$. We prove the following:

**Lemma 5.2.14.** *Let $\mathcal{D}$ be an $(L, R)$-well-behaved distribution. Let $\theta = \theta(\mathbf{w}^*, \mathbf{w}) \in (0, \pi/2)$. There exists a sufficiently small universal constant $c' > 0$ such that for any $\mathbf{w} \in \mathbb{R}^d$, if $\|\mathrm{proj}_{\mathbf{w}^\perp} \mathbf{w}^*\|_2 \geq 2K/R$, for $K \geq 1$, or $\|\mathrm{proj}_{\mathbf{w}} \mathbf{w}^*\|_2 \geq 2\|\mathbf{w}\|_2$, it holds:*

- *If $\|\mathbf{w}\|_2 \geq 2/R$, then $I_2 \geq \frac{\tau^2 L R^2}{72 \|\mathbf{w}\|_2^{1/2}} \|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}$.*
- *Otherwise, $I_2 \geq \frac{\tau^2 L R^3}{144} \|\mathbf{w} - \mathbf{w}^*\|_2$.*

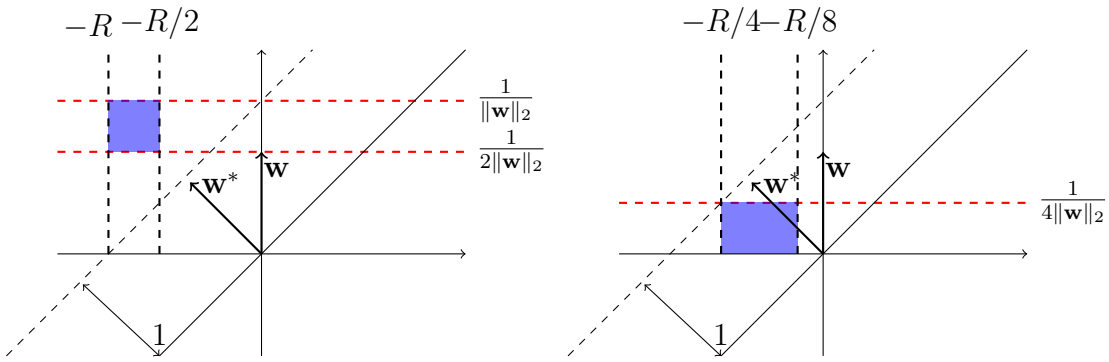



Figure 5.3: Using our distributional assumptions, there exists a region ("blue") that provides enough contribution to the gradient, the left plot corresponds to Lemma 5.2.14 where the angle between the current vector and the target one is large, in this case we take a region where $r(\mathbf{w}^* \cdot \mathbf{x}) = 1$, the right plot corresponds to Lemma 5.2.12, in this case the angle is small, and we take a region where $r(\mathbf{w}^* \cdot \mathbf{x}) = \mathbf{w}^* \cdot \mathbf{x}$.

*Proof.* From Claim 5.2.11, we have that

$$
\begin{aligned}
I_2 &\geq \tau^2 \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} [(r(\mathbf{w} \cdot \mathbf{x}) - r(\mathbf{w}^* \cdot \mathbf{x})) r'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} - \mathbf{w}^*) \cdot \mathbf{x}] \\
&= \tau^2 \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} [(\mathbf{w} \cdot \mathbf{x} - r(\mathbf{w}^* \cdot \mathbf{x}))(\mathbf{w} - \mathbf{w}^*) \cdot \mathbf{x} \, \mathbb{1}\{|\mathbf{w} \cdot \mathbf{x}| \leq 1\}] \\
&\geq \tau^2 \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} [(\mathbf{w} \cdot \mathbf{x} - r(\mathbf{w}^* \cdot \mathbf{x}))(\mathbf{w} - \mathbf{w}^*) \cdot \mathbf{x} \, \mathbb{1}\{0 \leq \mathbf{w} \cdot \mathbf{x} \leq 1\}] \, .
\end{aligned}
$$

Note that the last inequality holds because $r$ is a non-decreasing function, and thus we have that $(r(\mathbf{w} \cdot \mathbf{x}) - r(\mathbf{w}^* \cdot \mathbf{x}))(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$. Let $c = \min(1/\|\mathbf{w}\|_2, R)$,

and note that the condition $\|\text{proj}_{\mathbf{w}}\mathbf{w}^*\|_2 \geq 2\|\mathbf{w}\|_2$ is equivalent to $2\|\mathbf{w}\| \leq \cos\theta\|\mathbf{w}^*\|_2$. We consider the following subset $-R \leq \mathbf{x}_1 \leq -R/2$ and $c/2 \leq \mathbf{x}_2 \leq c$ which is chosen such $\mathbf{w}^* \cdot \mathbf{x} \geq 1$; which holds because $\mathbf{w}^* \cdot \mathbf{x} \geq \|\mathbf{w}^*\|_2(\cos\theta\mathbf{x}_2 + R\sin\theta/2) \geq \|\mathbf{w}^*\|_2(\cos\theta c/2 + R\sin\theta/2)$ and the last part is always greater than 1 if at least one of the following hold: $\|\mathbf{w}^*\|_2 \sin\theta \geq K/R$ or $2\|\mathbf{w}\| \leq \cos\theta\|\mathbf{w}^*\|_2$. Therefore, it holds that $r(\mathbf{w}^* \cdot \mathbf{x}) = 1$. Hence, we can write

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w}^*\cdot\mathbf{x}))(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}\ \mathbb{1}\{0 \leq \mathbf{w}\cdot\mathbf{x} \leq 1\}] \\
&\quad\geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w}^*\cdot\mathbf{x}))(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}\ \mathbb{1}\{c/2 \leq \mathbf{x}_2 \leq c, -R \leq \mathbf{x}_1 \leq -R/2\}] \\
&\quad\geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - 1)(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}\ \mathbb{1}\{c/2 \leq \mathbf{x}_2 \leq 2c/3, -R \leq \mathbf{x}_1 \leq -R/2\}]\ .
\end{aligned}
$$

Notice that in this case $\mathbf{w}\cdot\mathbf{x} \leq \mathbf{w}^*\cdot\mathbf{x}$ and that $(1 - \mathbf{w}\cdot\mathbf{x}) \geq 1/3$. Therefore,

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w}^*\cdot\mathbf{x}))(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}\ \mathbb{1}\{0 \leq \mathbf{w}\cdot\mathbf{x} \leq 1\}] \\
&\quad\geq (1/3)\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[|(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}|\mathbb{1}\{c/2 \leq \mathbf{x}_2 \leq 2c/3, -R \leq \mathbf{x}_1 \leq -R/2\}]\ .
\end{aligned}
$$

Let $\mathbf{q} = \mathbf{w} - \mathbf{w}^*$. By using the inequality $(a^2 + b^2)^{1/2} \geq (1/2)(|a| + |b|)$ for any $a, b \in \mathbb{R}$

and that $(\mathbf{q}\cdot\mathbf{x})^2 = (\mathbf{q}^{\parallel_\mathbf{w}}\cdot\mathbf{x})^2 + (\mathbf{q}^{\perp_\mathbf{w}}\cdot\mathbf{x})^2$, we have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}|\mathbb{1}\{c/2\le\mathbf{x}_2\le 2c/3, -R\le\mathbf{x}_1\le -R/2\}] \\
&\qquad = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\sqrt{(\mathbf{q}\cdot\mathbf{x})^2}\mathbb{1}\{c/2\le\mathbf{x}_2\le 2c/3, -R\le\mathbf{x}_1\le -R/2\}] \\
&\qquad \ge \frac{1}{2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(|\mathbf{q}^{\parallel_\mathbf{w}}\cdot\mathbf{x}| + |\mathbf{q}^{\perp_\mathbf{w}}\cdot\mathbf{x}|)\mathbb{1}\{c/2\le\mathbf{x}_2\le 2c/3, -R\le\mathbf{x}_1\le -R/2\}] \\
&\qquad \ge \frac{1}{2}\left(\|\mathbf{q}^{\parallel_\mathbf{w}}\|_2\frac{c}{2} + \|\mathbf{q}^{\perp_\mathbf{w}}\|_2\frac{R}{2}\right)\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{c/2\le\mathbf{x}_2\le 2c/3, -R\le\mathbf{x}_1\le -R/2\}] \\
&\qquad \ge \frac{cLR}{24}\left(\|\mathbf{q}^{\parallel_\mathbf{w}}\|_2\frac{c}{2} + \|\mathbf{q}^{\perp_\mathbf{w}}\|_2\frac{R}{2}\right) ,
\end{aligned}$$

where we used that the distribution is $(L, R)$-well-behaved. For the case that $c = 1/\|\mathbf{w}\|_2$, we have that

$$I_2 \ge \frac{\tau^2 LR}{72\|\mathbf{w}\|_2}\left(\|\mathbf{q}^{\parallel_\mathbf{w}}\|_2\frac{1}{2\|\mathbf{w}\|_2} + \|\mathbf{q}^{\perp_\mathbf{w}}\|_2\frac{R}{2}\right) \ge \frac{\tau^2 LR^2}{72\|\mathbf{w}\|_2^{1/2}}\|\mathbf{q}\|_\mathbf{w} .$$

Finally, for the case that $c = R$, we have that

$$I_2 \ge \frac{\tau^2 LR^2}{72}\left(\|\mathbf{q}^{\parallel_\mathbf{w}}\|_2\frac{R}{2} + \|\mathbf{q}^{\perp_\mathbf{w}}\|_2\frac{R}{2}\right) \ge \frac{\tau^2 LR^3}{144}\|\mathbf{q}\|_2 ,$$

where we used that $(a+b)^2 \ge a^2 + b^2$ for $a, b \ge 0$. This completes the proof of Lemma 5.2.14. □

To prove Proposition 5.2.7 for this case (the case for which the assumptions of Lemma 5.2.14 hold), we use the bound of the contribution of the noise to the gradient from Lemma 5.2.8, i.e., that $I_1 \ge -\sqrt{8\epsilon/\mu^3}(\xi/L^2)\|\mathbf{w}^* - \mathbf{w}\|_\mathbf{w}$, given that $\|\mathbf{w}\|_2 \ge 2/R$.

Putting together our estimates for $I_1$ and $I_2$, we obtain:

$$\begin{aligned} I_1 + I_2 &\geq \frac{\tau^2 LR^2}{72\|\mathbf{w}\|_2^{1/2}}\|\mathbf{q}\|_{\mathbf{w}} \left(1 - \sqrt{\epsilon\|\mathbf{w}\|_2}\frac{\xi 72\sqrt{8}}{L^3R^2\mu^{3/2}\tau^2}\right) \\ &\geq \frac{\tau^2 LR^2}{144\|\mathbf{w}\|_2^{1/2}}\|\mathbf{q}\|_{\mathbf{w}} \,, \end{aligned} \tag{5.3}$$

where, we used that $\sqrt{\epsilon\|\mathbf{w}\|_2}\xi^2/(L^3R^2\mu^{3/2}\tau^2)$ is less than a sufficiently small constant from the assumptions of Proposition 5.2.7. Observe that from the assumptions of Proposition 5.2.7, there exists a constant $C > 0$,such that $\|\mathbf{w}\|_2 \leq CL^6R^6\mu^3\tau^4/(\epsilon\xi^2)$ hence, we have that

$$I_1 + I_2 \geq \sqrt{\epsilon}\|\mathbf{q}\|_{\mathbf{w}} \,,$$

where we used that $L, R, \mu \leq 1$ and $\xi \geq 1$.

Finally, similar to the previous case, for the case that $\|\mathbf{w}\|_2 \leq 2/R$, it holds

$$I_1 + I_2 \geq \frac{\tau^2 LR^3}{144}\|\mathbf{q}\|_2 \left(1 - \sqrt{\epsilon}\frac{144\xi}{LR^3\tau^2}\right) \geq \frac{\tau^2 LR^3}{288}\|\mathbf{q}\|_2 \,,$$

where we used that $(1 - \sqrt{\epsilon}\frac{144\xi}{LR^3\tau^2}) \geq 1/2$.

**Case 3: angle of $\mathbf{w}$ and $\mathbf{w}^*$ is greater than $\pi/2$** We now handle the case where $\theta(\mathbf{w}, \mathbf{w}^*) \in (\pi/2, \pi)$. This case is very similar to the second case, i.e., Lemma 5.2.14.

**Lemma 5.2.15.** *Let $\mathcal{D}$ be an $(L, R)$-well-behaved distribution. For any vector $\mathbf{w} \in \mathbb{R}^d$ with $\theta(\mathbf{w}^*, \mathbf{w}) \in (\pi/2, \pi)$. We show*

- *If $\|\mathbf{w}\|_2 \geq 2/R$, then $I_2 \geq \frac{\tau^2 LR^2}{16\|\mathbf{w}\|_2^{1/2}}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}$.*
- *Otherwise, we have that $I_2 \geq \frac{\tau^2 LR^4}{16}\|\mathbf{w} - \mathbf{w}^*\|_2$.*

*Proof.* Let $c = \min(1/\|\mathbf{w}\|_2, R)$. We consider the $R/2 \leq \mathbf{x}_1 \leq R$ and $c/2 \leq \mathbf{x}_2 \leq c$ so that $0 \geq \mathbf{w}^* \cdot \mathbf{x}$. Similarly to the previous case, using Claim 5.2.11, we get that $I_2 \geq \tau^2 \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w}^*\cdot\mathbf{x}))(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}\ \mathbb{1}\{0 \leq \mathbf{w}\cdot\mathbf{x} \leq 1\}]$ and for this case it holds that $r(\mathbf{w}^*\cdot\mathbf{x}) \leq 0$. Hence, we have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w}^*\cdot\mathbf{x}))(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}\ \mathbb{1}\{0 \leq \mathbf{w}\cdot\mathbf{x} \leq 1\}] \\
&\quad \geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x} - r(\mathbf{w}^*\cdot\mathbf{x}))(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}\ \mathbb{1}\{c/2 \leq \mathbf{x}_2 \leq c, R/2 \leq \mathbf{x}_1 \leq R\}] \\
&\quad \geq \frac{c\|\mathbf{w}\|_2}{2} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x}\ \mathbb{1}\{c/2 \leq \mathbf{x}_2 \leq c, R/2 \leq \mathbf{x}_1 \leq R\}]\ .
\end{aligned}$$

Note that $(\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x} = (\|\mathbf{w}\|_2 - \cos\theta\|\mathbf{w}^*\|_2)\mathbf{x}_2 + \sin\theta\|\mathbf{w}^*\|_2\mathbf{x}_1 \geq (\|\mathbf{w}\|_2 - \cos\theta\|\mathbf{w}^*\|_2)c/2 + \sin\theta\|\mathbf{w}^*\|_2 R/2$, when $R/2 \leq \mathbf{x}_1 \leq R$ and $c/2 \leq \mathbf{x}_2 \leq c$. Denote by $\mathbf{q} = \mathbf{w} - \mathbf{w}^*$ we get that $(\|\mathbf{w}\|_2 - \cos\theta\|\mathbf{w}^*\|_2)c/2 + \sin\theta\|\mathbf{w}^*\|_2 R/2 = \|\mathbf{q}^{\|\mathbf{w}}\|_2 c/2 + \|\mathbf{q}^{\perp\mathbf{w}}\|_2 R/2$. Furthermore, using that the distribution $\mathcal{D}$ is $(L,R)$-well-behaved, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{c/2 \leq \mathbf{x}_2 \leq c, R/2 \leq \mathbf{x}_1 \leq R\}] \geq L\frac{Rc}{4}$. Hence,

$$I_2 \geq \frac{\tau^2 LRc^2\|\mathbf{w}\|_2}{16}\left(\|\mathbf{q}^{\|\mathbf{w}}\|_2 c + \|\mathbf{q}^{\perp\mathbf{w}}\|_2 R\right)\ ,$$

which completes the proof of Lemma 5.2.15. □

Therefore, if $\|\mathbf{w}\|_2 \geq 2/R$, we have $I_1 \geq -\sqrt{8\epsilon/\mu^3}(\xi/L^2)\|\mathbf{w}^* - \mathbf{w}\|_\mathbf{w}$. Putting together our estimates for $I_1$ and $I_2$, we obtain:

$$\begin{aligned}
I_1 + I_2 &\geq \frac{\tau^2 LR^2}{16\|\mathbf{w}\|_2^{1/2}}\|\mathbf{w}-\mathbf{w}^*\|_\mathbf{w}\left(1 - \sqrt{\epsilon\|\mathbf{w}\|_2}\frac{\xi 16}{L^3R^2\mu^{3/2}\tau^2}\right) \\
&\geq \frac{\tau^2 LR^2}{32\|\mathbf{w}\|_2^{1/2}}\|\mathbf{w}-\mathbf{w}^*\|_\mathbf{w}\ ,
\end{aligned}$$

where we used that $\sqrt{\epsilon\|\mathbf{w}\|_2}\xi/(L^3R^2\mu^{3/2}\tau^2)$ is less than a sufficiently small constant from the assumptions of Proposition 5.2.7. Note that from the assumptions of Proposition 5.2.7, there exists a constant $C > 0$ such that $\|\mathbf{w}\|_2 \leq CL^6R^6\mu^3\tau^4/(\epsilon\xi^2)$, and hence it holds

$$I_1 + I_2 \geq \sqrt{\epsilon}\|\mathbf{q}\|_{\mathbf{w}} \,,$$

where we used that $L, R, \mu \leq 1$ and $\xi \geq 1$.

Finally, similarly to the previous case, we have that if $\|\mathbf{w}\|_2 \leq 2/R$, it holds that

$$I_1 + I_2 \geq \frac{\tau^2 LR^4}{16}\|\mathbf{w} - \mathbf{w}^*\|_2 \left(1 - \sqrt{\epsilon}\frac{16\xi}{LR^4\tau^2}\right) \geq \frac{LR^4\tau^2}{32}\|\mathbf{w} - \mathbf{w}^*\|_2 \,,$$

where we used that $(1 - \sqrt{\epsilon}\frac{16\xi}{LR^4\tau^2}) \geq 1/2$. This completes the proof of Proposition 5.2.7.

□

### 5.2.1 Proof of Proposition 5.2.6

Similarly to the statement of Proposition 5.2.6, in our proof we shall consider several cases depending on how large $\|\mathbf{w}\|_2$ is. For the rest of the proof, let $c'$ be the absolute constant of Proposition 5.2.7 and denote by $\kappa = L^3R^6\mu^3\tau^4/\xi^2$, $\Lambda_1 = 16$, $\Lambda_2 = 1/\kappa \geq \xi^{6/5}/(c'\kappa L^2\mu^6)^{1/5}$ and $K = \Lambda_2^{2.5}U/(\sqrt{c'\kappa}LR)$.

First, we show Case 3 of Proposition 5.2.6 and bound from below the contribution of the gradient on the direction of $\mathbf{w}$, when the norm of $\mathbf{w}$ is large. In other words, we show that in this case the gradient contribution of the regularizer is large enough so that the gradient field of the regularized $L_2^2$-objective "pulls" $\mathbf{w}$ towards the origin.

**Claim 5.2.16.** *If* $\|\mathbf{w}\|_2 \geq c'\kappa/(\epsilon\Lambda_1)$, *then* $\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} \geq \sqrt{\epsilon}\|\mathbf{w}\|_{\mathbf{w}}$.

*Proof.* We have that

$$\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{w}\cdot\mathbf{x}] + \rho\|\mathbf{w}\|_2^2 \,.$$

First we bound from below the quantity $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{w}\cdot\mathbf{x}]$. We need to bound the maximum value that $\sigma$ can obtain.

**Fact 5.2.17.** *Let $\sigma$ be a $(\tau, \mu, \xi)$-sigmoidal activation, then $|\sigma(t_1) - \sigma(t_2)| \le 2\xi/\mu$ for any $t_1, t_2 \in \mathbb{R}$.*

*Proof.* Using the fundamental theorem of calculus, for any $t_1, t_2 \in \mathbb{R}$, it holds

$$|\sigma(t_1) - \sigma(t_2)| = \left|\int_{t_1}^{t_2} \sigma'(t)\mathrm{d}t\right| \le \int_{-\infty}^{\infty} \sigma'(t)\mathrm{d}t \le 2\int_0^{\infty} \xi e^{-t\mu}\mathrm{d}t \le 2\xi/\mu \,,$$

where we used that $\sigma$ is non-decreasing and that $\sigma'(t) \le \xi e^{-t\mu}$. □

Using Fact 5.2.17, we have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{w}\cdot\mathbf{x}] &\ge -\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)||\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{w}\cdot\mathbf{x}|] \\
&\ge -2\frac{\xi}{\mu}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\sigma'(\mathbf{w}\cdot\mathbf{x})^2|\mathbf{w}\cdot\mathbf{x}|] \\
&\ge -2\frac{\xi^3}{\mu}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\exp(-2\mu|\mathbf{w}\cdot\mathbf{x}| \le 1)|\mathbf{w}\cdot\mathbf{x}|] \\
&\ge -\frac{\xi^3}{L\mu^3\|\mathbf{w}\|_2} \,,
\end{aligned}$$

where in the second inequality we used that the maximal difference of $(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)$ is less than $2\xi/\mu$, and in the third the upper bound in the derivative of $\sigma$. Therefore, we

have that

$$\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} \geq (1/2)\rho\|\mathbf{w}\|_2^2 + (1/2)\|\mathbf{w}\|_2^2 \left(\rho - \frac{2\xi^3}{L\mu^3\|\mathbf{w}\|_2^3}\right)$$
$$\geq (1/2)\rho\|\mathbf{w}\|_2^2 \,,$$

where in the last inequality we used that $\rho \geq \frac{2\epsilon^3\Lambda_1^3\xi^3}{c'^3\kappa^3\mu^3 L} \geq \frac{2\xi^3}{L\mu^3\|\mathbf{w}\|_2^3}$.

Therefore, we have that $\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} \geq (1/2)\rho\|\mathbf{w}\|_2^{2.5}\|\mathbf{w}\|_{\mathbf{w}}$. By using that $\|\mathbf{w}\|_2 \geq c'\kappa/(\epsilon\Lambda_1)$ and $\rho \geq \frac{2\epsilon^3\Lambda_1^3\xi^3}{c'^3\kappa^3\mu^3 L}$, we get that

$$\nabla F_\rho(\mathbf{w}) \cdot \mathbf{w} \geq \sqrt{\epsilon\Lambda_1}\frac{\xi^3}{\mu^3\sqrt{c'\kappa}L}\|\mathbf{w}\|_{\mathbf{w}} \geq \sqrt{\epsilon}\|\mathbf{w}\|_{\mathbf{w}} \,,$$

where we used that $\xi, \Lambda_1 \geq 1, \mu, \kappa, L \leq 1$ (see Remark 5.2.3) and this completes the proof of Claim 5.2.16.

□

Next we bound from below the contribution of the gradient in the direction $\mathbf{w} - \mathbf{w}^*$ when $\mathbf{w}$ is close to the origin, i.e., when $\|\mathbf{w}\|_2 \leq 2/R$ (first case of Proposition 5.2.6). In this case, we show that our choice of regularization $\rho$, ensures that when $\mathbf{w}$ is not very large, then the gradient behaves qualitatively similarly to that of the vanilla $L_2^2$ objective, i.e., as in Proposition 5.2.7.

**Claim 5.2.18.** *If* $\|\mathbf{w}\|_2 \leq 2/R$ *and* $\|\mathbf{w}^* - \mathbf{w}\|_2 \geq \sqrt{\epsilon}\xi/(c'LR^4\tau^2)$*, then*

$$\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq (c'/2)\xi\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_2 \,.$$

*Proof.* From Proposition 5.2.7, we have that if $\|\mathbf{w}\|_2 \leq 2/R$ and $\|\mathbf{w}^* - \mathbf{w}\|_2 \geq$

$\sqrt{\epsilon}\xi/(c'LR^4\tau^2)$, then $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c'\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_2$. Therefore, we have that

$$\begin{aligned}\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) &\geq c'\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_2 + \rho\|\mathbf{w}\|_2(\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2 \cos\theta) \\ &\geq c'\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_2 - \rho\|\mathbf{w}\|_2\|\mathbf{w}^* - \mathbf{w}\|_2 \\ &\geq (c'/2)\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_2 \,,\end{aligned}$$

where in the third inequality we used that $\|\mathbf{w}\|_2 \leq 2/R$ and that $\rho = O(F^3(\mathbf{w}^*))$. □

Next we handle the case where $2/R \leq \|\mathbf{w}\|_2 \leq c'\kappa/(\epsilon\Lambda_2)$ (see Case 2 of Proposition 5.2.6). Again, the fact that we choose a small regularization parameter allows us to show that the gradient field in this case behaves similarly to that of the unregularized $L_2^2$ objective (see Case 2 of Proposition 5.2.7).

**Claim 5.2.19.** *If* $2/R \leq \|\mathbf{w}\|_2 \leq c'\kappa/(\epsilon\Lambda_2)$, *then*

$$\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq (\sqrt{c'}/2)\sqrt{\epsilon}\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} \,.$$

*Proof.* We compute the contribution of the regularizer in the direction $\mathbf{w} - \mathbf{w}^*$. This is equal to $\rho\|\mathbf{w}\|_2(\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2 \cos\theta)$. Note that this is positive when $\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2 \cos\theta \geq 0$ and negative otherwise. Hence, if $\theta \in (\pi/2, \pi)$ the contribution of the regularizer is positive, and therefore it is bounded from below by the contribution of the gradient without the regularizer. We need to choose the value of $\rho$ so that the regularizer cancels out the contribution of the noise when $\|\mathbf{w}\|_2$ is large (i.e., when the regularizer has positive contribution).

From Proposition 5.2.7, we have that if $\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} \geq \sqrt{\epsilon/\kappa}$, then $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq$

$c'\sqrt{\epsilon}\|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{w}}$, hence, it holds

$$\begin{aligned}\nabla F_\rho(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) &\geq c'\sqrt{\epsilon}\|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{w}} + \rho\|\mathbf{w}\|_2(\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta)\\ &\geq c'\sqrt{\epsilon}\|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{w}} - \rho\|\mathbf{w}\|_2|\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta|\\ &\geq \frac{\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta|}{\|\mathbf{w}\|_2^{3/2}}(c'\sqrt{\epsilon} - \rho\|\mathbf{w}\|_2^{2.5}) + c'\sqrt{\epsilon}\frac{\|\mathbf{w}^*\|_2\sin\theta}{\|\mathbf{w}\|_2}\\ &\geq (c'/2)\sqrt{\epsilon}\|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{w}}\,,\end{aligned}$$

where for the last inequality we used that $\|\mathbf{w}\|_2 \leq c'\kappa/(\epsilon\Lambda_2)$ and that $\rho \leq \frac{\Lambda_2^{2.5}\epsilon^3}{4c'^{1.5}\kappa^{2.5}}$.

□

Finally, we consider the case where $c'\kappa/(\epsilon\Lambda_2) \leq \|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$ and $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 \geq 2K/R$. We show that the contribution in the direction $\mathbf{w}-\mathbf{w}^*$ of the unregularized gradient of $L_2^2$ is greater than the contribution of the gradiend corresponding to the regularizer.

**Claim 5.2.20.** *If $c'\kappa/(\epsilon\Lambda_2) \leq \|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$ and $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 \geq 2K/R$ then*

$$\nabla F_\rho(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \geq (c'/2)\sqrt{\epsilon}\|\mathbf{w}^*-\mathbf{w}\|_{\mathbf{w}}\,.$$

*Proof.* First, we calculate the contribution of the regularizer in the direction $\mathbf{w}-\mathbf{w}^*$. This is equal to $2\rho\|\mathbf{w}\|_2^2(\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta)$. Note that this is positive when $\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta \geq 0$ and negative otherwise. Hence, if $\theta \in [\pi/2, \pi)$ the contribution of the regularizer is positive, and therefore it is bounded below by the contribution of the gradient without the regularizer. Moreover, if $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}^*\|_2$, then $\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2\cos\theta \geq \|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2 \geq 0$, therefore we can again bound from below the contribution of the gradient like before. For the rest of the proof, we consider the case

where $\theta \in (0, \pi/2)$ and $\|\mathbf{w}\|_2 \leq 2\|\mathbf{w}^*\|_2$.

From Proposition 5.2.7 and specifically Equation (5.3), we have that as long as $2/R \leq \|\mathbf{w}\|_2 \leq c'\kappa/(\epsilon$ and $\|\text{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 \geq 2K/R$ then $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq c' \frac{\tau^2 LR^2}{\|\mathbf{w}\|_2^{1/2}} \|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}$. Therefore, we have that

$$\begin{aligned}
\nabla F_\rho(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) &\geq c'\tau^2 LR^2 \frac{\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}}}{\|\mathbf{w}\|_2^{1/2}} + \rho\|\mathbf{w}\|_2(\|\mathbf{w}\|_2 - \|\mathbf{w}^*\|_2 \cos\theta) \\
&\geq c'\tau^2 LR^2 \frac{\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}}}{\|\mathbf{w}\|_2^{1/2}} - \rho\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2 \\
&\geq c'\tau^2 LR^2 \frac{\|(\mathbf{w} - \mathbf{w}^*)^{\|_{\mathbf{w}}}\|_2}{\|\mathbf{w}\|_2^2} + c'\tau^2 LR^2 \frac{1}{\|\mathbf{w}\|_2} \left( \|\text{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 - \frac{\rho\|\mathbf{w}\|_2^2\|\mathbf{w}^*\|_2}{c'\tau^2 LR^2} \right) .
\end{aligned}$$

To bound the $\|\text{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 - \rho\|\mathbf{w}\|_2^2\|\mathbf{w}^*\|_2/(c'\tau^2 LR^2)$, we have that

$$\frac{\|\text{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 c'\tau^2 LR^2}{2\rho\|\mathbf{w}\|_2^2\|\mathbf{w}^*\|_2} \geq \frac{Kc'\tau^2 LR}{\rho\|\mathbf{w}\|_2^2\|\mathbf{w}^*\|_2} \geq \frac{KF^3(\mathbf{w}^*)LR\tau^2}{\rho c'^2\kappa^2 U} \geq 1 ,$$

where the in the first inequality we used that $\|\text{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 \geq 2K/R$; in the second that $\|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$ and $\|\mathbf{w}^*\|_2\epsilon \leq U$; and in the last inequality that $\rho \leq \frac{K\epsilon^3 LR\tau^2}{c'^2\kappa^2 U}$. Therefore, we have that $(\|\text{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 - 2\rho\|\mathbf{w}\|_2^2\|\mathbf{w}^*\|_2/(c'LR^2\tau^2)) \geq \|\text{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2/2$ and the result follows similar to the proof of Claim 5.2.19. □

To prove the Proposition 5.2.6, it remains to show that $\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}$ is small when $\|\text{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 \leq 2K/R$ and $c'\kappa/(\epsilon\Lambda_2) \leq \|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$. Similarly to the proof of Claim 5.2.20, we need to consider only the case where $\theta \in (0, \pi/2)$ and $\|\mathbf{w}\|_2 \leq 2\|\mathbf{w}^*\|_2$.

In fact, we show that in this case the vector $\mathbf{w}$ is close to the target vector $\mathbf{w}^*$. We have:

$$\|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{w}} \leq \frac{2\|\mathbf{w}\|_2+\|\mathbf{w}^*\|_2}{\|\mathbf{w}\|_2^{3/2}} + \frac{\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2}{\|\mathbf{w}\|_2^{1/2}} \leq \frac{K/R+1}{\|\mathbf{w}\|_2^{1/2}} + \frac{\|\mathbf{w}^*\|_2}{\|\mathbf{w}\|_2^{3/2}}\,,$$

where in the first inequality we used the triangle inequality and in the second the fact that $\|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{w}^*\|_2 \leq 2K/R$. Moreover, recall that $c'\kappa/(\epsilon\Lambda_2) \leq \|\mathbf{w}\|_2$, $\|\mathbf{w}^*\|_2 \leq U/\epsilon$ and $K = \Lambda_2^{2.5}U/(\sqrt{c'\kappa}RL)$. Therefore, we have that $\|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{w}} \leq U\sqrt{\epsilon}\Lambda_2^3/(c'\kappa)^{3/2} \leq (cU/\kappa^5)\sqrt{\epsilon}$, for some absolute constant $c>0$.

### 5.2.2 Proof of Theorem 5.2.4

Before proving Theorem 5.2.4, we need to show that when the two vectors $\mathbf{w},\mathbf{v}$ have small distance with respect to an appropriate norm $\|\cdot\|$ (either the $\ell_2$ or the Weighted-Euclidean norm of Definition 5.2.2), then the $L_2^2$ distance of the corresponding sigmoidals, i.e., $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))^2]$, is small.

**Lemma 5.2.21** (Parameter vs $L_2^2$ Distance)**.** *Let $\mathcal{D}_\mathbf{x}$ be an $(L,R)$ well-behaved distribution. Let $\sigma$ be a $(\tau,\xi,\mu)$-sigmoidal activation. For any vectors $\mathbf{w},\mathbf{v}\in\mathbb{R}^d$, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))^2] \leq \xi^2\|\mathbf{w}-\mathbf{v}\|_2^2$. Moreover, if $\theta=\theta(\mathbf{w},\mathbf{v})<\pi/4$, $\|\mathbf{w}\|_2\leq\delta\|\mathbf{v}\|_2$ and $\delta\geq 1$, and $\|\mathbf{w}\|_2>2/R$, it holds*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))^2] \lesssim \frac{\xi^2\delta^3}{L^4\mu^3}\|\mathbf{w}-\mathbf{v}\|_{\mathbf{w}}^2\,.$$

*Proof.* First, we consider the case where $\|\mathbf{w}\|_2\geq 2/R$ and $\theta\in(0,\pi/4)$. Let $V$ be the subspace spanned by $\mathbf{w},\mathbf{v}$ and assume without loss of generality that $\mathbf{w}/\|\mathbf{w}\|_2=\mathbf{e}_2$ and

therefore $\mathbf{v} = \|\mathbf{v}\|_2(\cos\theta\mathbf{e}_2 - \sin\theta\mathbf{e}_1)$, we abuse the notation of $\mathcal{D}_\mathbf{x}$ to be the distribution projected on $V$. It holds that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{v}\cdot\mathbf{x}))^2] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\left(\int_{\mathbf{v}\cdot\mathbf{x}}^{\mathbf{w}\cdot\mathbf{x}} \sigma'(t)\mathrm{d}t\right)^2\right] \\ &\le \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\left(\int_{\mathbf{v}\cdot\mathbf{x}}^{\mathbf{w}\cdot\mathbf{x}} \xi\exp(-\mu|t|)\mathrm{d}t\right)^2\right] \\ &\le \xi^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{v}\cdot\mathbf{x})^2(\exp(-2\mu|\mathbf{w}\cdot\mathbf{x}|) + \exp(-2\mu|\mathbf{v}\cdot\mathbf{x}|))]\,,\end{aligned}$$

where we used that $\sigma$ is non-decreasing. Let $\mathbf{q} = \mathbf{w} - \mathbf{v}$. Using Claim 5.2.10, it holds that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{v}\cdot\mathbf{x})^2\exp(-2\mu|\mathbf{w}\cdot\mathbf{x})|)] &= \mathbf{E}[((\mathbf{q}^{\|\mathbf{w}}\cdot\mathbf{x})^2 + (\mathbf{q}^{\perp\mathbf{w}}\cdot\mathbf{x})^2)\exp(-2\mathbf{w}\cdot\mathbf{x}\mu)] \\ &\le \frac{8}{L^4}(\|\mathbf{q}^{\|\mathbf{w}}\|_2^2\frac{1}{\mu^3\|\mathbf{w}\|_2^3} + \|\mathbf{q}^{\perp\mathbf{w}}\|_2^2\frac{1}{\mu\|\mathbf{w}\|_2}) \le \frac{8\|\mathbf{q}\|_\mathbf{w}^2}{L^4\mu^3}\,.\end{aligned}$$

Moreover, for the second term, it holds that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{v}\cdot\mathbf{x})^2\exp(-2\mu|\mathbf{v}\cdot\mathbf{x})|)] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{q}^{\|\mathbf{w}}\cdot\mathbf{x})^2 + (\mathbf{q}^{\perp\mathbf{w}}\cdot\mathbf{x})^2)\exp(-2\mathbf{v}\cdot\mathbf{x}\mu)] \\ &\le \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{q}^{\|\mathbf{w}}\cdot\mathbf{x})^2 + (\mathbf{q}^{\perp\mathbf{w}}\cdot\mathbf{x})^2)\exp(-2\|\mathbf{v}\|_2\cos\theta\mathbf{x}_2\mu)] \\ &\le \frac{8}{L^4}(\|\mathbf{q}^{\|\mathbf{w}}\|_2^2\frac{1}{\mu^3\|\mathbf{v}\|_2^3\cos\theta^3} + \|\mathbf{q}^{\perp\mathbf{w}}\|_2^2\frac{1}{\mu\|\mathbf{v}\|_2\cos\theta}) \\ &\lesssim \frac{\delta^3\|\mathbf{q}\|_\mathbf{w}^2}{L^4\mu^3}\,,\end{aligned}$$

where in the last inequality, we used that $\cos\theta \ge 1/2$ and that $\|\mathbf{w}\|_2 \le \delta\|\mathbf{v}\|_2$. For the

other cases, we use the "trivial" upper-bound, i.e.,

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{v}\cdot\mathbf{x}))^2] &\le \xi^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x}-\mathbf{v}\cdot\mathbf{x})^2] \\ &\le \xi^2\|\mathbf{w}-\mathbf{v}\|_2^2 \sup_{\|\mathbf{w}\|_2=1} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\mathbf{w}\cdot\mathbf{x})^2] \\ &\le \xi^2\|\mathbf{w}-\mathbf{v}\|_2^2 \,,\end{aligned}$$

where we used that $\sigma$ is $\xi$-Lipschitz and that the distribution $\mathcal{D}_{\mathbf{x}}$ is isotropic. □

We now can now prove Theorem 5.2.4.

*Proof of Theorem 5.2.4.* We note that from Lemma 5.2.5, there exists a $\mathbf{w}^* \in \mathbb{R}^d$, such that $F(\mathbf{w}^*)\|\mathbf{w}^*\|_2 = 1 + O(\xi/(\mu L)) = U$ and $F(\mathbf{w}^*) \lesssim \epsilon$. We can assume that $\epsilon \le \mathrm{poly}(\xi/(LR\mu))$, since otherwise any vector $\mathbf{w}$, gets error $F(\mathbf{w}) \le 2\epsilon$. First we consider the case where $\|\mathbf{w}\|_2 \le 2/R$. From Proposition 5.2.6 for $c' > 0$ a sufficiently small absolute constant, we have that if $\|\mathbf{w}^* - \mathbf{w}\|_2 \ge \sqrt{\epsilon}/(c'\kappa^5)$, then

$$\|\nabla F(\mathbf{w})\|_2 \ge \nabla F(\mathbf{w}) \cdot \frac{\mathbf{w}-\mathbf{w}^*}{\|\mathbf{w}-\mathbf{w}^*\|_2} \ge c'\sqrt{\epsilon}\,.$$

In order to reach a contradiction, assume that $\|\nabla F(\mathbf{w})\|_2 < c'\sqrt{\epsilon}$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] > \xi^2\epsilon/(c'^2\kappa^{10})$. From Proposition 5.2.6, we have that $\|\mathbf{w}^* - \mathbf{w}\|_2 \le \sqrt{\epsilon}/(c'\kappa^5)$. Therefore, from Lemma 5.2.21 we have that

$$\xi^2\epsilon/(c'^2\kappa^{10}) < \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] \le \|\mathbf{w}-\mathbf{w}^*\|_2^2 \le \xi^2\epsilon/(c'^2\kappa^{10})\,,$$

which is a contradiction.

Next, we consider the case where $\|\mathbf{w}\|_2 \ge c'\kappa/(2\epsilon)$. From Proposition 5.2.6, we know

that there is no approximate stationary point in this region, i.e., there is no point $\mathbf{w}$ with $\|\nabla F(\mathbf{w})\|_{*,\mathbf{w}} \leq c'\sqrt{\epsilon}$.

For the last case we consider the case where $2/R \leq \|\mathbf{w}\|_2 \leq c'\kappa/\epsilon$. From Proposition 5.2.6, we have that if either $\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}} \geq U\sqrt{\epsilon}/(c'\kappa^5)$ or $\|\mathbf{w}\|_2 \geq 2\|\mathbf{w}^*\|_2$, then

$$\|\nabla F(\mathbf{w})\|_{*,\mathbf{w}} \geq \nabla F(\mathbf{w}) \cdot \frac{\mathbf{w} - \mathbf{w}^*}{\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}} \geq c'\sqrt{\epsilon}\,,$$

where we used that for any two vectors $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^d$, it holds $\mathbf{z}_1 \cdot \mathbf{z}_2 \leq \|\mathbf{z}_1\|_{*,\mathbf{w}}\|\mathbf{z}_2\|_{\mathbf{w}}$.

In order to reach a contradiction, assume that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] \geq C\frac{\xi^2U^2}{L^4\mu^3\kappa^{10}}\epsilon$ and that $\|\nabla F(\mathbf{w})\|_{*,\mathbf{w}} \leq c'\sqrt{\epsilon}$, where $C > 0$ is a sufficiently large absolute constant. It holds that $\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} \leq U\sqrt{\epsilon}/(c'\kappa^5)$ and $\|\mathbf{w}\|_2 \leq 2\|\mathbf{w}^*\|_2$. Hence, we have that

$$\|\mathbf{w}^* - \mathbf{w}\|_{\mathbf{w}} = \frac{|\|\mathbf{w}\|_2 - \cos\theta\|\mathbf{w}^*\|_2|}{\|\mathbf{w}\|_2^{3/2}} + \frac{\|\mathbf{w}^*\|_2\sin\theta}{\sqrt{\|\mathbf{w}\|_2}} \lesssim \frac{U\sqrt{\epsilon}}{\kappa^5}\,.$$

Therefore, it holds that

$$\sin\theta \lesssim \frac{\sqrt{\epsilon\|\mathbf{w}\|_2}}{\|\mathbf{w}^*\|_2\kappa^5} \lesssim \frac{\sqrt{\epsilon}}{\sqrt{\|\mathbf{w}^*\|_2}\kappa^5}\,,$$

where we used that $\|\mathbf{w}\|_2 \leq 2\|\mathbf{w}^*\|_2$, and using that $\sqrt{\epsilon/\|\mathbf{w}^*\|_2}$ is smaller than a sufficiently small absolute constant, we get that $\sin\theta \leq 1/2$, and hence, $\theta \in (0, \pi/4)$. Using again Lemma 5.2.21, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] \lesssim \frac{\xi^2}{L^4\mu^3}\|\mathbf{w} - \mathbf{w}^*\|_{\mathbf{w}}^2 \lesssim \frac{\xi^2U^2}{L^4\mu^3\kappa^{10}}\epsilon\,.$$

Therefore, we get again a contradiction. The proof then follows by noting that $F(\mathbf{w}) \leq 2\epsilon + 2\,\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2]$.

□

## 5.3 The Landscape of the $L_2^2$-Loss for Unbounded Activations

For unbounded activations we essentially characterize the optimization landscape of the $L_2^2$-loss as a function of the weight vector $\mathbf{w}$. Specifically, our main structural theorem in this section establishes (roughly) the following: even though the population $L_2^2$-loss $F(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2]$ is not convex, any approximate stationary point $\mathbf{w}$ of $F(\mathbf{w})$ will have error close to that of the optimal weight vector $\mathbf{w}^*$, i.e., $F(\mathbf{w}) \leq O(\epsilon)$. In more detail, while there exist "bad" stationary points in this case, they all lie in a cone around $-\mathbf{w}^*$, i.e., in the opposite direction of the optimal weight vector. As we will see in Section 5.4, gradient descent initialized at the origin will always avoid such stationary points. Our main structural result for unbounded activations is as follows:

**Theorem 5.3.1** (Stationary Points of $(\alpha,\lambda)$-Unbounded Activations)**.** *Let $\mathcal{D}$ be an $(\epsilon, W)$-corrupted, $(L,R)$-well-behaved distribution in $\mathbb{R}^d$. Let $\sigma$ be an $(\alpha,\lambda)$-unbounded activation and let $F(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2]$. Then, if for some $\mathbf{w}\in\mathbb{R}^d$, with $\mathbf{w}\cdot\mathbf{w}^* \geq 0$ and $\|\mathbf{w}\|_2 \leq W$ it holds $\|\nabla F(\mathbf{w})\|_2 \leq 2\lambda\sqrt{\epsilon}$, then $F(\mathbf{w}) \leq C\left(\frac{\lambda}{\alpha}\right)^4 \frac{1}{L^2R^8}\epsilon$, for some absolute constant $C > 0$.*

**Remark 5.3.2.** An important feature of the above theorem statement is that the error guarantee of the stationary points of the $L_2^2$-loss $F(\mathbf{w})$ does not depend on the size of the weight vector $\mathbf{w}^*$. It only depends on the ratio of the constants $\lambda$ and $\alpha$ (that for all activation functions discussed above is some small absolute constant). For the special case of ReLU activations, it holds that $\lambda/\alpha = 1$.

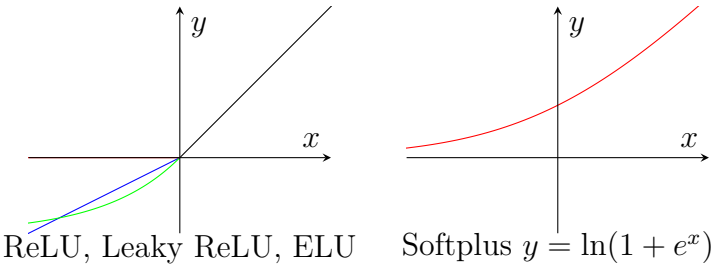



Figure 5.4: Some well-known unbounded activation functions that satisfy Definition 5.1.5 .

### Proof of Theorem 5.3.1

**Remark 5.3.3.** We remark that we assume $\sigma(0) = 0$ to simplify the presentation. If $\sigma(0) \neq 0$, then we can always consider the loss function $\tilde{\sigma}(t) = \sigma(t) - \sigma(0)$ and also subtract $\sigma(0)$ from the labels $y$, i.e., $y' = y - \sigma(0)$. Our results directly apply to the transformed instance.

For the proof of Theorem 5.3.1, we need the following structural result for the gradient field of the $L_2^2$ loss for unbounded activations. We show that as long as $\|\mathbf{w} - \mathbf{w}^*\|_2$ is larger than roughly $\sqrt{\epsilon}$ and $\mathbf{w} \cdot \mathbf{w}^* \geq 0$, then the gradient at $\mathbf{w}$ has large projection in the direction of $\mathbf{w} - \mathbf{w}^*$, i.e., it "points in the right direction".

**Proposition 5.3.4.** *Let $\mathcal{D}$ be an $(\epsilon, W)$-corrupted, $(L, R)$-well-behaved distribution and $\sigma$ be an $(\alpha, \lambda)$-unbounded activation. For any $\mathbf{w} \in \mathbb{R}^d$ with $\mathbf{w} \cdot \mathbf{w}^* \geq 0$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \geq C\lambda/(\alpha^2 L R^4)\sqrt{\epsilon}$, it holds $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \gtrsim \alpha^2 L R^4 \|\mathbf{w} - \mathbf{w}^*\|_2^2$.*

We remark that one can readily use the above proposition to obtain an efficient algorithm via black-box optimization methods. To achieve this, we can simply start by finding a stationary point $\mathbf{w}^{(0)}$, and then repeat our search constraining on the set $\{\mathbf{w} : \mathbf{w} \cdot \mathbf{w}^{(0)} \leq 0\}$ to obtain some stationary point $\mathbf{w}^{(1)}$. Then we continue in the set

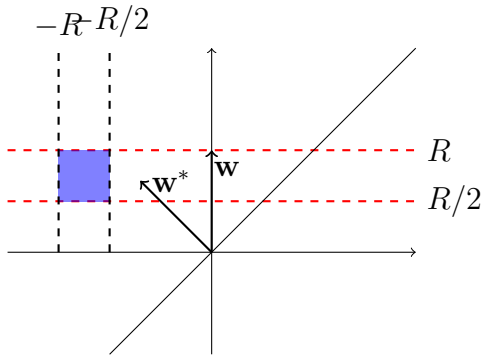



Figure 5.5: Using our distributional assumptions, we identify a region ("blue") that provides enough contribution to the gradient, so that an update step will decrease the distance with the optimal one.

$\{\mathbf{w}^{(0)} \cdot \mathbf{w} \leq 0, \mathbf{w}^{(1)} \cdot \mathbf{w} \leq 0\}$. Since for all the boundaries we have (using Proposition 5.3.4) that the gradient has non-zero projection onto the direction $\mathbf{w} - \mathbf{w}^*$, we obtain that by adding these linear constraints, we do not introduce new stationary points. It is not hard to see that after $O(d)$ such iterations, we obtain a stationary point that lies in the halfspace $\mathbf{w}^* \cdot \mathbf{w} \geq 0$, and therefore is an approximately optimal solution.

*Proof of Proposition 5.3.4.* We have

$$\begin{aligned}
&\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \\
&\qquad = \underbrace{\mathbf{E}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})]}_{I_1} \\
&\qquad + \underbrace{\mathbf{E}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - \sigma(\mathbf{w}^* \cdot \mathbf{x}))\sigma'(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})]}_{I_2} .
\end{aligned}$$

We start by bounding the contribution of $I_1$. We show that the "noisy" integral $I_1$ has small negative contribution that is bounded by some multiple of $\|\mathbf{w} - \mathbf{w}^*\|_2 \sqrt{\epsilon}$. We show the following claim.

**Claim 5.3.5.** *It holds that* $I_1 \geq -2\lambda\|\mathbf{w}-\mathbf{w}^*\|_2\sqrt{\epsilon}$.

*Proof.* Using the Cauchy-Schwarz inequality, we have that

$$\begin{aligned} I_1 &\geq -\mathbf{E}[|(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)\sigma'(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})|] \\ &\geq -\lambda\,\mathbf{E}[|\sigma(\mathbf{w}^*\cdot\mathbf{x})-y||\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x}|] \\ &\geq -\lambda(\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2])^{1/2}(\mathbf{E}[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2])^{1/2} \\ &\geq -\lambda\,\sqrt{\epsilon}\,\|\mathbf{w}-\mathbf{w}^*\|_2\left(\max_{\mathbf{v}\in\mathbb{R}^d,\|\mathbf{v}\|_2=1}\mathbf{E}[(\mathbf{v}\cdot\mathbf{x})^2]\right)^{1/2} \geq -2\lambda\|\mathbf{w}-\mathbf{w}^*\|_2\sqrt{\epsilon}\,, \end{aligned}$$

where in the first inequality we used the fact that $\sigma$ is $\lambda$-Lipschitz, in the second that $F(\mathbf{w}^*)=\epsilon$, and in the last the fact that the distribution is isotropic. Hence, we have that

$$I_1 \geq -2\lambda\|\mathbf{w}-\mathbf{w}^*\|_2\sqrt{\epsilon}\,. \tag{5.4}$$

This completes the proof of Claim 5.3.5. □

Next we bound from below the contribution of the "clean" examples, i.e., $I_2$. We show that this positive contribution is bounded below by a multiple of $\|\mathbf{w}-\mathbf{w}^*\|_2^2$, which is enough to surpass the contribution of the negative region $I_2$. We have the following claim.

**Claim 5.3.6.** *It holds that* $I_2 \geq \frac{7\alpha^2 LR^4}{64}\|\mathbf{w}-\mathbf{w}^*\|_2^2$.

*Proof.* We first notice that $(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))\sigma'(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})\geq 0$ for every $\mathbf{x}\in\mathbb{R}^d$, because $\sigma$ is an increasing function. Without loss of generality, we may assume that $\mathbf{w}/\|\mathbf{w}\|_2=\mathbf{e}_2$, and therefore $\mathbf{w}^*=\|\mathbf{w}^*\|_2(\cos\theta\mathbf{e}_2-\sin\theta\mathbf{e}_1)$. For simplicity, we abuse notation and denote $\mathcal{D}_{\mathbf{x}}$ the marginal distribution on the subspace spanned by

the vectors $\mathbf{w}, \mathbf{w}^*$. From the definition $(\alpha, \lambda)$-unbounded activations, we have that $\inf_{t\in(0,R/\|\mathbf{w}\|_2)} \sigma'(t) \geq \inf_{t\in(0,\infty)} \sigma'(t) \geq \alpha > 0$. We now have that

$$\begin{aligned} I_2 &= \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))\sigma'(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})] \\ &\geq \alpha^2\, \mathbf{E}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2 \mathbb{1}\{\mathbf{x}_1 \in (-R, -R/2), \mathbf{x}_2 \in (R/2, R)\}] \,, \end{aligned}$$

where the region inside the indicator is the "blue region" of Figure 5.5. We remark that this region is not particularly "special": we can use other regions that contain enough mass depending on the distributional assumptions. At a high level, we only need a rectangle that contains enough mass and the gradient in these points is non-zero. Therefore, using the fact that the distribution is $(L, R)$-well-behaved, we have

$$\begin{aligned} &\mathbf{E}\left[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\ \mathbb{1}\{\mathbf{x}_1 \in (-R, -R/2), \mathbf{x}_2 \in (R/2, R)\right] \\ &\qquad = \mathbf{E}\left[(\|\mathbf{q}^{\parallel_\mathbf{w}}\|_2^2 \mathbf{x}_2^2 + \|\mathbf{q}^{\perp_\mathbf{w}}\|_2^2 \mathbf{x}_1^2)\ \mathbb{1}\{\mathbf{x}_1 \in (-R, -R/2), \mathbf{x}_2 \in (R/2, R)\}\right] \\ &\qquad \geq \frac{7LR^4}{32}(\|\mathbf{q}^{\parallel_\mathbf{w}}\|_2^2 + \|\mathbf{q}^{\perp_\mathbf{w}}\|_2^2) = \frac{7LR^4}{64}\|\mathbf{w} - \mathbf{w}^*\|_2^2 \,, \end{aligned}$$

where $\mathbf{q} = \mathbf{w} - \mathbf{w}^*$. Therefore, we proved that

$$I_2 \geq \frac{7\alpha^2 LR^4}{64}\|\mathbf{w} - \mathbf{w}^*\|_2^2 \,. \tag{5.5}$$

□

Thus, combining Equations (5.4) and (5.5), we have that

$$\nabla F(\mathbf{w})\cdot(\mathbf{w} - \mathbf{w}^*) \geq \|\mathbf{w} - \mathbf{w}^*\|_2 \left(\alpha^2 \frac{7LR^4}{64}\|\mathbf{w} - \mathbf{w}^*\|_2 - 2\lambda\sqrt{\epsilon}\right) \,,$$

which completes the proof of Proposition 5.3.4. □

To prove Theorem 5.3.1, in order to reach a contradiction we assume that $F(\mathbf{w}) \geq C\frac{\lambda^4}{\alpha^4 L^2 R^8}\epsilon$ for absolute constant $C > 0$ and that $\|\nabla F(\mathbf{w})\|_2 \leq \frac{\alpha^2 L R^4}{100}\sqrt{\epsilon}$. From Proposition 5.3.4, we have that

$$\frac{7\alpha^2 L R^4}{64}\sqrt{\epsilon} \geq \nabla F(\mathbf{w}) \cdot \frac{(\mathbf{w} - \mathbf{w}^*)}{\|\mathbf{w} - \mathbf{w}^*\|_2} \geq \frac{7\alpha^2 L R^4}{64}\left(\|\mathbf{w} - \mathbf{w}^*\|_2 - \sqrt{\epsilon}\frac{\lambda 128}{7\alpha^2 L R^4}\right) ,$$

and therefore

$$\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \frac{128\lambda\sqrt{\epsilon}}{7\alpha^2 L R^4} . \tag{5.6}$$

Moreover, we have

$$\begin{aligned} F(\mathbf{w}) &= \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] \leq 2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2] + 2\,\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] \\ &\leq 2\epsilon + 2\lambda^2\|\mathbf{w} - \mathbf{w}^*\|_2^2 . \end{aligned}$$

Therefore, using Equation (5.6), we get

$$F(\mathbf{w}) \leq C\left(\frac{\lambda}{\alpha}\right)^4 \frac{1}{L^2 R^8}\epsilon ,$$

which leads to a contradiction. This completes the proof of Theorem 5.3.1. □

## 5.4 Optimizing the Empirical $L_2^2$-Loss

In this section, we show that given sample access to a distribution over labeled examples $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$, we can optimize the $L_2^2$ loss via (approximate) gradient descent. This allows us to efficiently find a weight vector that achieves error $O(\epsilon)$. Ideally, we would like to

perform gradient descent with the population gradients, i.e., $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta(\nabla F(\mathbf{w}^{(t)}))$. We do not have direct access to these gradients, but we show that it is not hard to estimate them using samples from $\mathcal{D}$; see Algorithm 8. We state our algorithmic result for sigmoidal activations in Section (5.4.1) and for unbounded activations in Section (5.4.2).

### 5.4.1 Sigmoidal Activations

In this section, we prove our main algorithmic result for learning the sigmoidal activations discussed in Section 5.2.

**Theorem 5.4.1** (Learning Sigmoidal Activations)**.** *Let $\mathcal{D}$ be an $\epsilon$-corrupted, $(L, R)$-well-behaved distribution on $\mathbb{R}^d \times \mathbb{R}$ and $\sigma(\cdot)$ be a $(\tau, \mu, \xi)$-sigmoidal activation. Set $\kappa = L^6 R^6 \mu^3 \tau^4 / \xi^2$ and let $c > 0$ be a sufficiently small absolute constant. Then gradient descent (Algorithm 8) with step size $\eta = c\epsilon^{2.5}$, regularization $\rho = (1/c)\epsilon^3/k^5$, truncation threshold $M = \xi/\mu$, $N = \widetilde{\Theta}(d/\epsilon \log(1/\delta))\ \mathrm{poly}(1/\kappa)$ samples, and $T = \mathrm{poly}(1/(\epsilon\kappa))$ iterations converges to a vector $\mathbf{w}^{(T)} \in \mathbb{R}^d$ that, with probability $1-\delta$, satisfies $F(\mathbf{w}^{(T)}) \leq \mathrm{poly}(1/\kappa)\ \epsilon$.*

**Corollary 5.4.2.** *Let $\mathcal{D}$ be an $(L, R)$-well-behaved distribution on $\mathbb{R}^d \times \mathbb{R}$ and $\sigma(\cdot)$ be a $(\tau, \mu, \xi)$-sigmoidal activation. Define $F(\mathbf{w}) = (1/2)\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2]$ and $\mathrm{OPT} = \inf_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w})$. Set $\kappa = \mathrm{poly}(LR\tau\mu/\xi)$. There exists an algorithm that given $N = \widetilde{\Theta}(d/\epsilon \log(1/\delta))\ \mathrm{poly}(1/\kappa)$ samples, runs in time $T = \mathrm{poly}(1/(\epsilon\kappa))$ and returns a vector $\widehat{\mathbf{w}} \in \mathbb{R}^d$ that, with probability $1-\delta$, satisfies*

$$F(\widehat{\mathbf{w}}) \leq \mathrm{poly}(1/\kappa)\ \mathrm{OPT} + \epsilon\,.$$

*Proof.* The proof of Corollary 5.4.2 is similar to the proof of Theorem 5.4.1, the main point which the two proofs diverge is that we have to guess the value $\rho$ of the regularizer. To do this, we construct a grid of the possible values of OPT, i.e., $\mathcal{G} = \{\epsilon, 2\epsilon, \ldots, \Theta(\xi/\mu)\}$. We choose the upper bound to be of size $\Theta(\xi/\mu)$ because this is the maximum value that the error can get, see e.g., Fact 5.2.17. Therefore, by running the Algorithm 8, with appropriate parameters as in Theorem 5.4.1 but by using for the regularizer, each value of $\mathcal{G}$ instead of OPT. There exist a $t \in \mathcal{G}$ such that $|\mathrm{OPT} - t| \leq \epsilon$, therefore from Proposition 5.2.6, this will converge to a stationary point with $(1+\epsilon)\mathrm{poly}(1/k)\mathrm{OPT}$, see e.g., Proposition 5.2.6 for $\Lambda = 1 + \epsilon$. Hence, by running Algorithm 8 for each value in $\mathcal{G}$, we output a list $U$ that contains $O(\xi/(\mu\epsilon))$ different hypothesis. We construct the empirical $\mathcal{D}_N$ of $\mathcal{D}$ by taking $N$ samples. We need enough samples, such that for each $\mathbf{w} \in U$, it will hold $\widehat{F}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}_N}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] \leq 2\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2]$. From Markov's inequality we have that

$$\mathbf{Pr}[|\widehat{F}(\mathbf{w}) - F(\mathbf{w})| \geq F(\mathbf{w})/2] \leq 4\frac{\mathbf{E}[\widehat{F}(\mathbf{w})^2]}{F(\mathbf{w})^2} \leq \frac{\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^4]}{NF(\mathbf{w})^2}\,.$$

Note that because $y, \sigma(\mathbf{w}\cdot\mathbf{x})$ is bounded by $\Theta(\xi/\mu)$, we have that $\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^4] \leq \mathrm{poly}(\xi/\mu)\,\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2]$, hence we have that

$$\mathbf{Pr}[|\widehat{F}(\mathbf{w}) - F(\mathbf{w})| \geq F(\mathbf{w})/2] \lesssim \mathrm{poly}(\xi/\mu)\frac{1}{NF(\mathbf{w})} \lesssim \mathrm{poly}(\xi/\mu)\frac{1}{N\epsilon}\,,$$

therefore by choosing $N = \mathrm{poly}(\xi/\mu)(1/\epsilon)$, we have with probability at least $2/3$ that $\widehat{F}(\mathbf{w}) \leq 2F(\mathbf{w}) + \epsilon$ and by standard boosting procedure we can increase the probability to $1 - \delta''$ with $O(\log(1/\delta''))$ samples. Hence, by choosing $\delta'' = \delta/|U|$, we have with probability $1 - \delta$, that for $\hat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in U}\,\widehat{F}(\mathbf{w})$, it holds that $F(\hat{\mathbf{w}}) \lesssim \mathrm{OPT} + \epsilon$. □

**Algorithm 8** Gradient Descent for the $L_2^2$-Loss with Sigmoidal Activations

**Input:** number of iterations $T$; samples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N \sim \mathcal{D}$; step size $\eta$; clipping bound $M$; regularization $\rho$.
Define $\widehat{F}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\Big(\sigma(\mathbf{w}\cdot\mathbf{x}^{(i)}) - \tilde{y}^{(i)}\Big)^2, \quad \tilde{y}^{(i)} = \mathrm{sign}(y^{(i)})\,\min\Big(|y^{(i)}|, M\Big)$.
Initialize $\mathbf{w}^{(0)} \leftarrow \mathbf{0}$.
**for** $t = 0$ **to** $T-1$ **do**
    $\mathbf{g}^{(t)} \leftarrow \nabla\widehat{F}\Big(\mathbf{w}^{(t)}\Big)$
    $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta\Big(\mathbf{g}^{(t)} + \rho\,\mathbf{w}^{(t)}\Big)$
**end for**
**Return:** $\mathbf{w}^{(T)}$

### Proof of Theorem 5.4.1

We will first assume that we have access to the population gradients of the $L_2^2$ objective and then show that given samples from $\mathcal{D}$, gradient descent again converges to some approximately optimal solution. Before we start, we observe that, without loss of generality, we may assume that $\sigma(0) = 0$ (otherwise we can subtract $\sigma(0)$ from $y$ and the activation, see Remark 5.3.3). Moreover, from Fact 5.2.17 we know that $|\sigma(t)| \leq \xi/\mu$. Therefore, for every example $(\mathbf{x}, y) \sim D$ we can "truncate" its label to $\widehat{y} = \mathrm{sign}(y)\min(|y|, \xi/\mu)$ and the new instance will still be (at most) $\epsilon$-corrupted. To simplify notation, from now on, we will overload the notation and use $y$ instead of $\widehat{y}$ (assuming that for every example $(\mathbf{x}, y)$ it holds that $|\sigma(\mathbf{w}\cdot\mathbf{x}) - y| \leq 2\xi/\mu$).

**Population Gradient Descent** Recall that in Proposition 5.2.6 we showed that the population gradient field "points" to the right direction, i.e., a step of gradient descent would decrease the distance between the current $\mathbf{w}$ and a target $\mathbf{v}$. Notice that in Proposition 5.2.6 we require that the target vector $\mathbf{v}$ satisfies $F(\mathbf{v}) = O(\epsilon)$ and $\|\mathbf{v}\|_2 \leq O(1/\epsilon)$. Indeed, from Lemma 5.2.5, we have that there exists an approximately

optimal target vector $\mathbf{v} \in \mathbb{R}^d$, such that $\|\mathbf{v}\|_2 \leq U/\epsilon$ for some $U = O(\xi/(\mu L))$ and $F(\mathbf{v}) = O(\epsilon)$.

Moreover, recall the way we measure distances changes depending on how far $\mathbf{w}$ is from the origin (see the three cases of Proposition 5.2.6). We first show that, given any gradient field that satisfies the properties given in Proposition 5.2.6, then gradient descent with an appropriate step size will converge to the target vector $\mathbf{v}$. The distance of $\mathbf{v}$ and the guess $\mathbf{w}^{(T)}$ after $T$ gradient iterations will be measured with respect to $\|\cdot\|_2$ when $\mathbf{w}^{(T)}$ is close to the origin and with respect to $\|\cdot\|_{\mathbf{w}^{(T)}}$ (see Definition 5.2.2) otherwise.

**Lemma 5.4.3** (Gradient Field Distance Reduction)**.** *Let $Z_1 > Z_0 \geq 1$. Let $\mathbf{g} : \mathbb{R}^d \mapsto \mathbb{R}^d$ be a vector field with $\|\mathbf{g}(\mathbf{w})\|_2 \leq B$ for every $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq 2Z_1$. We assume that $\mathbf{g}$ satisfies the following properties with respect to some unknown target vector $\mathbf{v}$:*

1. *If $\|\mathbf{w}\|_2 \leq Z_0$ and $\|\mathbf{w} - \mathbf{v}\|_2 \geq \alpha_1$ then it holds that $\mathbf{g}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{v}) \geq \alpha_2 \|\mathbf{w} - \mathbf{v}\|_2$, for $\alpha_1, \alpha_2 > 0$.*

2. *If $Z_0 < \|\mathbf{w}\|_2 \leq Z_1$ and $\|\mathbf{w} - \mathbf{v}\|_{\mathbf{w}} \geq \beta_1$ it holds that $\mathbf{g}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{v}) \geq \beta_2 \|\mathbf{w} - \mathbf{v}\|_{\mathbf{w}}$, for $\beta_1, \beta_2 > 0$.*

3. *For some $\zeta \in (0, 1), \gamma > 0$, we have that if $\|\mathbf{w}\|_2 \geq \zeta Z_1$, it holds that $\mathbf{g}(\mathbf{w}) \cdot \mathbf{w} \geq \gamma \|\mathbf{w}\|_{\mathbf{w}}$.*

*We consider the update rule $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \mathbf{g}(\mathbf{w}^{(t)})$ initialized with $\mathbf{w}^{(0)} = \mathbf{0}$ and step size*

$$\eta = \frac{1}{B^2} \min\left(\alpha_1\alpha_2, \frac{\beta_1\beta_2}{Z_1^{3/2}}, \frac{2\gamma}{Z_1}, (1-\zeta)Z_1 B\right) .$$

*Let $T$ be any integer larger than $\left\lceil \frac{\|\mathbf{v}\|^2}{\eta \min(\alpha_1\alpha_2, \beta_1\beta_2/Z_1^{3/2})} \right\rceil$. We have that $\|\mathbf{w}^{(T)}\|_2 \leq Z_1$.*

*Moreover, if* $\|\mathbf{w}^{(T)}\|_2 \leq Z_0$ *it holds that* $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq \eta B + \max(\alpha_1, (2Z_0)^{3/2}\beta_1)$ *and if* $\|\mathbf{w}^{(T)}\|_2 > Z_0$ *we have* $\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \sqrt{2}\eta B + \max(\sqrt{2}\alpha_1, e^{3Z_1^{3/2}\eta B}\beta_1)$ .

*Proof.* We first show by induction that for every $t \geq 0$ it holds that $\|\mathbf{w}^{(t)}\|_2 \leq Z_1$. For $t = 0$ we have $\mathbf{w}^{(0)} = \mathbf{0}$ so the claim holds. Assume first that $\|\mathbf{w}^{(t)}\|_2 \leq \zeta Z_1$. Then

$$\|\mathbf{w}^{(t+1)}\|_2 \leq \|\mathbf{w}^{(t)}\|_2 + \eta B \leq \zeta R_1 + (1-\zeta)Z_1 = Z_1,$$

where we used the fact that $\eta B \leq (1-\zeta)Z_1$. Now, if $\zeta Z_1 \leq \|\mathbf{w}^{(t)}\|_2 \leq Z_1$ by assumption 3 of the vector field $\mathbf{g}$, we have that

$$\begin{aligned}\|\mathbf{w}^{(t+1)}\|_2^2 &= \|\mathbf{w}^{(t)}\|_2^2 - 2\eta \mathbf{g}(\mathbf{w}^t)\cdot \mathbf{w}^{(t)} - \eta^2 \|\mathbf{g}(\mathbf{w}^{(t)})\|_2^2 \\ &\leq \|\mathbf{w}^{(t)}\|_2^2 - 2\eta\gamma/\sqrt{Z_1} - \eta^2 B^2 \,.\end{aligned}$$

Since $\eta \leq 2\gamma/(B^2\sqrt{Z_1})$, we have that $-2\eta\gamma/\sqrt{Z_1} - \eta^2 B^2 \leq 0$, and therefore $\|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 \leq Z_1^2$.

Now that we have that $\|\mathbf{w}^{(t)}\|_2$ is always smaller than $Z_1$, we know that the vector field $\mathbf{g}(\mathbf{w}^{(t)})$ has non-trivial component on the direction $\mathbf{w} - \mathbf{v}$, i.e., either condition 1 or condition 2 is true. We next show that when condition 1 or 2 of the Lemma 5.4.3 hold, we can improve the distance of $\mathbf{w}^{(t)}$ and the target $\mathbf{v}$. We first assume that $\|\mathbf{w}^{(t)}\|_2 \leq Z_0$ and that $\|\mathbf{w}^{(t)} - \mathbf{v}\|_2 \geq \alpha_1$. Then by assumption 1 of the lemma we have that it holds

$\mathbf{g}(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^{(t)} - \mathbf{v}) \geq \alpha_2 \|\mathbf{w}^{(t)} - \mathbf{v}\|_2$. We have

$$\begin{aligned}\|\mathbf{w}^{(t)} - \mathbf{v}\|_2^2 - \|\mathbf{w}^{(t+1)} - \mathbf{v}\|_2^2 &= 2\eta \mathbf{g}(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^{(t)} - \mathbf{v}) - \eta^2 \|\mathbf{g}(\mathbf{w}^{(t)})\|_2^2 \\ &\geq 2\eta\alpha_1\alpha_2 - \eta^2 \|\mathbf{g}(\mathbf{w}^{(t)})\|_2^2 \\ &\geq \eta\alpha_1\alpha_2 \,,\end{aligned}$$

where for the last inequality we used the fact that $\|\mathbf{g}(\mathbf{w}^{(t)})\|_2^2 \leq B^2$ and that $\eta \leq \alpha_1\alpha_2/B^2$. On the other hand, if $\|\mathbf{w}^{(t)}\|_2 \leq Z_0$ and $\|\mathbf{w}^{(t)} - \mathbf{v}\|_2 \geq \beta_1$, then by assumption 2 of the lemma we have that it holds $\mathbf{g}(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^{(t)} - \mathbf{v}) \geq \beta_2 \|\mathbf{w} - \mathbf{v}\|_{\mathbf{w}^{(t)}}$. Similarly to the previous case, we then have

$$\|\mathbf{w}^{(t)} - \mathbf{v}\|_2^2 - \|\mathbf{w}^{(t+1)} - \mathbf{v}\|_2^2 \geq 2\eta\beta_2 \|\mathbf{w}^{(t)} - \mathbf{v}\|_{\mathbf{w}^{(t)}} - \eta^2 \|\mathbf{g}(\mathbf{w}^{(t)})\|_2^2 \,.$$

We next bound from below the norm $\|\cdot\|_{\mathbf{w}}$ by the $\ell_2$ norm. The following rough estimate suffices.

**Claim 5.4.4.** *For every $\mathbf{x} \in \mathbb{R}^d$ it holds*

$$\min\Big(\frac{1}{\|\mathbf{w}\|_2^{3/2}}, \frac{1}{\|\mathbf{w}\|_2^{1/2}}\Big)\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_{\mathbf{w}} \leq \sqrt{2}\max\Big(\frac{1}{\|\mathbf{w}\|_2^{3/2}}, \frac{1}{\|\mathbf{w}\|_2^{1/2}}\Big)\|\mathbf{x}\|_2 \,.$$

*Proof.* The claim follows directly from the definition of the norm $\|\cdot\|_{\mathbf{w}}$ and the inequality $\sqrt{a^2+b^2} \leq a + b \leq \sqrt{2}\sqrt{a^2+b^2}$ that holds for all $a, b \geq 0$. □

Using the above bounds and the fact that $Z_1 \geq 1$, we conclude that $\|\mathbf{w}^{(t)} - \mathbf{v}\|_2^2 - \|\mathbf{w}^{(t+1)} - \mathbf{v}\|_2^2 \geq \eta\beta_1\beta_2/Z_1^{3/2}$, where we used the assumption that $\eta \leq \beta_1\beta_2/(Z_1^{3/2}B^2)$. Recall that we have set $\eta = 1/B^2 \min(\alpha_1\alpha_2, \beta_1\beta_2/Z_1^{3/2}, 2\gamma/Z_1, (1-\zeta)Z_1B)$ which implies that at every iteration where either $\|\mathbf{w}^{(t)}\|_2 \leq Z_0$ and $\|\mathbf{w}^{(t)} - \mathbf{v}\|_2 \geq \alpha_1$, or $\|\mathbf{w}^{(t)}\|_2 > Z_0$

and $\|\mathbf{w}^{(t)}-\mathbf{v}\|_{\mathbf{w}^{(t)}} \geq \alpha_1$ we have that $\|\mathbf{w}^{(t)}-\mathbf{v}\|_2^2-\|\mathbf{w}^{(t+1)}-\mathbf{v}\|_2^2 \geq \eta \min(\alpha_1\alpha_2, \beta_1\beta_2/Z_1^{3/2})$. Let us denote by $T$ the first iteration such that the distance reduction stops happening. Since we initialize at $\mathbf{w}^{(0)} = \mathbf{0}$ the first time that either of the above condition stops being true, namely, $T$ can be at most $\left\lceil \frac{\|\mathbf{v}\|^2}{\eta \min(\alpha_1\alpha_2,\beta_1\beta_2/Z_1^{3/2})} \right\rceil$.

After either $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq \alpha_1$ or $\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \beta_1$ we stop having the guarantee that updating the guess with $\mathbf{g}(\mathbf{w}^{(T)})$ will decrease the distance, and therefore we may move to the wrong direction. However, since our step size $\eta$ is small doing a small step in the wrong direction cannot increase the distance of $\mathbf{w}^{(T+1)}$ and $\mathbf{v}$ by a lot. We next show that even if we continue updating after the iteration $T$, we cannot make the distance of $\mathbf{w}^{(T)}$ and the target vector $\mathbf{v}$ much larger. We first assume that at iteration $T$ we have $\|\mathbf{w}^{(T)}\|_2 \leq Z_0$, and therefore, it must be the case that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq \alpha_1$. Notice that it suffices to show that after one iteration the claim holds, since after $\|\mathbf{w}^{(T+1)}\|_2$ grows larger than $\alpha_1$ the distance of $\mathbf{w}$ will start to decrease again. For the next iteration, using the bound $\|\mathbf{g}(\mathbf{w}^{(T)})\|_2 \leq B$, we obtain that $\|\mathbf{w}^{(T+1)} - \mathbf{w}^{(T)}\|_2 \leq \eta B$. Now if $\|\mathbf{w}^{(T+1)}\|_2 \leq Z_0$, using the triangle inequality for the $\ell_2$ norm, we have that $\|\mathbf{w}^{(T+1)} - \mathbf{v}\|_2 \leq \alpha_1 + \eta B$. On the other hand, if $\|\mathbf{w}^{(T+1)}\|_2 > Z_0$, using the triangle inequality of the norm $\|\cdot\|_{\mathbf{w}}$ we obtain

$$\begin{aligned} \|\mathbf{w}^{(T+1)} - \mathbf{v}\|_{\mathbf{w}^{(T+1)}} &\leq \|\mathbf{w}^{(T+1)} - \mathbf{w}^{(T)}\|_{\mathbf{w}^{(T+1)}} + \|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T+1)}} \\ &\leq \sqrt{2}(\eta B + \alpha_1)\,, \end{aligned} \tag{5.7}$$

where we used Claim 5.4.4, the fact that $\mathbf{w}^{(T+1)} - \mathbf{w}^{(T)} = \eta \mathbf{g}(\mathbf{w}^{(T)})$ and that $\|\mathbf{w}^{(T+1)}\|_2 > Z_0 \geq 1$. Next we consider the case $\|\mathbf{w}^{(T)}\|_2 > Z_0$ and $\|\mathbf{w}^{(T+1)}\|_2 > Z_0$. From Claim 5.4.4, we obtain that $\|\mathbf{w}^{(T+1)} - \mathbf{w}^{(T)}\|_{\mathbf{w}^{(T+1)}} \leq \sqrt{2}\eta B$. We prove the next lemma that bounds the ratio between two weighted euclidean norms with different bases.

**Lemma 5.4.5.** *Let $\mathbf{u}, \mathbf{v}, \mathbf{x} \in \mathbb{R}^d$ be non-zero vectors with $\|\mathbf{u}\|_2, \|\mathbf{v}\|_2 \leq Q$ for some $Q \geq 1$. Then it holds that*

$$\frac{\|\mathbf{x}\|_{\mathbf{u}}}{\|\mathbf{x}\|_{\mathbf{v}}} \leq \exp\left(Q^{3/2}\left(4\theta(\mathbf{u},\mathbf{v})\left(\frac{1}{\|\mathbf{v}\|^{3/2}} + \frac{1}{\|\mathbf{v}\|_2^{1/2}}\right) + \Delta(\mathbf{u},\mathbf{v})\right)\right),$$

*where* $\Delta(\mathbf{u},\mathbf{v}) = \left|\frac{1}{\|\mathbf{v}\|_2^{3/2}} - \frac{1}{\|\mathbf{u}\|_2^{3/2}}\right| + \left|\frac{1}{\|\mathbf{v}\|_2^{1/2}} - \frac{1}{\|\mathbf{u}\|_2^{1/2}}\right|$.

*Proof.* We first observe that, $\|\lambda\mathbf{x}\|_{\mathbf{u}} = |\lambda|\|\mathbf{x}\|_{\mathbf{u}}$, it suffices to consider $\mathbf{x}$ with unit norm $\|\mathbf{x}\|_2 = 1$. Moreover, from Claim 5.4.4 we have that since $\|\mathbf{v}\|_2, \|\mathbf{u}\|_2 \leq Q$ it holds that both $\|\mathbf{x}\|_{\mathbf{u}}, \|\mathbf{x}\|_{\mathbf{v}}$ are larger than $1/Q^{3/2}$. Next, using the fact that the function $t \mapsto \log(t)$ is $1/t$-Lipschitz, we obtain that $\log\left(\frac{\|\mathbf{x}\|_{\mathbf{u}}}{\|\mathbf{x}\|_{\mathbf{v}}}\right) \leq Q^{3/2}|\|\mathbf{x}\|_{\mathbf{u}} - \|\mathbf{x}\|_{\mathbf{v}}|$. Therefore, it suffices to bound the difference of the two norms. We have that

$$|\|\mathbf{x}\|_{\mathbf{u}} - \|\mathbf{x}\|_{\mathbf{v}}| \leq \left|\frac{\|\mathrm{proj}_{\mathbf{u}}\mathbf{x}\|_2}{\|\mathbf{u}\|^{3/2}} - \frac{\|\mathrm{proj}_{\mathbf{v}}\mathbf{x}\|_2}{\|\mathbf{v}\|^{3/2}}\right| + \left|\frac{\|\mathrm{proj}_{\mathbf{u}^\perp}\mathbf{x}\|_2}{\|\mathbf{u}\|^{1/2}} - \frac{\|\mathrm{proj}_{\mathbf{v}^\perp}\mathbf{x}\|_2}{\|\mathbf{v}\|^{1/2}}\right|.$$

We first bound the term $\left|\frac{\|\mathrm{proj}_{\mathbf{u}}\mathbf{x}\|_2}{\|\mathbf{u}\|^{3/2}} - \frac{\|\mathrm{proj}_{\mathbf{v}}\mathbf{x}\|_2}{\|\mathbf{v}\|^{3/2}}\right| \leq \left|\frac{\|\mathrm{proj}_{\mathbf{u}}\mathbf{x}\|_2}{\|\mathbf{v}\|^{3/2}} - \frac{\|\mathrm{proj}_{\mathbf{v}}\mathbf{x}\|_2}{\|\mathbf{v}\|^{3/2}}\right| + \left|\frac{1}{\|\mathbf{u}\|^{3/2}} - \frac{1}{\|\mathbf{v}\|^{3/2}}\right|$, since $\|\mathrm{proj}_{\mathbf{v}}\mathbf{x}\|_2 \leq \|\mathbf{x}\|_2 = 1$. Moreover, by Cauchy-Schwarz inequality, we have that $\|\mathrm{proj}_{\mathbf{v}}\mathbf{x} - \mathrm{proj}_{\mathbf{u}}\mathbf{x}\|_2 \leq 2\theta(v,u)$. Similarly, we can bound from above the term $\|\mathrm{proj}_{\mathbf{v}^\perp}\mathbf{x}$-$\mathrm{proj}_{\mathbf{u}^\perp}\|_2$ by $2\theta(\mathbf{v},u)$. The bound follows. □

We will now use Lemma 5.4.5. We denote $\theta(\mathbf{w}^{(T)}, \mathbf{w}^{(T+1)})$ the angle between $\mathbf{w}^{(T)}, \mathbf{w}^{(T+1)}$ and by $\Delta(\mathbf{w}^{(T)}, \mathbf{w}^{(T+1)})$ the corresponding difference defined in Lemma 5.4.5. Since $\|\mathbf{w}^{(T)}\|_2, \|\mathbf{w}^{(T+1)}\|_2 \geq Z_0 \geq 1$, we have that $\theta(\mathbf{w}^{(T)}, \mathbf{w}^{(T+1)}) \leq \|\mathbf{w}^{(T)} - \mathbf{w}^{(T+1)}\|_2 \leq \eta B$. Moreover, using the fact that the mappings $\mathbf{w} \mapsto \|\mathbf{w}\|_2^{-3/2}$ and $\mathbf{w} \mapsto \|\mathbf{w}\|_2^{-1/2}$ defined for $\|\mathbf{w}\|_2 \geq 1$ are $3/2$ and $1/2$-Lipschitz respectively, we obtain that $\Delta(\mathbf{w}^{(T)}, \mathbf{w}^{(T+1)}) \leq \|\mathbf{w}^{(T)} - \mathbf{w}^{(T+1)}\|_2 \leq \eta B$. Using Lemma 5.4.5, we conclude that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T+1)}} \leq$

$e^{3Z_1^{3/2}\eta B}\|\mathbf{w}^{(T)}-\mathbf{v}\|_{\mathbf{w}^{(T)}}$. Using the triangle inequality similarly to Equation (5.7) and the fact that $\|\mathbf{w}^{(T)}-\mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \beta_1$, we obtain

$$\|\mathbf{w}^{(T+1)}-\mathbf{v}\|_{\mathbf{w}^{(T+1)}} \leq \sqrt{2}\eta B + e^{3Z_1^{3/2}\eta B}\beta_1 \,.$$

Finally, we have to consider the case where $\|\mathbf{w}^{(T)}\|_2 > Z_0$ and $\|\mathbf{w}^{(T+1)}\|_2 \leq Z_0$. Using once more the triangle inequality and Claim 5.4.4, we obtain $\|\mathbf{w}^{(T+1)}-\mathbf{v}\|_2 \leq \eta B + \|\mathbf{w}^{(T)}-\mathbf{v}\|_2 \leq \eta B + (Z_0+\eta B)^{3/2}\|\mathbf{w}^{(T)}-\mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \eta B + (2Z_0)^{3/2}\beta_1$, where we used the fact that $\eta B \leq 1$ and $Z_0 \geq 1$. Combining the bounds for the two cases where $\|\mathbf{w}^{(T+1)}\|_2 \leq Z_0$, we obtain that in this case it holds $\|\mathbf{w}^{(T+1)}-\mathbf{v}\|_2 \leq \eta B + \max(\alpha_1, (2Z_0)^{3/2}\beta_1)$. Similarly, when $\|\mathbf{w}^{(T+1)}\|_2 > Z_0$ we have $\|\mathbf{w}^{(T+1)}-\mathbf{v}\|_{\mathbf{w}^{(T+1)}} \leq \sqrt{2}\eta B + \max(\sqrt{2}\alpha_1, e^{3Z_1^{3/2}\eta B}\beta_1)$.

□

We now show that the population gradient field, $\nabla F(\mathbf{w})$ satisfies the assumptions of Lemma 5.4.9 for any $\|\mathbf{w}\|_2 \leq \kappa/\epsilon$. We first show that $\|\nabla F(\mathbf{w})\|_2$ is bounded. We have

$$\begin{aligned}
\|\nabla F(\mathbf{w})\|_2 &= \|\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}] + \rho\|\mathbf{w}\|_2\mathbf{w}\|_2 \\
&\leq \max_{\|\mathbf{u}\|_2=1} \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{u}\cdot\mathbf{x}] + \rho\|\mathbf{w}\|_2^2 \\
&\leq (\xi^2/\mu)\max_{\|\mathbf{u}\|_2=1} \mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|] + \rho\|\mathbf{w}\|_2^2 \leq \xi^2/\mu + \rho\|\mathbf{w}\|_2^2 \lesssim 1/\kappa \,,
\end{aligned}$$

where we used the fact that $|\sigma(\mathbf{w}\cdot\mathbf{x})-y| \leq \xi/\mu$, $\sigma'(t) \leq \xi$, that the $\mathbf{x}$-marginal of $\mathcal{D}$ is isotropic and that $\rho = O(\epsilon^3)$.

From Proposition 5.2.6 we have that for any vector $\mathbf{w} \in \mathbb{R}^d$, with $\|\mathbf{w}\|_2 \leq 2/R$, it holds that if $\|\mathbf{w}-\mathbf{w}^*\|_2 \geq \sqrt{\epsilon}/(c'\kappa)$ for some absolute constant $c' > 0$ it holds $\nabla F_\rho(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \geq \sqrt{\epsilon}\|\mathbf{w}-\mathbf{w}^*\|_2$. Furthermore, for any vector $\mathbf{w} \in \mathbb{R}^d$, with $2/R \leq$

$\|\mathbf{w}\|_2 \le c'\kappa/\epsilon$, if $\|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{w}} \ge c'U\sqrt{\epsilon}/\kappa$, then $\nabla F_\rho(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \ge c'\sqrt{\epsilon}\|\mathbf{w}-\mathbf{w}^*\|_{\mathbf{w}}$. Finally, for any vector $\mathbf{w}\in\mathbb{R}^d$, with $c''\kappa/\epsilon \le \|\mathbf{w}\|_2$, it holds that $\nabla F_\rho(\mathbf{w})\cdot\mathbf{w} \ge c'\sqrt{\epsilon}\|\mathbf{w}\|_{\mathbf{w}}$, for some $0 < c'' \le c'$. Therefore, the true gradient field of $F$ satisfies the assumptions of Lemma 5.4.9 with $B = O(1/\kappa)$, $Z_0 = O(1/R)$, $Z_1 = \mathrm{poly}(LR\tau\mu/\xi)$, $\zeta = 1/2$, $\alpha_1, \beta_1 = \mathrm{poly}(\xi/(\mu\tau LR))\sqrt{\epsilon}$, and $\alpha_2, \beta_2, \gamma = O(\sqrt{\epsilon})$.

**Using Sample-Estimated Gradients** In the following claim we show that since $\nabla F(\mathbf{w})$ is a sub-exponential random variable, with roughly $N = \widetilde{O}(d/\epsilon'^2)$ samples we can get $\epsilon$-estimates of the population gradients.

**Claim 5.4.6** (Sample Complexity of Gradient Estimation)**.** *Fix $B, \epsilon' > 0$ with $\epsilon' \le 1/\sqrt{B}$. Using $N = \widetilde{O}((d/\epsilon'^2)\mathrm{poly}(\xi/(L\mu))\log(B/\delta))$ samples from $\mathcal{D}$, we define the empirical $L_2^2$-loss as $\widehat{F}(\mathbf{w}) = \sum_{i=1}^N (\sigma(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)})^2$ and its corresponding empirical gradient field $\nabla\widehat{F}(\mathbf{w})$. With probability at least $1-\delta$, for all $\mathbf{w}$ with $\|\mathbf{w}\| \le B$, it holds $\|\nabla\widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \le \epsilon'$ and $\|\nabla\widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})\|_{*,\mathbf{w}} \le \epsilon'$.*

*Proof.* The result follows from the fact that the gradient random variable $\nabla F(\mathbf{w})$ is sub-exponential and therefore, its empirical estimate achieves fast convergence rates. We draw use $N$ samples $(\mathbf{x}^{(i)}, y^{(i)})$ from $\mathcal{D}$ and form the standard empirical estimate

$$\mathbf{g}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{m}(\sigma(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)})\mathbf{x}^{(i)}\,.$$

We first prove the estimation error in the dual norm of $\|\cdot\|_{\mathbf{w}}$. Recall that the dual norm is equal to $\|\mathbf{u}\|_{*,\mathbf{w}} = \max(\|\mathrm{proj}_{\mathbf{w}}\mathbf{u}\|_2\|\mathbf{w}\|_2^{3/2}, \|\mathrm{proj}_{\mathbf{w}^\perp}\mathbf{u}\|_2\|\mathbf{w}\|_2^{1/2})$. Recall that $\sigma'(\mathbf{w}\cdot\mathbf{x}) \ge 0$ and $\sigma'(\mathbf{w}\cdot\mathbf{x}) \le \xi e^{-\mu|\mathbf{w}\cdot\mathbf{x}|}$ for all $\mathbf{x}$. We first show that the distribution of $\mathbf{x}(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})$ is sub-exponential. Fix any unit direction $\mathbf{v}$, we show an upper bound on the tail probability $|\mathbf{v}\cdot\mathbf{x}| \ge t$. We first consider the case $\mathbf{v}\cdot\mathbf{w} = 0$ and

without loss of generality we may assume that $\mathbf{v}$ is parallel to $\mathbf{e}_1$ and $\mathbf{w}$ is parallel to $\mathbf{e}_2$. In the following calculations we repeatedly use the properties of $(L, R)$-well-behaved distributions and $(\tau, \mu, \xi)$-sigmoidal activations. We have

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}\{|\mathbf{v}\cdot\mathbf{x}(\sigma(\mathbf{w}\cdot\mathbf{x})-y)|\geq t\}\sigma'(\mathbf{w}\cdot\mathbf{x})] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}\{|\mathbf{v}\cdot\mathbf{x}|\geq t(\xi/\mu)\}\sigma'(\mathbf{w}\cdot\mathbf{x})] \\
&\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}\{|\mathbf{v}\cdot\mathbf{x}|\geq t(\xi/\mu)\}\xi e^{-\mu|\mathbf{w}\cdot\mathbf{x}|}] \\
&\leq \frac{1}{L}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathbb{1}\{|\mathbf{x}_1|\geq t(\mu/\xi)\}e^{-\mu\|\mathbf{w}\|_2|\mathbf{x}_2|}e^{-L\sqrt{\mathbf{x}_1^2+\mathbf{x}_2^2}}d\mathbf{x}_1 d\mathbf{x}_2 \\
&\leq \frac{1}{L}\int_{-\infty}^{\infty}\mathbb{1}\{|\mathbf{x}_1|\geq t(\xi/\mu)\}e^{-L|\mathbf{x}_1|/2}d\mathbf{x}_1\int_{-\infty}^{\infty}e^{-\mu\|\mathbf{w}\|_2|\mathbf{x}_2|}e^{-L|\mathbf{x}_2|/2}d\mathbf{x}_2 \\
&= \frac{2}{L}\frac{e^{-L/2t\mu/\xi}}{L/2}\frac{1}{\mu\|\mathbf{w}\|_2+L/2} \\
&\lesssim \frac{1}{L^2\mu(\|\mathbf{w}\|_2+1)}e^{-L\mu/(2\xi)\ t},
\end{aligned}
$$

where we used again the fact that $|\sigma(\mathbf{w}\cdot\mathbf{x})-y|\leq\xi/\mu$, the upper bound on the derivative of sigmoidal activations, see Definition 5.1.2 and the fact that the density of any 2 dimensional projection of $\mathcal{D}_{\mathbf{x}}$ is upper bounded by $(1/L)\exp(-L\|\mathbf{x}\|_2)$, see Definition 5.1.1. Moreover, using the same properties as above we have that the variance of $\nabla F(\mathbf{w})=\mathbf{x}(y-\sigma(\mathbf{w}\cdot\mathbf{x})\sigma'(\mathbf{w}\cdot\mathbf{x})$ along the direction $\mathbf{v}$ is at most

$$
\begin{aligned}
\mathbf{Var}[\nabla F(\mathbf{w})\cdot\mathbf{v}] &\leq \left(\frac{\xi}{\mu}\right)^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{v}\cdot\mathbf{x})^2] \\
&\leq \left(\frac{\xi}{\mu}\right)^2\frac{1}{L}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}e^{-L\sqrt{\mathbf{x}_1^2+\mathbf{x}_2^2}}\ \xi e^{-2\mu\|\mathbf{w}\|_2|\mathbf{x}_1|}\ d\mathbf{x}_1 d\mathbf{x}_2 \\
&\leq \frac{\xi^3}{\mu^2 L}\int_{-\infty}^{+\infty}\mathbf{x}_2^2 e^{-L|\mathbf{x}_2|}\ d\mathbf{x}_2\ \int_{-\infty}^{+\infty}e^{-2\mu\|\mathbf{w}\|_2|\mathbf{x}_1|}\ d\mathbf{x}_1 \\
&\lesssim \frac{\xi^3}{\mu^3 L^4}\frac{1}{\|\mathbf{w}\|_2}.
\end{aligned}
$$

Therefore, along any direction $\mathbf{v}$ orthogonal to $\mathbf{w}$ we have that $\nabla F(\mathbf{w})$ is $(\sigma_1^2, b_1)$-sub-exponential with variance proxy $\sigma_1^2 = \mathrm{poly}(\xi/(L\mu))1/\|\mathbf{w}\|_2$ and rate $b_1 = O(\xi/(L\mu))$. Therefore, Bernstein's inequality (see, e.g., [Ver18]) implies that the empirical estimate $g(\mathbf{w})$ with $N$ samples satisfies

$$\mathbf{Pr}\left[|g(\mathbf{w}) - \nabla F(\mathbf{w})) \cdot \mathbf{v}| \geq t\right] \leq 2e^{-\frac{Nt^2/2}{\sigma_1^2 + b_1 t}}.$$

We next analyze the gradient $\nabla F(\mathbf{w})$ along the direction of $\mathbf{w}$. We denote by $\widehat{\mathbf{w}}$ the unit vector that is parallel to $\mathbf{w}$. $\|\mathrm{proj}_{\mathbf{w}} g(\mathbf{w}) - \mathrm{proj}_{\mathbf{w}} \nabla F(\mathbf{w})\|_2 = \|\widehat{\mathbf{w}} \cdot \mathbf{g}(\mathbf{w})\widehat{\mathbf{w}} - \widehat{\mathbf{w}} \cdot \mathbf{g}(\mathbf{w})\widehat{\mathbf{w}}\|_2 = |\widehat{\mathbf{w}} \cdot \mathbf{g}(\mathbf{w}) - \widehat{\mathbf{w}} \cdot \mathbf{g}(\mathbf{w})|$. We first show the sub-exponential tail bound:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)| \geq t\}\sigma'(\mathbf{w}\cdot\mathbf{x})] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}| \geq t(\xi/\mu)\}\sigma'(\mathbf{w}\cdot\mathbf{x})] \\
&\leq \frac{1}{L}\int_{-\infty}^{\infty} \mathbb{1}\{|\mathbf{x}_2| \geq t(\mu/\xi)\} e^{-\mu\|\mathbf{w}\|_2|\mathbf{x}_2|} e^{-L|\mathbf{x}_2|} d\mathbf{x}_2 \\
&= \frac{1}{L(\mu\|\mathbf{w}\|_2 + L)} e^{-(\mu\|\mathbf{w}\|_2 + L)\mu/\xi t}.
\end{aligned}$$

Similarly, we can compute the variance along the direction of $\mathbf{w}$.

$$\begin{aligned}
\mathbf{Var}[\nabla F(\mathbf{w}) \cdot \widehat{\mathbf{w}}] &\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[((\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\widehat{\mathbf{w}}\cdot\mathbf{x})^2] \\
&\leq \left(\frac{\xi}{\mu}\right)^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\sigma'(\mathbf{w}\cdot\mathbf{x})\widehat{\mathbf{w}}\cdot\mathbf{x})^2] \\
&\leq \frac{\xi^3}{\mu^2 L}\int_{-\infty}^{+\infty} e^{-2\mu\|\mathbf{w}\|_2|t|} t^2 dt \\
&= \frac{\xi^3}{2\mu^5 L}\frac{1}{\|\mathbf{w}\|_2^3},
\end{aligned}$$

where we used the fact that $|\sigma(\mathbf{w}\cdot\mathbf{x}) - y| \leq \xi/\mu$, the upper bound on the derivative of sigmoidal activations, see Definition 5.1.2 and the anti-concentration property of

$(L, R)$-well-behaved distributions, see Definition 5.1.1.

Therefore, along the direction $\mathbf{w}$ we have that $\nabla F(\mathbf{w})$ is $(\sigma_2^2, b_2)$-sub-exponential with variance proxy $\sigma_2^2 = \mathrm{poly}(\xi/(L\mu))1/\|\mathbf{w}\|_2^3$ and rate $b_1 = O(\xi/(L\mu^2\|\mathbf{w}\|_2))$. Therefore, Bernstein's inequality (see, e.g., [Ver18]) implies that the empirical estimate $g(\mathbf{w})$ with $N$ samples, satisfies

$$\mathbf{Pr}\left[|(g(\mathbf{w}) - \nabla F(\mathbf{w})) \cdot \widehat{\mathbf{w}}| \geq t\right] \leq 2e^{-\frac{Nt^2/2}{\sigma_2^2 + b_2 t}}.$$

Thus, we can now perform a union bound on every direction $\mathbf{u}$ and every $\|\mathbf{w}\| \leq B$ by using a $r = \mathrm{poly}(L\mu)/(\xi B)\epsilon')$-net of the unit sphere and the ball of radius $B$ (which will have size $r^O(d)$, see e.g., [Ver18]), we obtain that along any direction $\mathbf{v}$ orthogonal to $\mathbf{w}$ it holds that $|\mathbf{v}\cdot(\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w}))| \leq \epsilon/\|\mathbf{w}\|_2^{1/2}|$ with probability at least $1 - 2r^{O(d)}e^{-\frac{N\epsilon^2/2}{\|\mathbf{w}\|_2\sigma_1^2 + \|\mathbf{w}\|_2^{1/2} b_1 \epsilon}}$. Substituting the variance and rate values $\sigma_1^1, b_1$, we observe that $\|\mathbf{w}\|_2\sigma_1^2 + \|\mathbf{w}\|_2^{1/2} b_1 t \leq \mathrm{poly}(\xi/(L\mu))$, where we used our assumption that $\epsilon' \leq 1/\sqrt{B}$ (and that $\|\mathbf{w}\|_2 \leq B$). Picking $N = \widetilde{O}(d/\epsilon'^2 \mathrm{poly}(\xi/(L\mu))\log(B/\delta))$ we obtain that the above bound holds with probability at least $1 - \delta/2$. Similarly, performing the same union bound for the direction of $\mathbf{w}$ and using the sub-exponential bound with $\sigma_2^2, b_2$ and the fact that $\epsilon' \leq 1/\sqrt{B}$ we again obtain that $|\mathbf{v}\cdot(\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w}))| \leq \epsilon'/\|\mathbf{w}\|_2^{3/2}$ with probability at least $1 - \delta/2$ with $N = \widetilde{O}(d/\epsilon'^2 \mathrm{poly}(\xi/(L\mu))\log(B/\delta))$ samples. The bound for the $\ell_2$ distance of $\mathbf{g}(\mathbf{w})$ and $\nabla F(\mathbf{w})$ follows similarly from the above concentration bounds. □

Using Claim 5.4.6, we can estimate all the gradients with probability $1 - \delta'$ and accuracy $\epsilon'$, with $N = \widetilde{O}((d/\epsilon'^2)\mathrm{poly}(\xi/(L\mu))\log(B/\delta'))$ samples, for some parameters $\epsilon', \delta > 0$ that we will choose below. We now show that the empirical gradients will also

satisfy the required gradient field assumptions of Lemma 5.4.9. Assume that we have $\|\nabla \widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \epsilon'$ and $\|\nabla \widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})\|_{*,\mathbf{w}} \leq \epsilon'$ for some $\epsilon'$ to be specified later. Using the triangle inequality we have that

$$\|\nabla \widehat{F}(\mathbf{w})\|_2 \leq \|\nabla F(\mathbf{w})\|_2 + \epsilon' \,.$$

Therefore, for $\epsilon' \leq 1/\kappa$ we obtain that $\|\nabla \widehat{F}(\mathbf{w})\|_2 = O(1/\kappa) \leq O(B)$. We next show that the empirical gradient also points to the direction of $\mathbf{w} - \mathbf{w}^*$, i.e., satisfies the assumptions of Lemma 5.4.3. For the case where $\|\mathbf{w}\| \leq Z_0$ we have that

$$\nabla \widehat{F}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{v}) \geq \nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{v}) - \epsilon' \|\mathbf{w} - \mathbf{v}\|_2 \geq (\alpha_2 - \epsilon') \|\mathbf{w} - \mathbf{v}\|_2 \,. \tag{5.8}$$

Therefore, we need to choose $\epsilon' < \alpha_2$. Similarly, for the case where $Z_0 < \|\mathbf{w}\| \leq Z_1$ we have that

$$\begin{aligned} \nabla \widehat{F}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{v}) &\geq \nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{v}) - (\nabla \widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})) \cdot (\mathbf{w} - \mathbf{v}) \\ &\geq \beta_2 \|\mathbf{w} - \mathbf{v}\|_{\mathbf{w}} - \|\nabla \widehat{F}(\mathbf{w}) - \nabla F(\mathbf{w})\|_{*,\mathbf{w}} \|\mathbf{w} - \mathbf{v}\|_{\mathbf{w}} \\ &\geq (\beta_2 - \epsilon') \|\mathbf{w} - \mathbf{v}\|_{\mathbf{w}} \,. \end{aligned} \tag{5.9}$$

Finally, for $\|\mathbf{w}\|_2 \geq \zeta Z_1$ we similarly get the lower bound

$$\nabla \widehat{F}(\mathbf{w}) \cdot \mathbf{w} \geq (\gamma - \epsilon') \|\mathbf{w}\|_{\mathbf{w}} \,. \tag{5.10}$$

Therefore, it suffices to choose $\epsilon' \leq \min(\alpha_2, \beta_2, \gamma)/2 = (\sqrt{\epsilon})\text{poly}(LR\mu\tau/\xi)$. Assuming that all the empirical gradients used by the gradient descent satisfy the

error bound with $\epsilon'$ as above, from Lemma 5.4.3, we obtain that with step size $\eta = \text{poly}(\epsilon\tau\mu LR/\xi)$ after $T = \text{poly}(1/(\epsilon\eta))$ iterations, we will have that if $\|\mathbf{w}^{(T)}\|_2 \leq Z_0$, then $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq \text{poly}(\xi/(LR\tau\mu))\sqrt{\epsilon}$ which using Lemma 5.2.21, implies that $F(\mathbf{w}^{(T)}) \leq \text{poly}(\xi/(LR\mu\tau)(\epsilon)$. Similarly, if $\|\mathbf{w}^{(T)}\|_2 > Z_0$, we obtain that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_{\mathbf{w}^{(T)}} \leq \text{poly}(\xi/(LR\tau\mu))\sqrt{\epsilon}$ which again implies that $F(\mathbf{w}^{(T)}) \leq \text{poly}(\xi/(LR\mu\tau))\epsilon$.

### 5.4.2 Unbounded Activations

We now turn our attention to the case of unbounded activations. Perhaps surprisingly, we show that essentially the same algorithm (the only difference is that in this case we set the regularization parameter to 0) achieves sample complexity polylogarithmic in $1/\epsilon$.

**Theorem 5.4.7** (Unbounded Activations)**.** *Let $\mathcal{D}$ be an $(\epsilon, W)$-corrupted, $(L, R)$-well-behaved distribution on $\mathbb{R}^d \times \mathbb{R}$ and $\sigma(\cdot)$ be an $(\alpha, \lambda)$-unbounded activation. Set $\kappa = \text{poly}(LR\alpha/\lambda)/(W^2 \log(W))$. The gradient descent Algorithm 9 with step size $\eta = \kappa\epsilon$, truncation threshold $M = \widetilde{O}((W/L)\max(\log(\lambda^2 W^2/\epsilon), 1))$, $N = \widetilde{\Theta}(d/\kappa \log(1/\delta)\max(\text{poly}\log(1/\epsilon), 1))$ samples, and $T = \text{poly}(\log(1/\epsilon), 1/\kappa)$ iterations converges to a vector $\mathbf{w}^{(T)} \in \mathbb{R}^d$ that, with probability $1 - \delta$, satisfies $F(\mathbf{w}^{(T)}) \lesssim \frac{1}{LR^4}\left(\frac{\lambda}{\alpha}\right)^4 \epsilon$.*

#### Proof of Theorem 5.4.7

We consider the population loss $F(\mathbf{w}) = (1/2)\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2]$. As we discussed in Remark 5.2.3, we can assume that $\lambda \geq 1$ and $L \leq 1$ for making the presentation simpler. Moreover, we can also assume that $W \geq 1$, because if $W < 1$, then $\inf_{\|\mathbf{w}\|_2 \leq 1} F(\mathbf{w}) \leq \inf_{\|\mathbf{w}\|_2 \leq W} F(\mathbf{w})$, therefore the distribution is $(\epsilon, 1)$-corrupted. Denote $\mathbf{w}^*$ to be a vector with $\|\mathbf{w}^*\|_2 \leq W$ that achieves $F(\mathbf{w}^*) \leq \epsilon$. First, we assume that $\epsilon \geq \lambda^2 W^2/C$, for

**Algorithm 9** Gradient Descent Algorithm for Optimizing the $L_2^2$-Loss for Unbounded Activations

**Input:** number of iterations $T$, sample access to $\mathcal{D}$, step size $\eta$, bound $M$
$\mathbf{w}^{(0)} \leftarrow \mathbf{0}$
**for** $t = 0$ **to** $T$ **do**
    Use fresh $N$ samples from $\mathcal{D}$ truncated at $M$ to obtain $\mathbf{g}^{(t)}$, an approximation of the population gradient $\nabla F(\mathbf{w}^{(t)})$ (see Claim 5.4.10)
    $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta\, \mathbf{g}^{(t)}$
**end for**
**Return:** $\mathbf{w}^{(T)}$

some $C \geq 10$. Then, any solution $\mathbf{w} \in \mathbb{R}^d$ achieves error $F(\mathbf{w}) \leq 2\epsilon$. To see this, note that

$$F(\mathbf{w}) \leq 2F(\mathbf{w}^*) + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] \leq 2\epsilon + 4\lambda^2 W^2 \leq 2C\epsilon\ .$$

Next we consider the case where $\epsilon \leq \lambda^2 W^2/C$. First, using the exponential concentration properties of well-behaved distributions, we observe that we can truncate the labels $y$ such that $|y| \leq M$, for some $M > 0$, without increasing the level of corruption by a lot. Given the exponential concentration of the distribution and the fact that $\|\mathbf{w}^*\|_2 \leq W$, we show that we can pick $M = \Theta(W \max(\log(W/\epsilon), 1))$ so that the instance $\widetilde{D}$ of $(\mathbf{x}, \widetilde{y})$, where $\widetilde{y} = \mathrm{sign}(y)\min(|y|, M)$, is at most $O(\epsilon)$-corrupted. Formally, we show:

**Claim 5.4.8.** *Let $M = \Theta(W \max(\log(W/\epsilon), 1))$. Denote by $\mathrm{tr}(y) = \mathrm{sign}(y)\min(|y|, M)$. Then, it holds that:*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{tr}(y) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2] \leq O(\epsilon)\ .$$

*Proof.* Denote $\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{w}^*\cdot\mathbf{x}| \leq M\}$, i.e., the points $\mathbf{x}$ such that $|\mathbf{w}^*\cdot\mathbf{x}|$ is at most $M$. We have that for $\mathbf{x} \in \mathcal{B}$, it holds that $|\tilde{y} - \sigma(\mathbf{w}^*\cdot\mathbf{x})| \leq |y - \sigma(\mathbf{w}^*\cdot\mathbf{x})|$, where

$\tilde{y}$ is equal to $y$ truncated in $-M \leq y \leq M$. Therefore, we have that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\tilde{y} - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{\mathbf{x} \in \mathcal{B}\}] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{\mathbf{x} \in \mathcal{B}\}] \leq \epsilon \,.$$

Using the triangle inequality, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\tilde{y} - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\tilde{y} - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{\mathbf{x} \in \mathcal{B}\}] + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\tilde{y} - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{\mathbf{x} \notin \mathcal{B}\}] \\
&\leq \epsilon + 2M^2 \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x} \notin \mathcal{B}] + 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\sigma(\mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{\mathbf{x} \notin \mathcal{B}\}] \,.
\end{aligned}$$

To bound $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\sigma(\mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{\mathbf{x} \notin \mathcal{B}\}]$, note that $\|\mathbf{w}^*\|_2 \leq W$ and that $|\sigma(\mathbf{w}^* \cdot \mathbf{x})| \leq \lambda|\mathbf{w}^* \cdot \mathbf{x}|$. Assuming without loss of generality that $\mathbf{w}^*$ is parallel to $\mathbf{e}_1$ and let $V$ be the subspace spanned by $\mathbf{w}^*$ and any other vector orthogonal to $\mathbf{w}^*$. Using that the distribution $\mathcal{D}_\mathbf{x}$ is $(L, R)$-well-behaved, we have that it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\sigma(\mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{\mathbf{x} \notin \mathcal{B}\}] \leq \lambda^2 W^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x}_1^2 \mathbb{1}\{\mathbf{x}_1 \geq M/W\}] \leq \epsilon/2 \,,$$

which follows from $\lambda^2 \,\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{x}_1|^2 \mathbb{1}\{\mathbf{x}_1 \geq M/W\}] \leq \lambda^2 (M^2/(L^2W^2)) \exp(-LM/W) \leq \epsilon/2$, its proof is similar with the proof of Claim 5.2.10. It remains to bound the $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x} \notin \mathcal{B}]$, we have

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x} \notin \mathcal{B}] = \int_{|\mathbf{w}^* \cdot \mathbf{x}| \notin \mathcal{B}} \gamma_V(\mathbf{x}) \mathrm{d}\mathbf{x} \leq \frac{2}{L} \int_{t \geq M/\|\mathbf{w}^*\|_2} e^{-Lt} \mathrm{d}t = \frac{2}{L^2} e^{-LM/\|\mathbf{w}^*\|_2} \leq \frac{\epsilon}{2M^2} \,.$$

Hence, $(1/2)\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\tilde{y} - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] \lesssim \epsilon$. Therefore, by truncated the $y$ to values $-M \leq y \leq M$, we increase the total error at most $\epsilon' = O(\epsilon)$. For the rest of the proof

for simplicity we abuse the notation of the symbol $y$ for the $\tilde{y}$ and $\epsilon$ for the $\epsilon'$. $\square$

Thus, we will henceforth assume that $|y| \leq M$, and keep denoting the instance distribution as $\mathcal{D}$.

We first show that, given any gradient field that satisfies the properties given in Proposition 5.3.4, then gradient descent with an appropriate step size will converge to the target vector $\mathbf{v}$.

**Lemma 5.4.9** (Gradient Field Distance Reduction)**.** *Let $\mathbf{g} : \mathbb{R}^d \mapsto \mathbb{R}^d$ be a vector field and fix $W, B \geq 1$. We assume that $\mathbf{g}$ satisfies the following properties with respect to some unknown target vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 \leq W$. Fix parameters $\alpha_1, \alpha_2 > 0$. For every $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \leq 2W$ it holds that $\|g(\mathbf{w})\|_2 \leq B \max(\|\mathbf{w} - \mathbf{v}\|_2, \alpha_1)$. Moreover, if $\|\mathbf{w} - \mathbf{v}\|_2 \geq \alpha_1$ and $\theta(\mathbf{w}, \mathbf{v}) \in (0, \pi/2)$ then it holds that $\mathbf{g}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{v}) \geq \alpha_2 \|\mathbf{w} - \mathbf{v}\|_2^2$. We consider the update rule $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \mathbf{g}(\mathbf{w}^{(t)})$ initialized with $\mathbf{w}^{(0)} = \mathbf{0}$ and step size $\eta \leq \alpha_2/B$. Let $T$ be any integer larger than $\lceil (W^2 + \log(1/\alpha_1))/(\eta\alpha_2) \rceil$, it holds that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq (1 + \eta B)\alpha_1$.*

*Proof.* We first show that at every round the distance of $\mathbf{w}^{(t)}$ and the target $\mathbf{v}$ decreases. Since we initialize at $\mathbf{w} = \mathbf{0}$ (notice that in this case it holds trivially $\|\mathbf{w}^{(0)}\|_2 \leq W$) after the first update we have

$$\begin{aligned}
\|\mathbf{w}^{(1)} - \mathbf{v}\|_2^2 - \|\mathbf{w}^{(0)} - \mathbf{v}\|_2^2 &= -2\eta \mathbf{g}(\mathbf{w}^{(0)}) \cdot (\mathbf{w}^{(0)} - \mathbf{v}) + \eta^2 \|\mathbf{g}(\mathbf{w}^{(0)})\|_2^2 \\
&\leq -2\eta\alpha_2 \|\mathbf{w}^{(0)} - \mathbf{v}\|_2^2 + \eta^2 B^2 \|\mathbf{w}^{(0)} - \mathbf{v}\|_2^2 \\
&\leq -\eta\alpha_2 \|\mathbf{w}^{(0)} - \mathbf{v}\|_2^2 ,
\end{aligned}$$

where we used that since $\eta \leq \alpha_2/B$ it holds that $\eta\alpha_2 - \eta^2 B^2 \geq 0$. Moreover, we observe that after the first iteration we are going to have $\mathbf{w}^{(1)} \cdot \mathbf{v} \geq 0$ (since otherwise the distance

to $\mathbf{v}$ would increase). This implies that $\theta(\mathbf{w}^{(1)}, \mathbf{v}) \in (0, \pi/2)$. Observe now that it holds $\|\mathbf{w}^{(1)}\|_2 \leq 2W$ since $\mathbf{w}^{(1)}$ is closer to $\mathbf{v}$ than $\mathbf{w}^{(0)}$. Using induction, similarly to the previous case we can show that for all iterations $t$ where $g(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^{(t)} - \mathbf{v}) \geq \alpha_2 \|\mathbf{w}^{(t)} - \mathbf{v}\|_2^2$ it holds $\|\mathbf{w}^{(t+1)}\|_2 \leq 2W$ and $\theta(\mathbf{w}^{(t+1)}, \mathbf{v}) \in (0, \pi/2)$ and that the distance to the target vector $\mathbf{v}$ decreases at every iteration:

$$\|\mathbf{w}^{(t+1)} - \mathbf{v}\|_2^2 \leq \|\mathbf{w}^{(t)} - \mathbf{v}\|_2^2 (1 - \eta\alpha_2) \leq \|\mathbf{w}^{(0)} - \mathbf{v}\|_2^2 (1 - \eta\alpha_2)^t \,.$$

Let $T'$ be the first iteration where the assumption $g(\mathbf{w}^{(T')}) \cdot (\mathbf{w}^{(T')} - \mathbf{v}) \geq \alpha_2 \|\mathbf{w}^{(T')} - \mathbf{v}\|_2^2$ does not hold. Since at each round we decrease the distance of $\mathbf{w}^{(t)}$ and $\mathbf{v}$, and we initialize at $\mathbf{0}$ we have that the number of iterations $T' \leq \lceil (\|\mathbf{v}\|_2^2 + \log(1/\alpha))/(\eta\alpha_2) \rceil$. Moreover, even if we continue to use the update rule after the iteration $T'$ from the triangle inequality we have that

$$\|\mathbf{w}^{(T'+1)} - \mathbf{v}\|_2 \leq \|\mathbf{w}^{(T'+1)} - \mathbf{w}^{(T')}\|_2 + \|\mathbf{w}^{(T')} - \mathbf{v}\|_2 \leq \eta \|\mathbf{g}(\mathbf{w}^{(T')})\|_2 + \alpha_1 \leq (1 + \eta B)\alpha_1 \,.$$

Therefore, by induction we get that for all $T \geq T'$ it holds that $\|\mathbf{w}^{(T)} - \mathbf{v}\|_2 \leq (1 + \eta B)\alpha_1$. □

We first show that Algorithm 9 with the population gradients would converge after polylog$(1/\epsilon)$ iterations to an approximately optimal solution. We have

$$\|\nabla F(\mathbf{w})\|_2 = \|\, \mathbf{E}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})\mathbf{x}]\|_2 = \max_{\|\mathbf{u}\|_2 = 1} \mathbf{E}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)\sigma'(\mathbf{w} \cdot \mathbf{x})\mathbf{u} \cdot \mathbf{x}] \,.$$

Let $\mathbf{w}$ be any vector with $\|\mathbf{w}\|_2 \leq W$ and let $\mathbf{u}$ be any unit vector. By adding and

subtracting $\sigma(\mathbf{w}^* \cdot \mathbf{x})$, we get:

$$
\begin{aligned}
&\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{u}\cdot\mathbf{x}] \\
&= \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{u}\cdot\mathbf{x}] + \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{u}\cdot\mathbf{x}] \\
&\le \lambda\,\mathbf{E}[|(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)||\mathbf{u}\cdot\mathbf{x}|] + \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{u}\cdot\mathbf{x}] \\
&\le \lambda\sqrt{\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2]\,\mathbf{E}[(\mathbf{u}\cdot\mathbf{x})^2]} + \sqrt{\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2\sigma'(\mathbf{w}\cdot\mathbf{x})^2]\,\mathbf{E}[(\mathbf{u}\cdot\mathbf{x})^2]} \\
&\le \lambda\sqrt{\epsilon} + \lambda\|\mathbf{w} - \mathbf{w}^*\|_2\,,
\end{aligned}
$$

where for the third inequality we used the fact the $\mathbf{x}$-marginal of $\mathcal{D}$ is isotropic and that $\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2]^{1/2} \le \sqrt{\epsilon}$. Hence, we have that $\|\nabla F(\mathbf{w})\|_2 \le \lambda(\sqrt{\epsilon} + \|\mathbf{w} - \mathbf{w}^*\|_2)$. Using Proposition 5.3.4, we have that the true gradients point in the right direction: we have that for any vector $\mathbf{w} \in \mathbb{R}^d$, with $\theta(\mathbf{w}, \mathbf{w}^*) \in (0, \pi/2)$, it holds that if $\|\mathbf{w} - \mathbf{w}^*\|_2 \ge c'\lambda\sqrt{\epsilon}/(LR^4\alpha^2)$, for some absolute constant $c' > 0$, then it holds that $\nabla F(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \ge (1/2)\alpha^2 LR^4\|\mathbf{w} - \mathbf{w}^*\|_2^2$. Therefore, the true gradient field of $F$ satisfies the assumptions of Lemma 5.4.9 with $B = O(\lambda)$, $\alpha_1 = c'\sqrt{\epsilon}/(LR^4)\lambda/\alpha^2$ and $\alpha_2 = (1/2)\alpha^2 LR^4$.

In the following claim, we show that with roughly $\widetilde{O}(d\|\mathbf{w} - \mathbf{w}^*\|_2^2/\epsilon^2)$ samples in each iteration, we can get $\epsilon'$-estimates of the population gradients.

**Claim 5.4.10.** *Let* $\mathbf{w}$ *be such that* $\|\mathbf{w} - \mathbf{w}^*\| \ge \sqrt{\epsilon}$. *Using* $N = \widetilde{O}(d\lambda^4\|\mathbf{w} - \mathbf{w}^*\|_2^2 W^2 \max(\log^2(1/\epsilon), 1)/(L\epsilon')^2)$ *samples, we can compute an empirical gradient* $\mathbf{g}(\mathbf{w})$ *such that* $\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \le \epsilon$ *with probability* $1 - \delta$.

*Proof.* We start by bounding from above the variance in every direction. Let $\Sigma_i =$

$\mathbf{E}[((\sigma(\mathbf{w}\cdot\mathbf{x})-y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}_i)^2]$, we have that

$$\begin{aligned}\Sigma_i = \mathbf{E}[((\sigma(\mathbf{w}\cdot\mathbf{x})-y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}_j)^2] &\leq \lambda^2\,\mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2\mathbf{x}_j^2] \\ &\leq 2\lambda^2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\mathbf{x}_j^2] + 2\lambda^2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-\sigma(\mathbf{w}\cdot\mathbf{x}))^2\mathbf{x}_j^2]\,. \\ &\leq 2\lambda^2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\mathbf{x}_j^2] + 2\lambda^4\|\mathbf{w}^*-\mathbf{w}\|_2^2 \max_{\|\mathbf{u}\|_2=1}\mathbf{E}[(\mathbf{u}\cdot\mathbf{x})^2\mathbf{x}_j^2]\,.\end{aligned}$$

To bound the term $\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\mathbf{x}_j^2]$, we have that

$$\begin{aligned}&\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\mathbf{x}_j^2] \\ &= \mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\mathbf{x}_j^2\mathbb{1}\{\mathbf{x}_j\leq M\}] + \mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\mathbf{x}_j^2\mathbb{1}\{\mathbf{x}_j\geq M\}] \\ &\leq M^2\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2] + 4M^2\,\mathbf{E}[\mathbf{x}_j^2\mathbb{1}\{\mathbf{x}_j\geq M\}] \\ &\lesssim M^2\epsilon + 4M^2\,\mathbf{E}[\mathbf{x}_j^2\mathbb{1}\{\mathbf{x}_j\geq M'\}] \lesssim \epsilon M^2\,.\end{aligned}$$

Moreover, note that $\sqrt{\epsilon}\leq\|\mathbf{w}-\mathbf{w}^*\|_2$, hence $\Sigma_i\lesssim\lambda^4\|\mathbf{w}-\mathbf{w}^*\|_2^2M^2$. From Markov's inequality, we have that for each $j\leq d$, it holds

$$\begin{aligned}&\mathbf{Pr}\left[\left|\sum_{i=1}^N\frac{1}{N}(\sigma(\mathbf{w}\cdot\mathbf{x}^{(i)})-y^{(i)})\sigma'(\mathbf{w}\cdot\mathbf{x}^{(i)})\mathbf{x}_j^{(i)} - \mathbf{E}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}_j]\right|\geq\epsilon'/\sqrt{d}\right] \\ &\leq \frac{d}{N\epsilon'^2}\,\mathbf{E}[((\sigma(\mathbf{w}\cdot\mathbf{x})-y)\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}_j)^2] \leq \frac{d\|\mathbf{w}-\mathbf{w}^*\|_2^2\lambda^4M^2}{N\epsilon'^2}\,.\end{aligned}$$

Hence, with $N=O(\frac{d\lambda^4\|\mathbf{w}-\mathbf{w}^*\|_2^2M^2}{\epsilon'^2})$ samples, we can get an $\epsilon'/\sqrt{d}$-approximation to the $i$-th coordinate of the gradient $(\nabla F(\mathbf{w}))_i$ with constant probability, and by using a standard boosting procedure we can boost the probability to $1-\delta$ with a multiplicative overhead of $O(\log(1/\delta))$ samples. Finally, doing a union bound over all coordinates $j\in\{1,\ldots,d\}$ we obtain that $N=\widetilde{O}(d\lambda^4\|\mathbf{w}-\mathbf{w}^*\|_2^2W^2\max(\log^2(1/\epsilon),1)/(L^2\epsilon'^2))$

samples suffice. □

We now show that the empirical gradients will also satisfy the required gradient field assumptions of Lemma 5.4.9. Assume that we have $\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \epsilon'$ for some $\epsilon'$ to be specified later. Using the triangle inequality, we get for any $\mathbf{w}$ with $\|\mathbf{w}\| \leq 2W$ it holds that

$$\|\mathbf{g}(\mathbf{w})\|_2 \leq \|\nabla F(\mathbf{w})\|_2 + \epsilon' .$$

Therefore, for $\epsilon' \lesssim \lambda(\sqrt{\epsilon} + \|\mathbf{v} - \mathbf{w}\|_2)$ we obtain that $\|\mathbf{g}(\mathbf{w})\|_2 = O(\lambda(\sqrt{\epsilon} + \|\mathbf{v} - \mathbf{w}\|_2))$. We next show that the empirical gradient also points to the direction of $\mathbf{w} - \mathbf{w}^*$. It holds that

$$\mathbf{g}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \geq \nabla F(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) - \epsilon'\|\mathbf{w}-\mathbf{w}^*\|_2 \geq ((1/2)\alpha^2 L R^4 \|\mathbf{w}-\mathbf{w}^*\|_2 - \epsilon')\|\mathbf{w}-\mathbf{w}^*\|_2 .$$

Therefore, we need to choose $\epsilon' = O(\alpha^2 L R^4 \|\mathbf{w} - \mathbf{w}^*\|_2)$. Using that $\|\mathbf{w} - \mathbf{w}^*\|_2 \geq c'\lambda\sqrt{\epsilon}/(LR^4\alpha^2)$, we have that the estimated gradient field of $F$ satisfies the assumptions of Lemma 5.4.9 with $B = O(\lambda)$, $\alpha_1 = c'\sqrt{\epsilon}/(LR^4)\lambda/\alpha^2$ and $\alpha_2 = (1/2)\alpha^2 L R^4$. Conditioning that all the empirical gradients used by the gradient descent satisfy the error bound with $\epsilon'$ as above from Lemma 5.4.9 we obtain that with step size $\eta = \mathrm{poly}(LR\alpha/\lambda)$ after $T = \mathrm{poly}(1/(\alpha LR))W^2 \log(1/\alpha_1)$ iterations we will have that $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq \mathrm{poly}(1/(LR))\lambda/\alpha^2\sqrt{\epsilon}$ which implies that

$$\begin{aligned} F(\mathbf{w}^{(T)}) &\leq 2\,\mathbf{E}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2] + 2\,\mathbf{E}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - \sigma(\mathbf{w}^{(T)} \cdot \mathbf{x}))^2] \\ &= \mathrm{poly}(1/(LR))\ (\lambda/\alpha)^4\ O(\epsilon), \end{aligned}$$

where we used the fact that $\sigma$ is $\lambda$-Lipschitz and that the $\mathbf{x}$-marginal of $\mathcal{D}$ is isotropic.

Since we have to do a union bound over all $T$ iterations, and we draw fresh samples in each round, we need to divide $\delta$ by $T$ in each round; hence, the total sample complexity is $N = \text{poly}(\lambda/(\alpha LR))\ \widetilde{O}(dW^2 \log^3(1/\epsilon) \log(1/\delta))$ and the runtime $\text{poly}(\lambda/(\alpha LR)) \log(1/\epsilon) N$.

# Chapter 6

# Robustly Learning Monotone SIMs under Well-Behaved-Distributions

## 6.1 Introduction

This chapter addresses the challenging problem of agnostically learning single-index models (SIMs) where the monotone activation function is unknown and answering the following question:

*Can we obtain efficient* constant factor *learners for Problem 1.3.22 under mild distributional assumptions?*

Building on the setting of Chapter 5, we remove the assumption of a known activation (assumining that the unknown activation belongs to the class of Definition 5.1.5), providing the first efficient and provably robust algorithm that achieves a constant-factor approximation guarantee under well-behaved distributions (Definition 5.1.1).

A simplified version of our main algorithmic result is as follows (see Theorem 6.3.2

for a more detailed statement):

**Theorem 6.1.1** (Main Algorithmic Result, Informal)**.** *Given Problem 1.3.22, where $\mathcal{G}$ is the class of $(L,R)$-well behaved distributions with $L,R=O(1)$ and $\mathcal{F}=\mathcal{U}_{(a,b)}$ such that $(1/a),b=O(1)$, there is an algorithm that draws $N=\mathrm{poly}(W)\tilde{O}(d/\epsilon^2)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(N,d)$ time, and outputs a hypothesis $\hat{u}(\widehat{\mathbf{w}}\cdot\mathbf{x})$ with $\hat{u}\in\mathcal{U}_{(a,b)}$, $\|\widehat{\mathbf{w}}\|_2\le W$ such that $\mathcal{L}_2(\widehat{\mathbf{w}};\hat{u})=C\,\mathrm{OPT}+\epsilon$ with high probability, where $C>0$ is an absolute constant.*

We reiterate that the approximation factor $C$ in Theorem 6.1.1 is a universal constant, independent of the dimension and the diameter of the space. That is, our main result provides the first efficient learning algorithm achieving a constant factor approximation, even for the most basic case of Gaussian data and any non-trivial class of link functions.

**Technical Comparison to [GGKS23]** The only prior work addressing SIM learning (with unknown link functions) in the agnostic model is [GGKS23], thus here we provide a technical comparison. While both [GGKS23] and our work make use of the surrogate loss function from (1.2), on a technical level the two works are completely disjoint. [GGKS23] uses a framework of omnipredictors to minimize the surrogate loss and then relates this result to the $L_2^2$ loss. Although they handle more general distributions and activations, their learner outputs a hypothesis with error that cannot be considered constant factor approximation (see (1.1)) and is improper. By contrast, our work does not seek to minimize the surrogate loss. Instead, our main insight is that the gradient of the surrogate loss at a vector $\mathbf{w}$ conveys information about the direction of a target vector $\mathbf{w}^*$, for a *fixed* link function that minimizes *the* $L_2^2$ *loss*. We leverage this property to construct a proper learner achieving constant factor approximation.

### 6.1.1 Local Error Bounds and Alignment Sharpness

Given a generic optimization problem $\min_{\mathbf{w}} f(\mathbf{w})$ and a non-negative residual function $r(\mathbf{w})$ measuring the approximation error of the optimization problem, we say that the problem satisfies a local error bound if in some neighborhood of "test" (typically optimal) solutions $\mathcal{W}^*$ we have that

$$r(\mathbf{w}) \geq (\mu/\nu)\,\mathrm{dist}(\mathbf{w}, \mathcal{W}^*)^{\nu}. \tag{6.1}$$

In other words, low value of the residual function implies that $\mathbf{w}$ must be close to the test set $\mathcal{W}^*$.

Local error bounds have been studied in the optimization literature for decades, starting with the seminal works of [Hof52, Łoj63]; see, e.g., Chapter 6 in [FP03] for an overview of classical results and [BNPS17, KNS16, Rd17, LCP22] and references therein for a more cotemporary overview. While local error bounds can be shown to hold generically under fairly minimal assumptions on $f$ and for $r(\mathbf{w}) = f(\mathbf{w}) - \min_{\mathbf{w}'} f(\mathbf{w}')$ [Łoj63, Łoj93], it is rarely the case that they can be ensured to hold with a parameter $\mu$ that is not trivially small.

On the other hand, learning problems often possess very strong structural properties that can lead to stronger local error bounds. There are two main such examples we are aware of, where local error bounds can be shown to hold with $\nu = 2$ and an absolute constant $\mu > 0$. The first example are low-rank matrix problems such as matrix completion and matrix sensing, which are unrelated to our work [BNS16, ZL16, JGN+17]. More relevant to our work is the recent result in [WZDD23], which proved a local error

bound of the form

$$r(\mathbf{w}) \geq \frac{\mu}{2}\text{dist}(\mathbf{w}, \mathcal{W}^*)^2 \tag{6.2}$$

for the more restricted problem than ours (with a known activation function) but under somewhat more general distributional assumptions. In [WZDD23], the residual function was defined by $r(\mathbf{w}^t) = \nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; u^*) \cdot (\mathbf{w}^t - \mathbf{w}^*)$, where $\nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; u^*)$ is the gradient of an empirical surrogate loss, and the resulting local error bound referred to as "sharpness."[1]

Our structural result can be seen as a weak notion of a local error bound, where the residual function for the empirical surrogate loss expressed as $r(\mathbf{w}^t, \hat{u}^t) = \nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*)$ is bounded below as a function of the magnitude of the component of $\mathbf{w}^*$ that is orthogonal to $\mathbf{w}^t$. Compared to more traditional local error bounds and the bound from [WZDD23], which bound below the residual error function as a function of the distance to $\mathcal{W}^*$, this is a much weaker local error bound since it does not distinguish between vectors of varying magnitudes along the direction of $\mathbf{w}^*$. Since our lower bound is related to the "sharpness" notion studied in [WZDD23], we refer to it as the "alignment sharpness" to emphasize that it only relates the misalignment (as opposed to the distance) of vectors $\mathbf{w}^t$ and $\mathbf{w}^*$ to the residual error. To the best of our knowledge, such a form of a local error bound, which only bounds the alignment of vectors as opposed to their distance, is novel. We expect it to find a more broader use in learning theory and optimization.

[1] A local error utilizing the same type of a residual was introduced in [ZY13] under the name "restricted secant inequality."

## 6.2 Main Structural Result: Alignment Sharpness of Surrogate Loss

In this section, we establish our main structural result (Proposition 6.2.1), which is what crucially enables us to obtain the target $O(\mathrm{OPT}) + \epsilon$ error for the studied problem. Proposition 6.2.1 states that the empirical gradient of the surrogate loss (1.2) positively correlates with the direction of $\mathbf{w}^t - \mathbf{w}^*$ whenever $\mathbf{w}^t$ does not correspond to an $O(\mathrm{OPT}) + \epsilon$ error solution; and, moreover, the correlation is proportional to the quantity $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$. This is a key property that is leveraged in our algorithmic result (Theorem 6.3.2), both in obtaining an $O(\mathrm{OPT}) + \epsilon$ error result, and in arguing about the convergence and computational efficiency of our algorithm.

Intuitively, what Proposition 6.2.1 allows us to argue is that as long as the angle between $\mathbf{w}^t$ and $\mathbf{w}^*$ is not close to zero, we can update $\mathbf{w}^t$ to better align it with $\mathbf{w}^*$ (in the sense that we reduce the angle between these two vectors). To understand this statement better, note that when $\|\mathbf{w}^t\|_2 \approx \|\mathbf{w}^*\|_2$, we also have $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2 \approx \|\mathbf{w}^t - \mathbf{w}^*\|_2$. Additionally, $\|\mathbf{w}^t - \mathbf{w}^*\|_2 = O(\mathrm{OPT} + \epsilon)$ implies that the $L_2^2$ error of the hypothesis defined by $\hat{u}^t, \mathbf{w}^t$ is $O(\mathrm{OPT} + \epsilon)$ (see Claim 6.3.6). Thus, for a sufficiently good guess of the value of $\|\mathbf{w}^*\|_2$, Proposition 6.2.1 provides a local error bound of the form $\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) \gtrsim \mu \|\mathbf{w}^t - \mathbf{w}^*\|_2^2$ that holds outside of the set of $O(\mathrm{OPT} + \epsilon)$ error solutions, allowing us to contract the distance to this set.

**Proposition 6.2.1** (Alignment Sharpness of the Convex Surrogate)**.** *Suppose that $\mathcal{D}_{\mathbf{x}}$ is $(L, R)$-well-behaved, $\mathcal{U}_{(a,b)}$ is as in Definition 5.1.5, and $\epsilon, \delta > 0$. Let $\mu \gtrsim a^2 L R^4 / b$. Given any $\mathbf{w}^t \in \mathbb{B}(W)$, denote by $\hat{u}^t$ the optimal solution to (P) with respect to $\mathbf{w}^t$ and*

*the sample set* $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{m}$ *drawn i.i.d. from* $\mathcal{D}$. *If* $m$ *satisfies*

$$m \gtrsim dW^{9/2}b^4L^{-4}\log^4(d/(\epsilon\delta))(1/\epsilon^{3/2} + 1/(\epsilon\delta)) \ ,$$

*then, with probability at least* $1-\delta$,

$$\nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t)\cdot(\mathbf{w}^t - \mathbf{w}^*) \geq \mu\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2 - 2(\text{OPT}+\epsilon)/b - 2(\sqrt{\text{OPT}}+\sqrt{\epsilon})\|\mathbf{w}^t - \mathbf{w}^*\|_2 \ .$$

To prove Proposition 6.2.1, we rely on the following key ingredients. In Section (6.2.1), we prove our main technical lemma (Lemma 6.2.2), which states that the $L_2^2$ distance between the hypothesis $u(\mathbf{w}\cdot\mathbf{x})$ and the target $u^*(\mathbf{w}^*\cdot\mathbf{x})$ is bounded below by the misalignment of $\mathbf{w}^t$ and $\mathbf{w}^*$, i.e., the squared norm of the component of $\mathbf{w}^*$ that is orthogonal to $\mathbf{w}^t$, $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$. As will become apparent in the proof of Proposition 6.2.1, the inner product $\nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t)\cdot(\mathbf{w}^t - \mathbf{w}^*)$ can be bounded below as a function of the empirical $L_2^2$ error for $\mathbf{w}^t$ and a different (but related) activation $\hat{u}^{*t}$, which can in turn be argued to be close to the population $L_2^2$ error for a sufficiently large sample size, using concentration. Thus, Lemma 6.2.2 can be leveraged to obtain a term scaling with $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$ in the lower bound on $\nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t)\cdot(\mathbf{w}^t - \mathbf{w}^*)$.

In Section (6.2.2), we characterize structural properties of the population-optimal link functions $u^t$ and $u^{*t}$ (see (EP) and (EP*)), which play a crucial role in the proof of Proposition 6.2.1. Specifically, we show that the activation $u^t$ is close to the idealized activation $u^{*t}$ (the optimal activation without noise, given $\mathbf{w}^t$) in $L_2^2$ distance (Lemma 6.2.3). Since by standard uniform convergence results we have that $\hat{u}^t$ and $\hat{u}^{*t}$ are close to their population counterparts $u^t$ and $u^{*t}$, respectively, Lemma 6.2.3 certifies that $\hat{u}^t$ is not far from $\hat{u}^{*t}$. This property enables us to replace $\hat{u}^t$ by (the idealized) $\hat{u}^{*t}$

in the empirical surrogate gradient $\nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t;\hat{u}^t)$, which is easier to analyze, since $\hat{u}^{*t}$ is defined with respect to the "ideal" dataset (with uncorrupted labels).

Finally, as a simple corollary of Lemma 6.2.3, we obtain Corollary 6.2.5, which gives a clear explanation of why our algorithm, which alternates between updating $\mathbf{w}^t$ and $\hat{u}^t$, works: we show that the $L_2^2$ loss between the hypothesis generated by our algorithm $\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})$ and the underlying optimal hypothesis $u^*(\mathbf{w}^*\cdot\mathbf{x})$ is bounded above by the distance between $\mathbf{w}^t$ and $\mathbf{w}^*$. Since our structural sharpness result (Proposition 6.2.1) enables us to decrease $\|\mathbf{w}^t-\mathbf{w}^*\|_2$, Corollary 6.2.5 certifies that choosing the empirically-optimal activation leads to convergence of the hypothesis $\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})$.

Equipped with these technical lemmas, we prove our main structural result (Proposition 6.2.1) in Section (6.2.3).

### 6.2.1 $L_2^2$ Error and Misalignment

Our first key result is Lemma 6.2.2 below, which plays a critical role in the proof of Proposition 6.2.1. As discussed, for two different activations $u$ and $u^*$ and parameters $\mathbf{w}$ and $\mathbf{w}^*$ such that $\mathbf{w}$ and $\mathbf{w}^*$ are parallel, even when the $L_2^2$ error is $\Omega(1)$, the gradient $\nabla\mathcal{L}_{\text{sur}}(\mathbf{w};u)$ might not significantly align with the direction of $\mathbf{w}-\mathbf{w}^*$, and thus cannot provide sufficient information about the direction to decrease $\|\mathbf{w}-\mathbf{w}^*\|_2$. Intuitively, the following lemma shows that this is the only thing that can go wrong, and it happens when $\mathbf{w}$ and $\mathbf{w}^*$ are parallel. In particular, Lemma 6.2.2 shows that for any square integrable link function $f$, we can relate the $L_2^2$ distance $\mathbf{E}[(f(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]$ to the magnitude of the component of $\mathbf{w}^*$ that is *orthogonal* to $\mathbf{w}$. Although its proof is quite technical, this lemma is the main supporting result allowing us to prove Proposition 6.2.1, thus we provide its full proof below. It is however possible to follow the rest of this

section by only relying on its statement.

**Lemma 6.2.2** (Lower Bound on $L_2^2$ Error by Misalignment)**.** *Let $u^* \in \mathcal{U}_{(a,b)}$, $\mathcal{D}_{\mathbf{x}}$ be $(L,R)$-well-behaved, and $f : \mathbb{R} \mapsto \mathbb{R}$ be square-integrable with respect to the measure of the distribution $\mathcal{D}_{\mathbf{x}}$. Then, for any $\mathbf{w}, \mathbf{w}^* \in \mathbb{R}^d$,*

$$\mathbf{E}[(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \gtrsim a^2 L R^4 \|(\mathbf{w}^*)^{\perp\mathbf{w}}\|_2^2 \,.$$

*Proof.* The statement holds trivially if $\mathbf{w}$ is parallel to $\mathbf{w}^*$, so assume this is not the case. Let $\mathbf{v} = (\mathbf{w}^*)^{\perp\mathbf{w}} = \mathbf{w}^* - (\mathbf{w}^*\cdot\mathbf{w})\mathbf{w}/\|\mathbf{w}\|_2^2$. Suppose first that $\mathbf{w}\cdot\mathbf{w}^* \geq 0$. Then $\mathbf{w}^* = \alpha\mathbf{w} + \mathbf{v}$, for some $\alpha > 0$. Let $V$ be the subspace spanned by $\mathbf{w}, \mathbf{v}$. Then,

$$\mathbf{E}[(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \geq \mathbf{E}[(f(\mathbf{w}\cdot\mathbf{x}_V) - u^*(\mathbf{w}^*\cdot\mathbf{x}_V))^2 \mathbb{1}\{\mathbf{x}_V \in A\}] \,,$$

for any $A \subseteq \mathbb{R}^d$. For ease of notation, we drop the subscript $V$, and we assume that all $\mathbf{x}$ are projected to the subspace $V$. We denote by $\tilde{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|_2$ (resp. $\tilde{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|_2$) the unit vector in the direction of $\mathbf{w}$ (resp. $\mathbf{v}$). We choose $A = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}\cdot\mathbf{x} \geq 0, \tilde{\mathbf{v}}\cdot\mathbf{x} \in (R/16, R/8) \cup (3R/8, R/2)\}$.

The idea of the proof is to utilize the non-decreasing property of $u^*$ and the fact that the marginal distribution $\mathcal{D}_{\mathbf{x}}$ is anti-concentrated on the subspace $V$. In short, for any $\mathbf{x}$ such that $|\tilde{\mathbf{v}}\cdot\mathbf{x}| \leq R$, by the non-decreasing property of $u^*$ we know that $f(\mathbf{w}\cdot\mathbf{x})$ falls into one of the following four intervals:

$$(-\infty,\ u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R/32)] \,, \quad (u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R/32),\ u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R/4)] \,,$$
$$(u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R/4),\ u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R)] \,, \quad (u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R),\ +\infty) \ .$$

When $f(\mathbf{w} \cdot \mathbf{x})$ belongs to any of the intervals above, we can show that with some constant probability, the difference between $\mathbf{w}^* \cdot \mathbf{x}$ and $\mathbf{w} \cdot \mathbf{x}$ is proportional to $\|\mathbf{v}\|_2$, and hence $u^*(\mathbf{w}^* \cdot \mathbf{x})$ is far from $f(\mathbf{w} \cdot \mathbf{x})$ (due to the well-behaved property of the marginal $\mathcal{D}_{\mathbf{x}}$).

To indicate that $f(\mathbf{w} \cdot \mathbf{x})$ belongs to one of the intervals above, denote

$$
\begin{aligned}
I_1(\mathbf{x}) &= f(\mathbf{w} \cdot \mathbf{x}) - u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/32) \, , \\
I_2(\mathbf{x}) &= f(\mathbf{w} \cdot \mathbf{x}) - u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4) \, , \\
I_3(\mathbf{x}) &= f(\mathbf{w} \cdot \mathbf{x}) - u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R) \, .
\end{aligned}
$$

For any $\mathbf{x} \in \mathbb{R}^d$, using the assumption that $u^*$ is non-decreasing, we have that $I_1(\mathbf{x}) \geq I_2(\mathbf{x}) \geq I_3(\mathbf{x})$; as a consequence, it must be that $I_1(\mathbf{x}) I_2(\mathbf{x}) \geq 0$ or $I_2(\mathbf{x}) I_3(\mathbf{x}) \geq 0$.

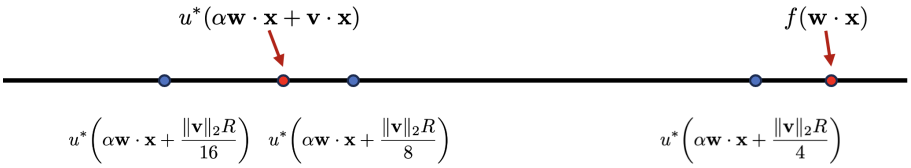



Figure 6.1: *Under the assumption that* $\tilde{\mathbf{v}} \cdot \mathbf{x} \in (R/16, R/8)$*, and* $I_1(\mathbf{x}) \geq 0, I_2(\mathbf{x}) \geq 0$*, the distance between* $f(\mathbf{w} \cdot \mathbf{x})$ *and* $u^*(\mathbf{w}^* \cdot \mathbf{x})$ *is at least* $|u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4) - u^*(\mathbf{w}^* \cdot \mathbf{x})| \geq a\|\mathbf{v}\|_2 R/8$.

**Case 1:** $f(\mathbf{w} \cdot \mathbf{x}) \in (u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4), \infty)$. Then $I_1(\mathbf{x}) \geq I_2(\mathbf{x}) \geq 0$. Let

$$
B := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{x} \geq 0, \tilde{\mathbf{v}} \cdot \mathbf{x} \in (R/16, R/8)\}
$$

and notice that $B \subseteq A$. We have that when $\mathbf{x} \in B$,

$$u^*(\mathbf{w}^* \cdot \mathbf{x}) = u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 \tilde{\mathbf{v}} \cdot \mathbf{x}) \in (u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/16),\ u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/8)),$$

thus we can conclude that

$$\begin{aligned}
&(f(\mathbf{w} \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{\mathbf{x} \in B\} \\
&= \Big(\{f(\mathbf{w} \cdot \mathbf{x}) - u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4)\} + \{u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4) - u^*(\mathbf{w}^* \cdot \mathbf{x})\}\Big)^2 \mathbb{1}\{\mathbf{x} \in B\} \\
&\geq (u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{\mathbf{x} \in B\}\ ,
\end{aligned}$$

where in the last inequality we used that $I_2(\mathbf{x}) = f(\mathbf{w} \cdot \mathbf{x}) - u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4) \geq 0$ and $u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4) - u^*(\mathbf{w}^* \cdot \mathbf{x}) \geq 0$ by the non-decreasing property of $u^*$, and the elementary inequality $(a+b)^2 \geq \max(a,b)^2$ for $a, b \geq 0$. Further, using $u^*(t) - u^*(t') \geq a(t - t')$ for $t \geq t' \geq 0$ (which holds by assumption) and $\mathbf{w}^* = \alpha \mathbf{w} + \mathbf{v}$, we have

$$\begin{aligned}
(u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{\mathbf{x} \in B\} &\geq a^2 (\|\mathbf{v}\|_2 R/4 - \mathbf{v} \cdot \mathbf{x})^2 \mathbb{1}\{\mathbf{x} \in B\} \\
&\geq a^2 \|\mathbf{v}\|_2^2 (R/8)^2 \mathbb{1}\{\mathbf{x} \in B\}\ ,
\end{aligned}$$

where in the last inequality we used that $0 \leq \tilde{\mathbf{v}} \cdot \mathbf{x} \leq R/8$ (by the definition of the event $B$). A visual illustration of the argument above is given in Figure 6.1.

**Case 2:** $f(\mathbf{w} \cdot \mathbf{x}) \in (-\infty,\ u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/32))$. Then $0 \geq I_1(\mathbf{x}) \geq I_2(\mathbf{x})$. We follow

a similar argument as in the previous case. In particular, we begin with

$$\begin{aligned}&(f(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\}=\\&(\{f(\mathbf{w}\cdot\mathbf{x})-u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2R/32)\}+\\&\{u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2R/32)-u^*(\mathbf{w}^*\cdot\mathbf{x})\})^2\mathbb{1}\{\mathbf{x}\in B\}.\end{aligned}\tag{6.3}$$

Note that $I_1(\mathbf{x})\le 0$ and $u^*(\mathbf{w}^*\cdot\mathbf{x})=u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2\tilde{\mathbf{v}}\cdot\mathbf{x})\ge u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2R/32)$ since $\tilde{\mathbf{v}}\cdot\mathbf{x}\ge R/16\ge R/32$ for $\mathbf{x}\in B$; thus, the two terms in curly brackets in (6.3) have the same sign and we further have:

$$\begin{aligned}(f(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\}&\ge(u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2R/32)-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\}\\&\ge a^2\|\mathbf{v}\|_2^2(R/32)^2\mathbb{1}\{\mathbf{x}\in B\}\ ,\end{aligned}$$

where in the first inequality we used the fact that $(a+b)^2\ge\max\{a^2,b^2\}$ when both $a,b\le 0$.

By the analysis of Case 1 and Case 2, we can conclude that when $I_1(\mathbf{x})I_2(\mathbf{x})\ge 0$, it must be:

$$(f(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{x}\in B\}\ge a^2\|\mathbf{v}\|_2^2R^2/2^{10}\mathbb{1}\{\mathbf{x}\in B\}\ .\tag{6.4}$$

**Case 3:** $f(\mathbf{w}\cdot\mathbf{x})\in(u^*(\alpha\mathbf{w}\cdot\mathbf{x}+\|\mathbf{v}\|_2R),\,+\infty)$. Then $I_2(\mathbf{x})\ge I_3(\mathbf{x})\ge 0$ and we choose

$$B'=\{\mathbf{x}\in\mathbb{R}^d:\mathbf{w}\cdot\mathbf{x}\ge 0,\tilde{\mathbf{v}}\cdot\mathbf{x}\in(3R/8,R/2)\}\ .$$

Following the same reasoning as in the previous two cases, we have

$$
\begin{aligned}
&(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2 \mathbb{1}\{\mathbf{x}\in B'\} \\
&= \Big(\{f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R)\} + \{u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R) - u^*(\mathbf{w}^*\cdot\mathbf{x})\}\Big)^2 \mathbb{1}\{\mathbf{x}\in B'\} \\
&\geq (u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2 \mathbb{1}\{\mathbf{x}\in B'\} \\
&\geq a^2\|\mathbf{v}\|_2^2 (R/2)^2 \mathbb{1}\{\mathbf{x}\in B'\} \,.
\end{aligned}
$$

**Case 4:** $f(\mathbf{w}\cdot\mathbf{x}) \in (-\infty, u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R/4))$. Then $0 \geq I_2(\mathbf{x}) \geq I_3(\mathbf{x})$. It follows that

$$
\begin{aligned}
&(f(\mathbf{w}\cdot\mathbf{x}) - u(\mathbf{w}^*\cdot\mathbf{x}))^2 \mathbb{1}\{\mathbf{x}\in B'\} \\
&= \Big(\{f(\mathbf{w}\cdot\mathbf{x}) - u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R/4)\} + \{u^*(\alpha\mathbf{w}\cdot\mathbf{x} + \|\mathbf{v}\|_2 R/4) - u(\mathbf{w}^*\cdot\mathbf{x})\}\Big)^2 \mathbb{1}\{\mathbf{x}\in B'\} \\
&\geq a^2\|\mathbf{v}\|_2^2 (R/8)^2 \mathbb{1}\{\mathbf{x}\in B'\} \,.
\end{aligned}
$$

Thus, from the analysis of Case 3 and Case 4, we conclude that when $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$, we have

$$
(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2 \mathbb{1}\{\mathbf{x}\in B'\} \geq a^2\|\mathbf{v}\|_2^2 (R^2/64) \mathbb{1}\{\mathbf{x}\in B'\} \,. \tag{6.5}
$$

Recall that for any $\mathbf{x}$, at least one of the inequalities $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$ or $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$ happens, thus, $\mathbb{1}\{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\} \geq 1 - \mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}$. Therefore, the probability mass of the region

$$
(B \cap \{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\}) \cup (B' \cap \{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\})
$$

can be bounded below by:

$$\begin{aligned}
&\mathbf{Pr}\left[\mathbf{x} \in (B \cap \{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\}) \cup (B' \cap \{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\})\right] \\
&= \int_V \left(\mathbb{1}\{\mathbf{x} \in B\}\mathbb{1}\{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\} + \mathbb{1}\{\mathbf{x} \in B'\}\mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\right)\gamma(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&\geq \int_{V,\|\mathbf{x}\|_\infty \leq R} \left(\mathbb{1}\{\mathbf{x} \in B\}\mathbb{1}\{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\} + \mathbb{1}\{\mathbf{x} \in B'\}\mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\right)L\,\mathrm{d}\mathbf{x} \\
&\geq L\int_{V,\|\mathbf{x}\|_\infty \leq R} \left(\mathbb{1}\{\mathbf{x} \in B\} + (\mathbb{1}\{\mathbf{x} \in B'\} - \mathbb{1}\{\mathbf{x} \in B\})\mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\right)\mathrm{d}\mathbf{x}\,, \quad (6.6)
\end{aligned}$$

where in the first inequality we used the assumption that $\mathcal{D}_{\mathbf{x}}$ is $(L, R)$- well-behaved. As a visual illustration of the lower bound argument above, the reader is referred to Figure 6.2.

To finish bounding below the probability in (6.6), it remains to bound the integral from its final inequality, which now does not involve the probability density function anymore, as we used the anti- concentration property of $\mathcal{D}_{\mathbf{x}}$ to uniformly bound below $\gamma(\mathbf{x})$. Recall that by definition, $I_1(\mathbf{x}), I_2(\mathbf{x}), I_3(\mathbf{x})$ are functions of $\mathbf{w} \cdot \mathbf{x}$ that do not depend on $\tilde{\mathbf{v}} \cdot \mathbf{x}$. Denote the projection of $\mathbf{x}$ on the standard basis of space $V$ by $\mathbf{x}_{\tilde{\mathbf{w}}} = \tilde{\mathbf{w}} \cdot \mathbf{x}$ and $\mathbf{x}_{\tilde{\mathbf{v}}} = \tilde{\mathbf{v}} \cdot \mathbf{x}$. Then, we have:

$$\begin{aligned}
&\int_{V,\|\mathbf{x}\|_\infty \leq R} \left(\mathbb{1}\{\mathbf{x} \in B'\} - \mathbb{1}\{\mathbf{x} \in B\}\right)\mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\,\mathrm{d}\mathbf{x} = \\
&\int_{|\mathbf{x}_{\tilde{\mathbf{w}}}| \leq R}\int_{|\mathbf{x}_{\tilde{\mathbf{v}}}| \leq R} (\mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{v}}} \in (\frac{3R}{8}, \frac{R}{2})\} - \mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{v}}} \in (\frac{R}{16}, \frac{R}{8})\})\,\mathrm{d}\mathbf{x}_{\tilde{\mathbf{v}}}\mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{w}}} \geq 0, I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\,\mathrm{d}\mathbf{x}_{\tilde{\mathbf{w}}} \\
&= \int_{|\mathbf{x}_{\tilde{\mathbf{w}}}| \leq R} \mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{w}}} \geq 0, I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\,\mathrm{d}\mathbf{x}_{\tilde{\mathbf{w}}} \int_{|\mathbf{x}_{\tilde{\mathbf{v}}}| \leq R} (\mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{v}}} \in (\frac{3R}{8}, \frac{R}{2})\} - \mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{v}}} \in (\frac{R}{16}, \frac{R}{8})\})\,\mathrm{d}\mathbf{x}_{\tilde{\mathbf{v}}} \\
&\geq 0\,.
\end{aligned}$$

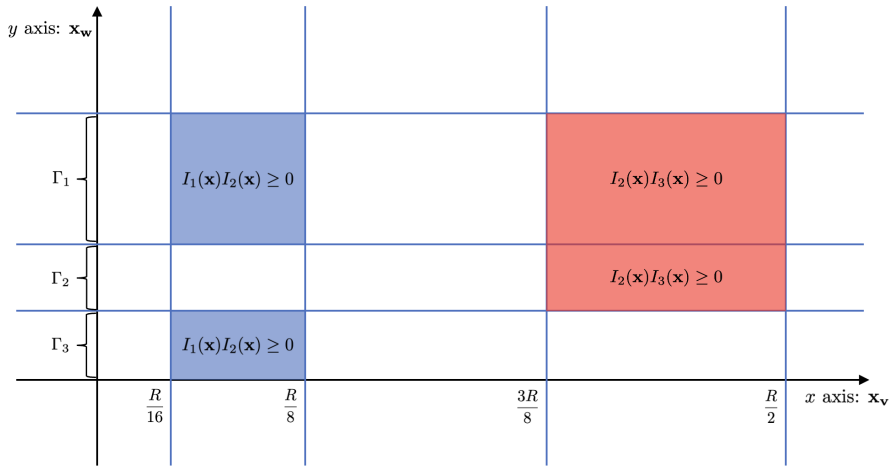



Figure 6.2: *On the 2-dimensional space $V$ spanned by $(\mathbf{x_v}, \mathbf{x_w})$, at each point $\mathbf{x} \in B \cup B'$, it must be that $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$ or $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$. $\Gamma_1$ denotes the interval of $\mathbf{x_w} = \mathbf{w} \cdot \mathbf{x}$ such that $f(\mathbf{w} \cdot \mathbf{x}) \geq u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R)$, hence both $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$, $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$; $\Gamma_2$ denotes the interval of $\mathbf{x_w}$ such that $f(\mathbf{w} \cdot \mathbf{x}) \in (u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/32), u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4))$, hence $I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0$; finally, $\Gamma_3$ denotes the interval of $\mathbf{x_w}$ such that $f(\mathbf{w} \cdot \mathbf{x}) \in (u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/4), u^*(\alpha \mathbf{w} \cdot \mathbf{x} + \|\mathbf{v}\|_2 R/))$, hence $I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0$. The area of the union of the red and blue regions is the lower bound on the probability in (6.6). As displayed in the figure, the sum of the blue and red region is lower bounded by $\mathbb{1}\{\mathbf{x} \in B\} + (\mathbb{1}\{\mathbf{x} \in B'\} - \mathbb{1}\{\mathbf{x} \in B\})\mathbb{1}\{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}$.*

Plugging the inequality above back into (6.6), we get:

$$\begin{aligned}
&\mathbf{Pr}\left[\mathbf{x} \in \Big(B \cap \{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\}\Big) \cup \Big(B' \cap \{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\Big)\right] \\
&\geq L \int_{V, \|\mathbf{x}\|_\infty \leq R} \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \geq 0, \tilde{\mathbf{v}} \cdot \mathbf{x} \in (R/16, R/8)\}\, \mathrm{d}\mathbf{x} \\
&= L \iint (\mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{w}}} \in (0, R)\}\, \mathrm{d}\mathbf{x}_{\tilde{\mathbf{w}}}) \mathbb{1}\{\mathbf{x}_{\tilde{\mathbf{v}}} \in (R/16, R/8)\}\, \mathrm{d}\mathbf{x}_{\tilde{\mathbf{v}}} = LR^2/16\,. \qquad (6.7)
\end{aligned}$$

We are now ready to provide a lower bound on the $L_2^2$ distance between $f(\mathbf{w} \cdot \mathbf{x})$ and

$u^*(\mathbf{w}^* \cdot \mathbf{x})$. Combining the inequalities from (6.4) and (6.5), we get

$$\begin{aligned}
&\mathbf{E}[(f(\mathbf{w}\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \\
&\geq \mathbf{E}[(f(\mathbf{w}\cdot\mathbf{x}_V) - u^*(\mathbf{w}^*\cdot\mathbf{x}_V))^2 \mathbb{1}\{\mathbf{x}_V \in A\}] \\
&\geq a^2(R^2/1024)\|\mathbf{v}\|_2^2\, \mathbf{E}[\mathbb{1}\Big\{\{\mathbf{x}_V \in B \cap \{I_1(\mathbf{x})I_2(\mathbf{x}) \geq 0\}\} \cup \{B' \cap \{I_2(\mathbf{x})I_3(\mathbf{x}) \geq 0\}\}\Big\}] \\
&\geq a^2(R^4/2^{13})L\|\mathbf{v}\|_2^2\,,
\end{aligned}$$

where we used (6.7) in the last inequality.

Now for the case where $\mathbf{w}\cdot\mathbf{w}^* \leq 0$, it holds $\mathbf{w}^* = \alpha\mathbf{w} + \mathbf{v}$ with $\alpha \leq 0$. Considering instead $A = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}\cdot\mathbf{x} \leq 0, \tilde{\mathbf{v}}\cdot\mathbf{x} \in (R/16, R/8) \cup (3R/8, R/2)\}$ and similarly $B = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}\cdot\mathbf{x} \leq 0, \tilde{\mathbf{v}}\cdot\mathbf{x} \in (R/16, R/8)\}$, $B' = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}\cdot\mathbf{x} \leq 0, \tilde{\mathbf{v}}\cdot\mathbf{x} \in (R/3, R/2)\}$, then all the steps above remains valid without modification. This completes the proof of Lemma 6.2.2. □

### 6.2.2 Closeness of Idealized and Attainable Activations

In this section, we bound the contribution of the error incurred from working with attainable link functions $\hat{u}^t$ in the iterations of the algorithm. The error incurred is due to both the arbitrary noise in the labels and due to using a finite sample set. In bounding the error, for analysis purposes, we introduce auxiliary population-level link functions.

Concretely, given $\mathbf{w} \in \mathbb{B}(W)$, a population-optimal activation is a solution to the following stochastic convex program:

$$u_{\mathbf{w}} \in \operatorname*{argmin}_{u \in \mathcal{U}_{(a,b)}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}\cdot\mathbf{x}) - y)^2]. \tag{EP}$$

We further introduce auxiliary "idealized, noiseless" activations, which, given noiseless labels $y^* = u^*(\mathbf{w}^* \cdot \mathbf{x})$ and a parameter weight vector $\mathbf{w}$, are defined via

$$u^*_{\mathbf{w}} \in \operatorname*{argmin}_{u \in \mathcal{U}_{(a,b)}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}\cdot\mathbf{x}) - y^*)^2]. \tag{EP*}$$

Below we relate $u^t := u_{\mathbf{w}^t}$ and $u^{*t} := u^*_{\mathbf{w}_t}$ and show that their $L_2^2$ error for the parameter vector $\mathbf{w}^t$ is bounded by OPT.

**Lemma 6.2.3** (Closeness of Population-Optimal Activations)**.** *Let $\mathbf{w}^t \in \mathbb{B}(W)$ and let $u^{*t}$, $u^t$ be defined as solutions to (EP*), (EP), respectively. Then,*

$$\mathbf{E}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] \le \mathrm{OPT}.$$

*Proof.* To prove Lemma 6.2.3, we first prove the following auxiliary claim, which is inspired by [KKSK11, Lemma 9].

**Claim 6.2.4.** *Let $\mathbf{w}^t \in \mathbb{B}(W)$ and let $u^{*t}, u^t$ be defined as solutions to (EP*), (EP), respectively. Then,*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - v(\mathbf{w}^t\cdot\mathbf{x}))(y - u^t(\mathbf{w}^t\cdot\mathbf{x}))] \ge 0, \quad \forall \in \mathcal{U}_{(a,b)}.$$

*Similarly,*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - v'(\mathbf{w}^t\cdot\mathbf{x}))(y^* - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))] \ge 0, \quad \forall v' \in \mathcal{U}_{(a,b)}.$$

*Proof of Claim 6.2.4.* Denote by $\mathcal{F}_t$ the set of functions of the form $f(\mathbf{x}) = u(\mathbf{w}^t\cdot\mathbf{x})$, where $u \in \mathcal{U}_{(a,b)}$ and $\mathbf{w}^t$ is a fixed vector in $\mathbb{B}(W)$. We first argue that $\mathcal{F}_t$ is a convex set,

using the definition of convexity. In particular, for any $\alpha \in (0,1)$ and any $f_1, f_2 \in \mathcal{F}_t$ such that $f_1(\mathbf{x}) = u_1(\mathbf{w}^t \cdot \mathbf{x})$, $f_2 = u_2(\mathbf{w}^t \cdot \mathbf{x})$, let $u_3(\cdot) = \alpha u_1(\cdot) + (1-\alpha)u_2(\cdot)$. Then:

$$\alpha f_1(\mathbf{x}) + (1-\alpha) f_2(\mathbf{x}) = \alpha u_1(\mathbf{w}^t \cdot \mathbf{x}) + (1-\alpha) u_2(\mathbf{w}^t \cdot \mathbf{x}) = u_3(\mathbf{w}^t \cdot \mathbf{x}).$$

It is immediate that $u_3$ is also $(a,b)$-bounded, non-decreasing, and $u_3(0) = 0$, hence $u_3 \in \mathcal{U}_{(a,b)}$ and $f_3(\mathbf{x}) = u_3(\mathbf{w}^t \cdot \mathbf{x}) \in \mathcal{F}_t$. Thus, $\mathcal{F}_t$ is convex.

Since $\mathcal{F}_t$ is a convex set of functions, we can regard $u^t(\mathbf{w}^t \cdot \mathbf{x})$ as the orthogonal projection of $y$ (which is a function of $\mathbf{x}$) onto the convex set $\mathcal{F}_t$. Classic inequalities for orthogonal projections can then be applied to our case. In particular, below we prove that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - v(\mathbf{w}^t \cdot \mathbf{x}))(y - u^t(\mathbf{w}^t \cdot \mathbf{x}))] \geq 0, \quad \forall v \in \mathcal{U}_{(a,b)}. \tag{6.8}$$

To prove (6.8), note first that $f_u(\mathbf{x}) = u^t(\mathbf{w}^t \cdot \mathbf{x}) \in \mathcal{F}_t$ and $f_v(\mathbf{x}) = v(\mathbf{w}^t \cdot \mathbf{x}) \in \mathcal{F}_t$ since $u^t, v \in \mathcal{U}_{(a,b)}$. Thus, for any $\alpha \in (0,1)$, we have $\alpha f_v(\mathbf{x}) + (1-\alpha) f_u(\mathbf{x}) \in \mathcal{F}_t$. Furthermore, by definition of $u^t$, $\forall f \in \mathcal{F}_t$ we have $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2] \leq \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - y)^2]$, therefore, it holds:

$$\begin{aligned}
0 &\leq \frac{1}{\alpha} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\alpha f_v(\mathbf{x}) + (1-\alpha) f_u(\mathbf{x}) - y)^2] - \frac{1}{\alpha} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2] \\
&= \frac{1}{\alpha} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - y + \alpha(v(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x})))^2 - (u^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[2(u^t(\mathbf{w}^t \cdot \mathbf{x}) - y)(v(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x})) + \alpha(v(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2].
\end{aligned}$$

Let $\alpha \downarrow 0$, and note that $\mathbf{E}[(v(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2] < +\infty$, we thus have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - v(\mathbf{w}^t \cdot \mathbf{x}))(y - u^t(\mathbf{w}^t \cdot \mathbf{x}))] \geq 0,$$

proving the claim.

The second claim can be proved following the same argument and is omitted for brevity. □

Summing up the first and second statement of Claim 6.2.4 with $v = u^{*t} \in \mathcal{U}_{(a,b)}$ in (6.8) and $v' = u^t \in \mathcal{U}_{(a,b)}$, we get:

$$
\begin{aligned}
0 \leq & \\
& \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(y - u^t(\mathbf{w}^t\cdot\mathbf{x})) + (u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))(y^* - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))] \\
& = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(y - y^* + u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))] \\
& = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(y - y^*)] - \mathbf{E}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]
\end{aligned}
$$

Rearranging and applying the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\mathbf{E}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] &\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(y - y^*)] \\
&\leq \sqrt{\mathbf{E}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]\,\mathbf{E}[(y - y^*)^2]}.
\end{aligned}
$$

To complete the proof, it remains to recall that $\mathbf{E}[(y - y^*)^2] = \mathrm{OPT}$ and rearrange the last inequality. □

As a consequence of the lemma above, we are able to relate $\hat{u}^t$ to the "noiseless" labels $y^* = u^*(\mathbf{w}^*\cdot\mathbf{x})$ by showing that the $L_2^2$ distance between $u^*(\mathbf{w}^*\cdot\mathbf{x})$ and the sample-optimal activation $\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})$ is bounded by $\|\mathbf{w}^t - \mathbf{w}^*\|_2^2$. Although Corollary 6.2.5 is not used in the proof of Proposition 6.2.1, we still present it here as it justifies the mechanism of our approach alternating between updates for $\mathbf{w}^t$ and $\hat{u}^t$.

**Corollary 6.2.5** (Closeness of Idealized and Attainable Activations)**.** *Let $\epsilon, \delta > 0$. Given a parameter $\mathbf{w}^t \in \mathbb{B}(W)$ and $m \gtrsim d \log^4(d/(\epsilon\delta))(b^2W^3/(L^2\epsilon))^{3/2}$ samples from $\mathcal{D}$, let $\hat{u}^t$ be the sample-optimal activation on these samples given $\mathbf{w}^t$, as defined in (P). Then, with probability at least $1-\delta$,*

$$\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq 3(\epsilon + \mathrm{OPT} + b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2) \,.$$

*Proof.* The corollary follows directly from the combination of Lemma D.3.4 and Lemma 6.2.3, as we have:

$$\begin{aligned}
&\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \\
&= \mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}) + u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) + u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \\
&\leq 3(\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2] + \mathbf{E}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2]) \\
&\quad + 3\,\mathbf{E}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \\
&\leq 3(\epsilon + \mathrm{OPT} + b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2),
\end{aligned}$$

where we used that because $u^{*t} \in \mathrm{argmin}_{u \in \mathcal{U}_{(a,b)}} \mathbf{E}[(u(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]$, we have $\mathbf{E}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq \mathbf{E}[(u^*(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2$, with the last inequality following from the fact that $u^* \in \mathcal{U}_{(a,b)}$. □

### 6.2.3 Proof of Proposition 6.2.1

*Proof of Proposition 6.2.1.* Given any weight parameter $\mathbf{w}^t \in \mathbb{B}(W)$ and $\hat{u}^t$ chosen as its corresponding sample-optimal solution to problem (P), let $u^t$ be the population-optimal activation, as defined by Problem (EP). Given a sample set $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$, consider

an idealized, "noise-free" set $S^*$ that assigns realizable labels to data vectors from $S$, i.e., $S^* = \{(\mathbf{x}^{(i)}, y^{*(i)})\}_{i=1}^m$, $y^{*(i)} = u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$. Further define idealized sample-optimal activations by

$$\hat{u}^*_{\mathbf{w}} \in \operatorname*{argmin}_{u \in \mathcal{U}_{(a,b)}} \frac{1}{m} \sum_{i=1}^{m} (u(\mathbf{w} \cdot \mathbf{x}^{(i)}) - y^{*(i)})^2. \tag{P*}$$

For a parameter $\mathbf{w}^t$, denote $\hat{u}^{*t} := \hat{u}^*_{\mathbf{w}^t}$, for simplicity, and recall that the population version of $\hat{u}^{*t}$ was defined by (EP*). To prove Proposition 6.2.1, we decompose $\nabla \widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*)$ into three summation terms:

$$\begin{aligned}
&\nabla \widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot (\mathbf{w}^t - \mathbf{w}^*) \\
&= \frac{1}{m} \sum_{i=1}^{m} (\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - y^{(i)})(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)} \\
&= \underbrace{\frac{1}{m} \sum_{i=1}^{m} (\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)} - \hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}))(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)}}_{Q_1} \\
&+ \underbrace{\frac{1}{m} \sum_{i=1}^{m} (\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - y^{*(i)})(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)}}_{Q_2} \\
&\quad + \underbrace{\frac{1}{m} \sum_{i=1}^{m} (y^{*(i)} - y^{(i)})(\mathbf{w}^t \cdot \mathbf{x}^{(i)} - \mathbf{w}^* \cdot \mathbf{x}^{(i)})}_{Q_3} .
\end{aligned} \tag{6.9}$$

We tackle each term $Q_1$ to $Q_3$ in (6.9) separately, using the following arguments relying on three auxiliary claims. The first claim states that $Q_1$ is of the order $(\sqrt{\epsilon} + \sqrt{\text{OPT}})\|\mathbf{w}^t - \mathbf{w}^*\|_2 + (\text{OPT} + \epsilon)/b$ with high probability.

**Claim 6.2.6.** *Let $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ be i.i.d. samples from $\mathcal{D}$ where $m$ is as specified in the statement of Proposition 6.2.1. Let $\hat{u}^t$ be the solution of optimization problem (P) given $\mathbf{w}^t \in \mathbb{B}(W)$ and $S$. Furthermore, denote the idealized version of $S$ by*

$S^* = \{(\mathbf{x}^{(i)}, y^{*(i)})\}_{i=1}^m$, *where* $y^{*(i)} = u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$. *Let* $\hat{u}^{*t}$ *be the solution of problem* (P*). *Then, with probability at least* $1-\delta$,

$$Q_1 = \frac{1}{m}\sum_{i=1}^{m}((\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)} \geq -(\sqrt{\epsilon}+\sqrt{\mathrm{OPT}})\|\mathbf{w}^t-\mathbf{w}^*\|_2-\frac{(\epsilon+\mathrm{OPT})}{b}$$

*Proof.* Adding and subtracting $u^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)})$ and $u^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)})$, we have

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}((\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - \hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}))(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m}(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - u^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}))(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)} \\
&+ \frac{1}{m}\sum_{i=1}^{m}(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - \hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}))(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)} \\
&+ \frac{1}{m}\sum_{i=1}^{m}(u^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}))(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)}. \qquad (6.10)
\end{aligned}$$

To proceed, we use that both $\hat{u}^{*t}(z)$ and $\hat{u}^t(z)$ are close to their population counterparts $u^t(z)$ and $u^{*t}(z)$, respectively. In particular, in Lemma D.3.4 and Lemma D.3.2, we show that using a dataset $S$ of $m$ samples such that

$$m \gtrsim d\log^4(d/(\epsilon\delta))\left(\frac{b^2W^3}{L^2\epsilon}\right)^{3/2},$$

we have that with probability at least $1-\delta$, for all $\mathbf{w}^t, \mathbf{w}^* \in \mathbb{B}(W)$ it holds

$$\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2] \leq \epsilon,\ \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \leq \epsilon. \qquad (6.11)$$

Now suppose that the inequalities in (6.11) hold for the given $\mathbf{w}^t \in \mathbb{B}(W)$ (which happens with probability at least $1-\delta$). Applying Chebyshev's inequality to the first

summation term in (6.10), we get:

$$\mathbf{Pr}\Bigg[|\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}- \mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}]|\geq s\Bigg] \leq \frac{1}{ms^2}\,\mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2], \quad (6.12)$$

since $\mathbf{x}^{(i)}$ are i.i.d. random variables. The next step is to bound the variance. Note that $\mathcal{D}_{\mathbf{x}}$ possesses a $1/L$-sub-exponential tail, thus we have $\mathbf{Pr}[|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\geq\|\mathbf{w}^t-\mathbf{w}^*\|_2 r]\leq(2/L^2)\exp(-Lr)$. Choose $r=\frac{2W}{L}\log(2/(L^2\epsilon'))$; then, we have $\mathbf{Pr}[|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\geq r]\leq\epsilon'$. Now we separate the variance under the event $A=\{\mathbf{x}:|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\leq r\}$ and its complement.

$$\begin{aligned}&\mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2]\\ =\;&\mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{A\}]\\ &+\mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2(1-\mathbb{1}\{A\})].\end{aligned} \quad (6.13)$$

Using that $\mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2]\leq\epsilon$, the first term in (6.13) can be bounded as follows:

$$\begin{aligned}\mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{A\}] &\leq r^2\,\mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2]\\ &\leq r^2\epsilon=\frac{4W^2\epsilon}{L^2}\log^2(2/(L^2\epsilon')).\end{aligned} \quad (6.14)$$

The second term in (6.13) can be bounded using that both $\hat{u}^t$ and $u^t$ are non-decreasing $b$-Lipschitz and vanish at zero (thus $|\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})|\leq b|\mathbf{w}^t\cdot\mathbf{x}|$ and $|u^t(\mathbf{w}^t\cdot\mathbf{x})|\leq b|\mathbf{w}^t\cdot\mathbf{x}|$,

with their signs determined by the sign of $\mathbf{w}^t \cdot \mathbf{x}$), and then applying Young's inequality:

$$\begin{aligned}
&\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2 (\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 (1 - \mathbb{1}\{A\})] \\
&\leq b^2 \, \mathbf{E}[(\mathbf{w}^t \cdot \mathbf{x})^2 (\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 (1 - \mathbb{1}\{A\})] \\
&\leq 2b^2 \, \mathbf{E}[((\mathbf{w}^t \cdot \mathbf{x})^4 + (\mathbf{w}^t \cdot \mathbf{x})^2 (\mathbf{w}^* \cdot \mathbf{x})^2)(1 - \mathbb{1}\{A\})] \, .
\end{aligned}$$

Since $\mathcal{D}_{\mathbf{x}}$ is sub-exponential, we have $\mathbf{E}[(\mathbf{v} \cdot \mathbf{x})^8] \leq c^2/L^8$ for some absolute constant $c$, hence

$$\mathbf{E}[(\mathbf{w}^t \cdot \mathbf{x})^4 (1 - \mathbb{1}\{A\})] \leq \sqrt{\mathbf{E}[W^8 ((\mathbf{w}^t / \|\mathbf{w}^t\|_2) \cdot \mathbf{x})^8] \, \mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| \geq r]} \leq cW^4 \sqrt{\epsilon'}/L^4.$$

Similarly, for $\mathbf{E}[(\mathbf{w}^t \cdot \mathbf{x})^2 (\mathbf{w}^* \cdot \mathbf{x})^2 (1 - \mathbb{1}\{A\})]$, we have:

$$\mathbf{E}[(\mathbf{w}^t \cdot \mathbf{x})^2 (\mathbf{w}^* \cdot \mathbf{x})^2 (1 - \mathbb{1}\{A\})] \leq 2 \, \mathbf{E}[((\mathbf{w}^t \cdot \mathbf{x})^4 + (\mathbf{w}^* \cdot \mathbf{x})^4)(1 - \mathbb{1}\{A\})] \leq 2c(W/L)^4 \sqrt{\epsilon'}.$$

Combining the inequalities above with (6.14), we get the final upper bound on the variance in (6.13):

$$\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2 (\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2] \leq \frac{4W^2 \epsilon}{L^2} \log^2(2/(L^2 \epsilon')) + 6cb^2 (W/L)^4 \sqrt{\epsilon'}.$$

Thus, choosing $s = \epsilon/b$ in (6.12), $\epsilon' = \epsilon^2$, and using $m \gtrsim W^4 b^4 \log^2(1/\epsilon)/(\epsilon \delta L^4)$ samples we get

$$\frac{1}{ms^2} \left( \frac{4W^2 \epsilon}{L^2} \log^2(2/(L\epsilon')) + \frac{12cb^2 W^4 \sqrt{\epsilon'}}{L^4} \right) \lesssim \frac{b^2 L^4 \epsilon \delta}{\epsilon^2 W^4 b^4 \log^2(1/\epsilon)} \left( \frac{W^2 \epsilon}{L^2} \log^2 \left( \frac{1}{L\epsilon} \right) + \frac{b^2 W^4 \epsilon}{L^4} \right) \leq \delta \, .$$

Plugging the inequality above back into (6.12) and recalling that $\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot$

$\mathbf{x}))^2] \leq \epsilon$ (from (6.11)), we finally have with probability at least $1-\delta$,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)} \\
&\geq \mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}] - \epsilon/b \\
&\geq -\sqrt{\mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}) - u^t(\mathbf{w}^t\cdot\mathbf{x}))^2]\,\mathbf{E}[(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2]} - \epsilon/b \\
&\geq -\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2 - \epsilon/b,
\end{aligned}$$

where in the second inequality we used the Cauchy-Schwarz inequality and in the last inequality we used the assumption that $\mathbf{E}[\mathbf{x}\mathbf{x}^\intercal] \preccurlyeq \mathbf{I}$. Finally, noting that (6.11) holds with probability at least $1-\delta$, applying a union bound we get that with probability at least $1-2\delta$, we have

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)} \geq -\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2 - \epsilon/b\,.$$

In summary, to guarantee that the inequality above remains valid, we need the batch size to be:

$$m \gtrsim \frac{dW^{9/2}b^4\log^4(d/(\epsilon\delta))}{L^4}\left(\frac{1}{\epsilon^{3/2}}+\frac{1}{\epsilon\delta}\right). \tag{6.15}$$

We finished bounding the first term in (6.10).

Since the same statements hold for the relationship between $\hat{u}^{*t}$ and $u^{*t}$ as they do for $\hat{u}^t$ and $u^t$, using the same argument we also get that with probability at least $1-2\delta$,

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)} \geq -\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2 - \epsilon/b,$$

which is the lower bound for the second term in (6.10).

Lastly, for the third term in (6.10), since in Lemma 6.2.3 we showed that for any $\mathbf{w}^t$ it always holds:

$$\mathbf{E}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \leq \mathrm{OPT},$$

the only change of the previous steps is at the right-hand side of (6.14), where instead of having the upper bound of $r^2\epsilon$, we have

$$\mathbf{E}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{A\}] \leq r^2\mathrm{OPT} = \frac{4W^2\mathrm{OPT}}{L^2}\log^2(2/(L^2\epsilon')).$$

By the same token, we have

$$\mathbf{E}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2(1 - \mathbb{1}\{A\})] \leq 6cb^2(W/L)^4\sqrt{\epsilon'}.$$

As a result, Chebyshev's inequality yields:

$$\begin{aligned}
\mathbf{Pr}\Bigg[\Bigg|\frac{1}{m}\sum_{i=1}^{m}&(u^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}))(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)} \\
&- \mathbf{E}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}]\Bigg| \geq s\Bigg] \\
&\leq \frac{1}{ms^2}\mathbf{E}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2] \\
&\leq \frac{1}{ms^2}\left(\frac{4W^2\mathrm{OPT}}{L^2}\log^2(2/(L^2\epsilon')) + \frac{6cb^2W^4\sqrt{\epsilon'}}{L^4}\right).
\end{aligned}$$

Now instead of choosing $s = \epsilon$, we let $s = (\mathrm{OPT} + \epsilon)/b$ and keep $\epsilon'$ as $\epsilon^2$ to get

$$\begin{aligned}
&\frac{1}{ms^2}\left(\frac{4W^2\mathrm{OPT}}{L^2}\log^2\left(\frac{2}{L^2\epsilon'}\right) + \frac{12cb^2W\sqrt{\epsilon'}}{L^4}\right) \\
\lesssim\ &\frac{b^2L^4\epsilon\delta}{dW^{9/2}b^4\log^4(d/(\epsilon\delta))(\mathrm{OPT}+\epsilon)^2}\left(\frac{W^2\mathrm{OPT}}{L^2}\log^2\left(\frac{1}{L\epsilon}\right) + \frac{b^2W^4\epsilon}{L^4}\right) \\
\leq\ &\delta,
\end{aligned}$$

under our choice of $m$ as specified in (6.15). Thus, we have that with probability at least $1-\delta$, it holds

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}(u^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)} \\
&\geq \mathbf{E}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}] - (\mathrm{OPT}+\epsilon)/b \\
&\geq -\sqrt{\mathrm{OPT}}\|\mathbf{w}^t-\mathbf{w}^*\|_2 - (\mathrm{OPT}+\epsilon)/b,
\end{aligned}$$

where in the last inequality we used the fact that

$$\begin{aligned}
|\,\mathbf{E}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}]| &\leq \sqrt{\mathbf{E}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2]\,\mathbf{E}[((\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x})^2]} \\
&\leq \sqrt{\mathrm{OPT}}\|\mathbf{w}^t-\mathbf{w}^*\|_2,
\end{aligned}$$

since $\mathbf{E}[(u^t(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] \leq \mathrm{OPT}$ by Lemma 6.2.3.

Therefore, combining the upper bounds on the three terms in (6.10), we get that with probability at least $1-5\delta$, it holds:

$$\frac{1}{m}\sum_{i=1}^{m}((\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - \hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)} \geq -(2\sqrt{\epsilon}+\sqrt{\mathrm{OPT}})\|\mathbf{w}^t-\mathbf{w}^*\|_2 - (3\epsilon+\mathrm{OPT})/b. \tag{6.16}$$

Since (6.16) was proved using arbitrary $\epsilon, \delta > 0$, it remains to replace $\delta \leftarrow \delta/5$ and $\epsilon \leftarrow \epsilon/4$ to complete the proof of Claim 6.2.6. □

The second claim leverages the misalignment lemma (Lemma 6.2.2) and shows that, up to small errors, $Q_2$ is a constant multiple of $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$.

**Claim 6.2.7.** *Let $S^* = \{(\mathbf{x}^{(i)}, y^{*(i)})\}_{i=1}^m$ be a sample set such that $\mathbf{x}^{(i)}$'s are i.i.d. samples from $\mathcal{D}_{\mathbf{x}}$ and $y^{*(i)} = u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$ for each $i$. Let $m$ be the value specified in the statement of Proposition 6.2.1. Then, given a parameter $\mathbf{w}^t \in \mathbb{B}(W)$, with probability at least $1 - \delta$,*

$$Q_2 = \frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)} \geq \frac{Ca^2LR^4}{b}\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2-\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2-\epsilon/b\,,$$

*where $C$ is an absolute constant.*

*Proof.* Before we proceed to the proof of the claim, let us consider first the inverse of $u^*$. Since $u^*(z) \in \mathcal{U}_{(a,b)}$ is strictly increasing when $z \geq 0$, $(u^*)^{-1}(\alpha)$ exists for $\alpha \geq 0$. However, when $z \leq 0$, $u^*(z)$ could be constant on some intervals, hence $(u^*)^{-1}(\alpha)$ might not exist for every $\alpha \leq 0$. We consider instead an 'empirical' version of $(u^*)^{-1}(\alpha)$ based on $S^*$, which is defined on every $\alpha \in \mathbb{R}$. Given a sample set $S^* = \{(\mathbf{x}^{(i)}, y^{*(i)})\}$ where $y^{*(i)} = u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$, let us sort the index $i$ in the increasing order of $\mathbf{w}^* \cdot \mathbf{x}^{(i)}$, i.e., $\mathbf{w}^* \cdot \mathbf{x}^{(1)} \leq \cdots \leq \mathbf{w}^* \cdot \mathbf{x}^{(m)}$. Since $u^*$ is a monotone function, this implies $y^{*(i)}$'s are also in increasing order, i.e., we have $y^{*(1)} \leq \cdots \leq y^{*(m)}$. We then partition the set $\{y^{*(i)}\}_{i=1}^m$ into blocks

$$\Delta_s = \{y^{*(k_{s-1}+1)}, \ldots, y^{*(k_s)}\}, \text{ s.t. } y^{*(k_{s-1}+1)} = \cdots = y^{*(k_s)} = \tau_s,$$

for $s = 1, \ldots, s'$. Since $\{y^{*(i)}\}$ is sorted in increasing order, we have $\tau_{s-1} < \tau_s$ for $s = 2, \ldots, s'$. Note that since $u^*(z)$ is strictly increasing when $z \geq 0$ and as $u^*(0) = 0$, $\Delta_s$ is a singleton set whenever $\tau_s > 0$. Furthermore, let us denote by $s^*$ the largest index among $1, \ldots, s'$ such that $\tau_{s^*} \leq 0$.

Suppose first that $\tau_{s^*} < 0$ and define a function $\hat{f} : \mathbb{R} \to \mathbb{R}$ in the following way:

$$\hat{f}(\alpha) = \begin{cases} (u^*)^{-1}(\alpha), & \alpha > 0 \\ \mathbf{w}^* \cdot \mathbf{x}^{(k_{s^*})} + \frac{\alpha - \tau_{s^*}}{\tau_{s^*}}(\mathbf{w}^* \cdot \mathbf{x}^{(k_{s^*})}), & \alpha \in [\tau_{s^*}, 0] \\ \mathbf{w}^* \cdot \mathbf{x}^{(k_s)}, & \alpha = \tau_s, s = 1, \ldots, s^* - 1 \\ \mathbf{w}^* \cdot \mathbf{x}^{(k_{s-1})} + \frac{\alpha - \tau_{s-1}}{\tau_s - \tau_{s-1}}(\mathbf{w}^* \cdot \mathbf{x}^{(k_{s-1}+1)} - \mathbf{w}^* \cdot \mathbf{x}^{(k_{s-1})}), & \alpha \in (\tau_{s-1}, \tau_s), s = 2, \ldots, s^* \\ \mathbf{w}^* \cdot \mathbf{x}^{(1)} + \frac{1}{b}(\alpha - \tau_1)\ . & \alpha \in (-\infty, \tau_1) \end{cases} \tag{6.17}$$

When $\tau_{s^*} = 0$, we define $(0 - \tau_{s^*})/\tau_{s^*} = -1$, and hence $\hat{f}(0) = 0$. The rest remains unchanged. A visualization of $\hat{f}$ with respect to the ReLU activation is presented in Figure 6.3.

The function $\hat{f}$ has the following properties. First, $\hat{f}(\alpha)$ satisfies $\hat{f}(0) = 0$, $(\alpha_1 - \alpha_2)/a \geq \hat{f}(\alpha_1) - \hat{f}(\alpha_2)$, for all $\alpha_1 \geq \alpha_2 \geq 0$, since $\hat{f}(\alpha) = (u^*)^{-1}(\alpha)$ when $\alpha > 0$ and $u^* \in \mathcal{U}_{(a,b)}$. Second, $\hat{f}(\alpha_1) - \hat{f}(\alpha_2) \geq (\alpha_1 - \alpha_2)/b$ for all $\alpha_1, \alpha_2 \in \mathbb{R}$, $\alpha_1 \geq \alpha_2$. This is because each segment of $\hat{f}$ has slope at least $1/b$. Third, for any $\alpha \geq u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$, it holds that $\hat{f}(\alpha) - \mathbf{w}^* \cdot \mathbf{x}^{(i)} \geq (\alpha - u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}))/b$. To see this, suppose $u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}) \in \Delta_s$. Then for any $\alpha \geq \tau_s$, we have

$$\hat{f}(\alpha) - \mathbf{w}^* \cdot \mathbf{x}^{(i)} \geq \hat{f}(\alpha) - \mathbf{w}^* \cdot \mathbf{x}^{(k_s)} = \hat{f}(\alpha) - \hat{f}(\tau_s) = \hat{f}(\alpha) - \hat{f}(u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})) \geq (\alpha - u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}))/b,$$

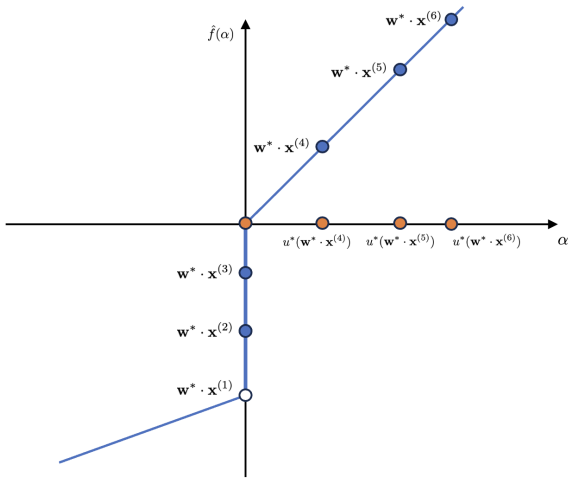



Figure 6.3: *An illustration of* $\hat{f}$ *for* $u^*(z) = \max\{0, z\}$ *and a dataset* $S^* = \{(\mathbf{x}^{(1)}, u^*(\mathbf{w}^* \cdot \mathbf{x}^{(1)})), \ldots, (\mathbf{x}^{(6)}, u^*(\mathbf{w}^* \cdot \mathbf{x}^{(6)}))\}$ *where* $\mathbf{w}^* \cdot \mathbf{x}^{(1)} < \mathbf{w}^* \cdot \mathbf{x}^{(2)} < \mathbf{w}^* \cdot \mathbf{x}^{(3)} < 0$.

using that the slope of $\hat{f}(\alpha)$ is at least $1/b$. On the other hand, when $\alpha < u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$, we have $\mathbf{w}^* \cdot \mathbf{x}^{(i)} - \hat{f}(\alpha) \geq (u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}) - \alpha)/b$. This can be seen similarly from the construction of $\hat{f}$.

Finally, suppose $u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}) \in \Delta_s$. Then for any $\alpha < \tau_s$, we have

$$\begin{aligned}\mathbf{w}^* \cdot \mathbf{x}^{(i)} - \hat{f}(\alpha) \geq \mathbf{w}^* \cdot \mathbf{x}^{(k_{s-1}+1)} - \hat{f}(\alpha) = \hat{f}(\tau_s) - \hat{f}(\alpha) = \hat{f}(u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})) - \hat{f}(\alpha) \\ \geq (u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}) - \alpha)/b.\end{aligned}$$

Again, we used the fact that $\hat{f}(\alpha_1) - \hat{f}(\alpha_2) \geq (\alpha_1 - \alpha_2)/b$ for all $\alpha_1, \alpha_2 \in \mathbb{R}$, $\alpha_1 \geq \alpha_2$ in the last inequality.

Now we turn to the summation displayed in the statement of the claim. To proceed, we add and subtract $\hat{f}(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}))$ in the second component in the inner product, which

yields:

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\\
=&\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})))\\
&+\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(\hat{f}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))-\mathbf{w}^*\cdot\mathbf{x}^{(i)}). \qquad (6.18)
\end{aligned}$$

The statement considers a smaller activation class and a function $f$ with different properties compared to [KKSK11], and the proof is based on a rigorous KKT argument.

**Fact 6.2.8.** *Let $\mathbf{w}^t\in\mathbb{B}(W)$. Given $m$ samples $S=\{(\mathbf{x}^{(1)},y^{*(1)}),\cdots,(\mathbf{x}^{(m)},y^{*(m)})\}$, let $\hat{u}^{*t}$ be one of the solutions to the optimization problem (P*) , i.e.,*

$$\hat{u}^{*t}\in\operatorname*{argmin}_{u\in\mathcal{U}_{(a,b)}}(1/m)\sum_{i=1}^{m}(u(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})^2\ .$$

*Then*

$$\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-f(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})))\geq 0,$$

*for any function $f:\mathbb{R}\to\mathbb{R}$ such that $f(0)=0$, $(\alpha_1-\alpha_2)/a\geq f(\alpha_1)-f(\alpha_2)$ for all $\alpha_1\geq\alpha_2\geq 0$, and $f(\alpha_1)-f(\alpha_2)\geq(\alpha_1-\alpha_2)/b$, $\forall\alpha_1,\alpha_2\in\mathbb{R}$, $\alpha_1\geq\alpha_2$.*

*Proof.* We transform the optimization problem (P*) to a quadratic optimization problem with linear constraints. To guarantee that the solution of this quadratic problem corresponds to a function that is $(a,b)$-unbounded, we add a sample $(\mathbf{x}^{(k)},y^{*(k)})=(\mathbf{0},0)$ to the sample set. Let $z_i=\mathbf{w}^t\cdot\mathbf{x}^{(i)}$ such that (after sorting the indices) $z_1\leq z_2\leq\cdots\leq$

$z_m$ and $z_k = 0$. We solve the following optimization problem:

$$\begin{aligned} \min_{\tilde{y}^{(i)}, i\in[m]} \ & \sum_{i=1}^{m} (\tilde{y}^{(i)} - y^{*(i)})^2 \\ \text{s.t.} \ & 0 \le \tilde{y}^{(i+1)} - \tilde{y}^{(i)}, && 1 \le i \le k-1 \ , \\ & a(z_{i+1} - z_i) \le \tilde{y}^{(i+1)} - \tilde{y}^{(i)}, && k \le i \le m-1 \ , \\ & \tilde{y}^{(i+1)} - \tilde{y}^{(i)} \le b(z_{i+1} - z_i), && 1 \le i \le m-1 \ , \\ & \tilde{y}^{(k)} = 0 \ . \end{aligned} \tag{6.19}$$

Denote the solution of (6.19) as $\hat{y}^{*(i)}$, $i = 1, \cdots, m$. Let $\hat{u}^{*t}(z)$ be the linear interpolation function of $(z_i, \hat{y}^{*(i)})$, then $\hat{u}^{*t} \in \mathcal{U}_{(a,b)}$ since $\hat{u}^{*t}(0) = \hat{u}^{*t}(z_k) = \hat{y}^{*(k)} = 0$, $\hat{u}^{*t}$ is $b$-Lipschitz and $\hat{u}^{*t}(z) - \hat{u}^{*t}(z') \ge a(z - z')$ for all $z \ge z' \ge 0$. In other words, finding a solution of (P*) is equivalent to solving (6.19).

Now observe that the summation $\sum_{i=1}^{m} (\hat{y}^{*(i)} - y^{*(i)})(z_i - f(\hat{y}^{*(i)}))$ can be transformed into the following:

$$\sum_{i=1}^{m} (\hat{y}^{*(i)} - y^{*(i)})(z_i - f(\hat{y}^{*(i)})) = \sum_{i=1}^{m} \left( \sum_{j=1}^{i} (\hat{y}^{*(j)} - y^{*(j)}) \right) (z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))), \tag{6.20}$$

where we let $z_{m+1} = 0$, $\hat{y}^{*}_{m+1} = 0$ (and hence $f(\hat{y}^{*}_{m+1}) = 0$ as $f(0) = 0$).

To utilize the information that $\hat{y}^{*(i)}$ is the minimizer of the optimization problem (6.19), we write down the KKT conditions for the optimization problem (6.19) described

above:

$$\hat{y}^{*(i)} = y^{*(i)} + (\lambda'_i - \lambda'_{i-1})/2 - (\lambda_i - \lambda_{i-1})/2 - (\nu_k/2)\mathbb{1}\{i = k\}, \quad i = 1, \cdots, m; \tag{6.21}$$

$$-\lambda_i(\hat{y}^{*(i+1)} - \hat{y}^{*(i)}) = 0, \qquad i = 1, \cdots, k-1; \tag{6.22}$$

$$\lambda_i(a(z_{i+1} - z_i) - (\hat{y}^{*(i+1)} - \hat{y}^{*(i)})) = 0, \qquad i = k, \cdots, m-1; \tag{6.23}$$

$$\lambda'_i((\hat{y}^{*(i+1)} - \hat{y}^{*(i)}) - b(z_{i+1} - z_i)) = 0, \qquad i = 1, \cdots, m-1; \tag{6.24}$$

$$\nu_k \hat{y}^{*(k)} = 0\ , \tag{6.25}$$

where $\lambda_i, \lambda'_i \geq 0$, for $i = 1, \ldots, m-1$, and $\nu_k \in \mathbb{R}$ are dual variables, and we let $\lambda_0 = \lambda'_0 = 0$ for the convenience of presenting (6.21).

Summing up (6.21) recursively, we immediately get that

$$\sum_{j=1}^{i}(\hat{y}^{*(i)} - y^{*(i)}) = \frac{1}{2}((\lambda'_i - \lambda_i) - \nu_k \mathbb{1}\{i \geq k\}).$$

Plugging the equality above back into (6.20), we have

$$\begin{aligned}
&\sum_{i=1}^{m}(\hat{y}^{*(i)} - y^{*(i)})(z_i - f(\hat{y}^{*(i)})) = \\
&\frac{1}{2}\sum_{i=1}^{m}(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \\
&+ \frac{1}{2}\sum_{i=k}^{m}\nu_k(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \\
&= \frac{1}{2}\sum_{i=1}^{m}(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \\
&+ \nu_k(z_k - f(\hat{y}^{*(k)}) - (z_{m+1} - f(\hat{y}^{*(m+1)}))) \qquad (6.26)
\end{aligned}$$

Since by definition, $z_{m+1} = f(\hat{y}^{*(m+1)}) = 0$, $z_k = 0$, and as $\hat{y}^{*(i)}$, $i \in [m]$, is a feasible solution of (6.19), it holds $\hat{y}^{*(k)} = 0$, we thus have

$$\nu_k(z_k - f(\hat{y}^{*(k)}) - (z_{m+1} - f(\hat{y}^{*(m+1)}))) = 0.$$

Plugging this back into (6.26), we get

$$\begin{aligned}
\sum_{i=1}^{m}(\hat{y}^{*(i)} - y^{*(i)})(z_i - f(\hat{y}^{*(i)})) &= \frac{1}{2}\sum_{i=1}^{m}(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \\
&= \underbrace{\frac{1}{2}\sum_{i=1}^{k-1}(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)})))}_{S_1} \\
&\quad + \underbrace{\frac{1}{2}\sum_{i=k}^{m}(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)})))}_{S_2}. \qquad (6.27)
\end{aligned}$$

Consider first $S_1$. Suppose that for some $i \in \{1, \ldots, k-1\}$ we have $\lambda_i', \lambda_i > 0$. Then, according to the complementary slackness condition (6.22) and (6.23), it holds that

$0 = \hat{y}^{*(i+1)} - \hat{y}^{*(i)} = b(z_{i+1} - z_i)$. Therefore,

$$(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \geq 0.$$

Suppose now that for some $i \in \{1, \ldots, k-1\}$, it holds $\lambda_i' > 0, \lambda = 0$. Then, it must be the case that

$$\hat{y}^{*(i+1)} - \hat{y}^{*(i)} = b(z_{i+1} - z_i) \geq 0,$$

according to the KKT condition (6.24). Since

$$f(\hat{y}^{*(i+1)}) - f(\hat{y}^{*(i)}) \geq (\hat{y}^{*(i+1)} - \hat{y}^{*(i)})/b$$

by assumption on $f$, we thus have

$$(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \geq 0.$$

Finally, if $\lambda_i > 0, \lambda' = 0$, then (6.22) indicates that $0 = \hat{y}^{*(i+1)} - \hat{y}^{*(i)}$. Therefore, as $z_{i+1} \geq z_i$, the $i^{\text{th}}$ summand is also positive. In summary, $S_1 \geq 0$.

Now consider $S_2$. Observe that if for some $i \in \{k, \ldots, m\}$ it holds $\lambda_i > 0$ and $\lambda_i' > 0$ at the same time, then KKT conditions (6.23) and (6.24) imply that

$$a(z_{i+1} - z_i) = \hat{y}^{*(i+1)} - \hat{y}^{*(i)} = b(z_{i+1} - z_i),$$

as $a < b$ and it has to be $z_{i+1} - z_i = \hat{y}^{*(i+1)} - \hat{y}^{*(i)} = 0$, which indicates that the $i^{\text{th}}$

summand in the second term must be 0, i.e.,

$$(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) = 0.$$

Now suppose for some $i \in \{1, \ldots, m\}$, $\lambda_i' > 0$ and $\lambda_i = 0$. Then by the complementary slackness conditions (6.23) and (6.24), it must be that

$$\hat{y}^{*(i+1)} - \hat{y}^{*(i)} = b(z_{i+1} - z_i) \geq 0.$$

Again, since $f$ satisfies

$$f(\hat{y}^{*(i+1)}) - f(\hat{y}^{*(i)}) \geq (\hat{y}^{*(i+1)} - \hat{y}^{*(i)})/b$$

for any $\hat{y}^{*(i+1)} \geq \hat{y}^{*(i)}$, we thus have

$$z_i - z_{i+1} + (f(\hat{y}^{*(i+1)}) - f(\hat{y}^{*(i)})) \geq 0.$$

Thus, it holds that

$$(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \geq 0.$$

On the other hand, if $\lambda_i' = 0$ and $\lambda_i > 0$, then complementary slackness implies that

$$\hat{y}^{*(i+1)} - \hat{y}^{*(i)} = a(z_{i+1} - z_i) \geq 0.$$

Furthermore, since $\hat{y}^{*(i)} \geq \hat{y}^{*(k)} \geq 0$ when $i \geq k$, using the assumption that

$$(\alpha_1 - \alpha_2)/a \geq f(\alpha_1) - f(\alpha_2)$$

when $\alpha_1 \geq \alpha_2 \geq 0$, we get

$$z_i - z_{i+1} + (f(\hat{y}^{*(i+1)}) - f(\hat{y}^{*(i)})) \leq 0,$$

and hence

$$(\lambda_i' - \lambda_i)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \geq 0$$

holds as well. Thus we conclude that $S_2 \geq 0$.

In summary, since each summand in (6.27) is non-negative, we finally get that

$$\sum_{i=1}^{m}(\hat{y}^{*(i)} - y^{*(i)})(z_i - f(\hat{y}^{*(i)})) = \sum_{i=1}^{m}\left(\sum_{j=1}^{i}(\hat{y}^{*(j)} - y^{*(j)})\right)(z_i - f(\hat{y}^{*(i)}) - (z_{i+1} - f(\hat{y}^{*(i+1)}))) \geq 0.$$

This completes the proof of Fact 6.2.8. □

As we have already argued, $\hat{f}$ satisfies the assumptions of Fact 6.2.8, hence

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - y^{*(i)})(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)} \geq \frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - y^{*(i)})(\hat{f}(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)})) - \mathbf{w}^* \cdot \mathbf{x}^{(i)}). \tag{6.28}$$

Recall that we have shown the function $\hat{f}$ satisfies $\hat{f}(\alpha) - \mathbf{w}^* \cdot \mathbf{x}^{(i)} \geq (\alpha - u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}))/b \geq 0$ whenever $\alpha \geq u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$, and moreover, $\mathbf{w}^* \cdot \mathbf{x}^{(i)} - \hat{f}(\alpha) \geq (u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)}) - \alpha)/b \geq 0$ when $\alpha < u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$. Therefore, letting $\alpha = \hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}^{(i)})$ and combining

these results we get

$$(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))(f(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))-\mathbf{w}^*\cdot\mathbf{x}^{(i)})\geq\frac{1}{b}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)}))^2.$$

Plugging the inequality above back into (6.28) we then get

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}^{(i)}\geq\frac{1}{mb}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})^2\,. \tag{6.29}$$

The goal now is to bound below the right-hand side of (6.29) by $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]$ and some small error terms using Chebyshev inequality as we did in Claim 6.2.6. Plugging in Lemma 6.2.2, we can further lower bound $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]$ by $\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$ and then we are done with the proof of this claim. Note that Chebyshev's inequality yields

$$\mathbf{Pr}[|\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-y^{*(i)})^2-\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]|\geq s]\leq\frac{1}{ms^2}\,\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^4]. \tag{6.30}$$

We now bound $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^4]$. Observe that

$$\begin{aligned}\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^4]&=\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x})+u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2]\\&\leq 4\,\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2((\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2+(y^*)^2)]\\&\quad+4\,\mathbf{E}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y^*)^2((\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2+(y^*)^2)].\end{aligned} \tag{6.31}$$

We focus on the two terms in (6.31) separately. Again, choosing $r=\frac{2W}{L}\log(2/(L^2\epsilon'))$, then by the $L$-sub-exponential tail bound of $\mathcal{D}_{\mathbf{x}}$, it holds $\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}|\geq r]\leq\epsilon',\mathbf{Pr}[|\mathbf{w}^*\cdot\mathbf{x}|\geq r]\leq\epsilon'$. Since $y^*=u^*(\mathbf{w}^*\cdot\mathbf{x})$ and both $u^*$ and $\hat{u}^{*t}$ are non-decreasing $b$-Lipschitz, it

holds:

$$
\begin{aligned}
&\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2((\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2 + (y^*)^2)] \\
&\leq b^2\, \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2((\mathbf{w}^t \cdot \mathbf{x})^2 + (\mathbf{w}^* \cdot \mathbf{x})^2)] \\
&= b^2\, \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2((\mathbf{w}^t \cdot \mathbf{x})^2 + (\mathbf{w}^* \cdot \mathbf{x})^2)\mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \leq r, |\mathbf{w}^* \cdot \mathbf{x}| \leq r\}] \\
&\quad + b^2\, \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2((\mathbf{w}^t \cdot \mathbf{x})^2 + (\mathbf{w}^* \cdot \mathbf{x})^2)\mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \geq r \text{ or } |\mathbf{w}^* \cdot \mathbf{x}| \geq r\}] \\
&\leq 2b^2r^2\, \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \\
&\quad + 2b^4\, \mathbf{E}[2(\mathbf{w}^t \cdot \mathbf{x})^2((\mathbf{w}^t \cdot \mathbf{x})^2 + (\mathbf{w}^* \cdot \mathbf{x})^2)\mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \geq r \text{ or } |\mathbf{w}^* \cdot \mathbf{x}| \geq r\}]. \qquad (6.32)
\end{aligned}
$$

The first term in (6.32) can be upper bounded using Lemma D.3.2, which states that when

$$m \gtrsim d \log(1/\delta)(b^2 W^3 \log^2(d/\epsilon)/(L^2\epsilon))^{3/2}),$$

with probability at least $1-\delta$ it holds $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \leq \epsilon$ for all $\mathbf{w}^t \in \mathbb{B}(W)$. Now suppose this inequality is valid given $\mathbf{w}^t \in \mathbb{B}(W)$ (which happens with probability at least $1-\delta$). For the second term in (6.32), note that for any unit vector $\mathbf{a}$ it holds $\mathbf{E}[(\mathbf{a} \cdot \mathbf{x})^8] \leq c^2/L^8$ for some absolute constant $c > 0$, and furthermore, the magnitude of $r$ ensures that $\mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| \geq r \text{ or } |\mathbf{w}^* \cdot \mathbf{x}| \geq r] \leq 2\epsilon'$; therefore, combining these bounds, we get:

$$
\begin{aligned}
&\mathbf{E}[2(\mathbf{w}^t \cdot \mathbf{x})^2((\mathbf{w}^t \cdot \mathbf{x})^2 + (\mathbf{w}^* \cdot \mathbf{x})^2)\mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \geq r \text{ or } |\mathbf{w}^* \cdot \mathbf{x}| \geq r\}] \\
&\leq 2\sqrt{\mathbf{E}[(\mathbf{w}^t \cdot \mathbf{x})^8]\, \mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| \geq r \text{ or } |\mathbf{w}^* \cdot \mathbf{x}| \geq r]} \\
&\quad + 2\sqrt{2(\mathbf{E}[(\mathbf{w}^t \cdot \mathbf{x})^8] + \mathbf{E}[(\mathbf{w}^* \cdot \mathbf{x})^8])\, \mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| \geq r \text{ or } |\mathbf{w}^* \cdot \mathbf{x}| \geq r]} \\
&\leq 24c(W/L)^4\sqrt{\epsilon'}.
\end{aligned}
$$

Plugging back into (6.32), we have

$$\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2((\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2 + (y^*)^2)] \leq 2b^2r^2\epsilon + 48c(bW/L)^4\sqrt{\epsilon'},$$

which is the upper bound on the first term of (6.31).

For the second term in (6.31), since by definition we have $u^{*t} \in \operatorname{argmin}_{u \in \mathcal{U}_{(a,b)}} \mathbf{E}[(u(\mathbf{w}^t \cdot \mathbf{x}) - y^*)^2]$, it holds that

$$\mathbf{E}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - y^*)^2] \leq \mathbf{E}[(u^*(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq b^2\,\mathbf{E}[((\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x})^2] \leq b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2,$$

noting in addition that $\mathbf{E}[\mathbf{x}\mathbf{x}^\intercal] \preccurlyeq \mathbf{I}$. Thus, using similar steps as in (6.32), we have

$$\begin{aligned}
&\mathbf{E}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - y^*)^2((\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2 + (y^*)^2)] \\
&\leq 2b^2r^2\,\mathbf{E}[(u(\mathbf{w}^t \cdot \mathbf{x}) - y^*)^2] \\
&\quad + 2b^4\,\mathbf{E}[2((\mathbf{w}^t \cdot \mathbf{x})^2 + (\mathbf{w}^* \cdot \mathbf{x})^2)^2 \mathbb{1}\{|\mathbf{w}^t \cdot \mathbf{x}| \geq r \text{ or } |\mathbf{w}^* \cdot \mathbf{x}| \geq r\}] \\
&\leq 2b^4r^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 48c(bW/L)^4\sqrt{\epsilon}.
\end{aligned}$$

In summary, combining all the results and plugging them back into (6.31), we finally get the upper bound for the variance:

$$\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - y^*)^4] \leq \frac{32b^2W^2}{L^2}\log^2(2/(L^2\epsilon'))(b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \epsilon) + 384c(bW/L)^4\sqrt{\epsilon'}.$$

Let $s = b\sqrt{\epsilon}\|\mathbf{w}^t - \mathbf{w}^*\|_2 + \epsilon/b$ and plug the last inequality back into (6.30) to get:

$$\mathbf{Pr}\left[\left|\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - y^{*(i)})^2 - \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - y^*)^2]\right| \geq b\sqrt{\epsilon}\|\mathbf{w}^t - \mathbf{w}^*\|_2 + \epsilon/b\right]$$
$$\leq \frac{1}{m(\epsilon b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \epsilon^2/b^2)}\left(\frac{32b^2W^2}{L^2}\log^2(\frac{2}{L^2\epsilon'})(b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \epsilon) + 384c(bW/L)^4\sqrt{\epsilon'}\right)$$

Choosing $\epsilon' = \epsilon^2/b^4$ and using similar arguments as in Claim 6.2.6, we get that the right-hand side of the inequality above is bounded by $\delta$, given our choice of $m \gtrsim db^4W^{9/2}\log^4(d/(\epsilon\delta))(1/\epsilon^{3/2} + 1/(\epsilon\delta))$ as specified in the statement of Proposition 6.2.1. In summary, after a union bound on the probability above and the event that $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] \leq \epsilon$, we have with probability at least $1 - 2\delta$,

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - y^{*(i)})^2 \geq \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - y^*)^2] - \sqrt{\epsilon}b\|\mathbf{w}^t - \mathbf{w}^*\|_2 - \epsilon/b.$$

Recall that in Lemma 6.2.2 we showed that $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \geq Ca^2LR^4\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2$ for an absolute constant $C$; thus, our final result is that with probability at least $1 - \delta$,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - y^{*(i)})(\mathbf{w}^t - \mathbf{w}^*)\cdot\mathbf{x}^{(i)} &\geq \frac{1}{mb}\sum_{i=1}^{m}(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}^{(i)}) - y^{*(i)})^2 \\ &\geq \frac{Ca^2LR^4}{b}\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2 - \sqrt{\epsilon}\|\mathbf{w}^t - \mathbf{w}^*\|_2 - \epsilon/b.\end{aligned}$$

This completes the proof of Claim 6.2.7. □

Similar to Claim 6.2.6, the last claim shows that $Q_3$ is of the order $\sqrt{\mathrm{OPT}}\|\mathbf{w}^* - \mathbf{w}^t\|_2$, which is small compared to the positive term in Claim 6.2.7 outside the set of $O(\mathrm{OPT}) + \epsilon$ error solutions.

**Claim 6.2.9.** *Let $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ be i.i.d. samples from $\mathcal{D}$, and denote by $S^* =$*

$\{(\mathbf{x}^{(i)}, y^{*(i)})\}_{i=1}^m$ *the idealized version of* $S$*, where* $y^{*(i)} = u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})$*. Under the condition of Proposition 6.2.1, given a parameter* $\mathbf{w}^t \in \mathbb{B}(W)$*, with probability at least* $1 - \delta$*,*

$$Q_3 = \frac{1}{m}\sum_{i=1}^m (y^{*(i)} - y^{(i)})(\mathbf{w}^t \cdot \mathbf{x}^{(i)} - \mathbf{w}^* \cdot \mathbf{x}^{(i)}) \geq -\sqrt{\mathrm{OPT}}\|\mathbf{w}^* - \mathbf{w}^t\|_2 - (\mathrm{OPT} + \epsilon)/b\ .$$

*Proof.* By Chebyshev's inequality, we can write

$$\mathbf{Pr}\left[\left|\frac{1}{m}\sum_{i=1}^m (y^{*(i)} - y^{(i)})(\mathbf{w}^t \cdot \mathbf{x}^{(i)} - \mathbf{w}^* \cdot \mathbf{x}^{(i)}) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y^* - y)(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}]\right| \geq s\right] \leq \frac{\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y^* - y)^2(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2]}{ms^2}.$$

Let $r = \frac{2W}{L}\log(2/(L^2\epsilon'))$, then by the fact that $\mathcal{D}_{\mathbf{x}}$ is sub-exponential, we have $\mathbf{Pr}[|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}| \geq r] \leq \epsilon'$. Furthermore, since $|y| \leq M$ where $M = \frac{bW}{L}\log(16b^4W^4/\epsilon^2)$, as stated in Fact D.3.3, the variance can be bounded as follows:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y^* - y)^2(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2] \\
\leq &\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y^* - y)^2(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}| \leq r\}] \\
&+ \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y^* - y)^2(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}| \geq r\}] \\
\leq &\, r^2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^*(\mathbf{w}^* \cdot \mathbf{x}) - y)^2] \\
&+ \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(2(u^*(\mathbf{w}^* \cdot \mathbf{x}))^2 + y^2)(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}| \geq r\}] \\
\leq &\, r^2\mathrm{OPT} + \mathbf{E}[2(b^2(\mathbf{w}^t \cdot \mathbf{x})^2 + M^2)(\mathbf{w}^t \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}| \geq r\}].
\end{aligned}$$

Since for any unit vectors $\mathbf{a}, \mathbf{b}$ we have $\mathbf{E}[(\mathbf{a} \cdot \mathbf{x})^4] \leq c^2/L^4$ and $\mathbf{E}[(\mathbf{a} \cdot \mathbf{x})^4(\mathbf{b} \cdot \mathbf{x})^4] \leq c^2/L^8$,

we have:

$$
\begin{aligned}
&2b^2\,\mathbf{E}[(\mathbf{w}^t\cdot\mathbf{x})^2(\mathbf{w}^t\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x}^2)^2\mathbb{1}\{|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\geq r\}]\\
&\leq 4b^2(W/L)^4\sqrt{\mathbf{E}[((\mathbf{w}^t/\|\mathbf{w}^t\|_2)\cdot\mathbf{x})^4(((\mathbf{w}^t-\mathbf{w}^*)/\|\mathbf{w}^t-\mathbf{w}^*\|_2)\cdot\mathbf{x})^4]\,\mathbf{Pr}[|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\geq r]}\\
&\leq 4cb^2(W/L)^4\sqrt{\epsilon'},
\end{aligned}
$$

and in addition,

$$
\begin{aligned}
&\mathbf{E}[M^2((\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x})^2\mathbb{1}\{|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\geq r\}]\\
&\leq 2M^2W^2\sqrt{\mathbf{E}[((\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x})^4]\,\mathbf{Pr}[|(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}|\geq r]}\leq cM^2(W/L)^2\sqrt{\epsilon'}.
\end{aligned}
$$

Let $s=(\mathrm{OPT}+\epsilon)/b$, $\epsilon'=\epsilon^2$, under our choice of $m\gtrsim db^4W^{9/2}\log^4(d/(\epsilon\delta))(1/\epsilon^{3/2}+1/(\epsilon\delta))$, it holds that

$$
\frac{1}{ms^2}\left(\frac{4W^2\log^2(1/(L^2\epsilon'))\mathrm{OPT}}{L^2}+(4cb^2(W/L)^4+cM^2(W/L)^2)\sqrt{\epsilon'}\right)\leq\delta.
$$

Thus, with probability at least $1-\delta$ it holds that

$$
\frac{1}{m}\sum_{i=1}^{m}(y^{*(i)}-y^{(i)})(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\mathbf{w}^*\cdot\mathbf{x}^{(i)})\geq\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-y^*)(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}]-(\mathrm{OPT}+\epsilon)/b.
$$

Since

$$
\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-y^*)(\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x}]\right|\leq\sqrt{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-y^*)^2]\,\mathbf{E}[((\mathbf{w}^t-\mathbf{w}^*)\cdot\mathbf{x})^2]}\leq\sqrt{\mathrm{OPT}}\|\mathbf{w}^*-\mathbf{w}^t\|_2,
$$

we finally have

$$\frac{1}{m}\sum_{i=1}^{m}(y^{*(i)}-y^{(i)})(\mathbf{w}^t\cdot\mathbf{x}^{(i)}-\mathbf{w}^*\cdot\mathbf{x}^{(i)})\geq-\sqrt{\mathrm{OPT}}\|\mathbf{w}^*-\mathbf{w}^t\|_2-(\mathrm{OPT}+\epsilon)/b,$$

completing the proof of Claim 6.2.9. □

Plugging the bounds from Claim 6.2.6, Claim 6.2.7, and Claim 6.2.9 back into (6.9) and using a union bound, we get that with probability at least $1-3\delta$,

$$\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t;\hat{u}^t)\cdot(\mathbf{w}^t-\mathbf{w}^*)\geq\frac{Ca^2LR^4}{b}\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2-2(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})\|\mathbf{w}^t-\mathbf{w}^*\|_2-2(\mathrm{OPT}+\epsilon)/b,$$

for some absolute constant $C$, completing the proof. □

## 6.3 Robust SIM Learning via Alignment Sharpness

Our algorithm can be viewed as employing an alternating procedure: taking a Riemannian gradient descent step on a sphere with respect to the empirical surrogate, given an estimate of the activation, and optimizing the activation function on the sample set for a given parameter weight vector. This procedure is performed using a fine grid of guesses of the scale of $\|\mathbf{w}^*\|_2$. For this process to converge with the desired linear rate (even for a known value of $\|\mathbf{w}^*\|_2$), the algorithm needs to be properly initialized to ensure that the initial weight vector has a nontrivial alignment with the optimal vector $\mathbf{w}^*$. The initialization process is handled in the following subsection.

### 6.3.1 Initialization

We begin by showing that the Initialization subroutine stated in Algorithm 10 returns a point $\bar{\mathbf{w}}^0$ that has a sufficient alignment with $\mathbf{w}^*$. As will become apparent later in the proof of Theorem 6.3.2, this property of the initial point is critical for Theorem 6.3.2 to converge at a linear rate.

**Algorithm 10** Initialization

**Input:** Initial $\mathbf{w}^0 = 0$; tolerances $\epsilon, \delta > 0$; parameters $a, b, L, R, W$; choose $\mu \lesssim \frac{a^2 L R^4}{b}$, step-size $\eta = \frac{\mu^3}{2^7 b^4}$, and iteration count $t_0 \lesssim \left(\frac{b}{\mu}\right)^6 \log\left(\frac{b}{\mu}\right)$.
**for** $t = 0$ **to** $t_0$ **do**
  Draw $m_0 \gtrsim \frac{W^{9/2} b^{10} d \log^4(d/(\epsilon\delta))}{L^4 \mu^6 \delta \epsilon^{3/2}}$ i.i.d. samples from $\mathcal{D}$.
  $\hat{u}^t \leftarrow \arg\min_{u \in \mathcal{U}} \frac{1}{m_0} \sum_{i=1}^{m_0} \left(u(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - y^{(i)}\right)^2$
  $\nabla \widehat{\mathcal{L}}^{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) \leftarrow \frac{1}{m_0} \sum_{i=1}^{m_0} \left(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - y^{(i)}\right) \mathbf{x}^{(i)}$
  $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta \nabla \widehat{\mathcal{L}}^{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)$
**end for**
**Return:** $\{\mathbf{w}^0, \ldots, \mathbf{w}^{t_0}\}$

**Lemma 6.3.1** (Initialization)**.** *Let $\mu = Ca^2 LR^4/b$ for an absolute constant $C > 0$ and let $\epsilon, \delta > 0$. Choose the step size $\eta = \mu^3/(2^7 b^4)$ in Algorithm 10. Then, drawing $m_0$ i.i.d. samples from $\mathcal{D}$ at each iteration such that*

$$m_0 \gtrsim \frac{W^{9/2} b^{10} d \log^4(d/(\epsilon\delta))}{L^4 \mu^6 \delta \epsilon^{3/2}},$$

*ensures that within $t_0 \lesssim b^6 \log(b/\mu)/\mu^6$ iterations, the initialization subroutine Algorithm 10 generates a list of size $t_0$ that contains a point $\bar{\mathbf{w}}^0$ such that $\|(\mathbf{w}^*)^{\perp_{\bar{\mathbf{w}}^0}}\|_2 \leq \max\{\mu\|\mathbf{w}^*\|_2/(4b), 64b^2/\mu^3(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\}$, with probability at least $1 - \delta$. The total number of samples required for Algorithm 10 is $N_0 = t_0 m_0$.*

*Proof.* Consider first the case that $\|\mathbf{w}^*\|_2 \leq 64b^2/\mu^3(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$. Then, for the parameter vector $\mathbf{w}^0 = 0$, we have

$$\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^0}}\|_2 = \|\mathbf{w}^*\|_2 \leq 64b^2/\mu^3(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$$

and the claimed statement holds trivially.

Thus, in the rest of the proof we assume $\|\mathbf{w}^*\|_2 \geq 64b^2/\mu^3(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$. Let $\mathbf{v}^t$ denote the component of $\mathbf{w}^*$ that is orthogonal to $\mathbf{w}^t$; i.e., $\mathbf{v}^t = \mathbf{w}^* - (\mathbf{w}^* \cdot \mathbf{w}^t)\mathbf{w}^t/\|\mathbf{w}^t\|_2^2 = (\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}$, where $\mathbf{w}^t$ is defined in Algorithm 10. Our goal is to show that when $\|\mathbf{v}^t\|_2 \geq \mu\|\mathbf{w}^*\|_2/(4b)$ at iteration $t$, the distance between $\mathbf{w}^{t+1}$ and $\mathbf{w}^*$ contracts by a constant factor $1-c$ for some $c < 1$, i.e., $\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq (1-c)\|\mathbf{w}^t - \mathbf{w}^*\|_2$. This implies that when $\|\mathbf{v}^t\|_2$ is greater than $\mu\|\mathbf{w}^*\|_2/(4b)$, $\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_2$ contracts until $\|\mathbf{v}^t\|_2 \geq \mu\|\mathbf{w}^*\|_2/(4b)$ is violated at step $t_0$; this $\mathbf{w}^{t_0}$ is exactly the initial point we are seeking to initialize the optimization subroutine.

Applying Proposition 6.2.1, we get that under our choice of batch size $m$, with probability at least $1 - \delta$, at each iteration it holds

$$\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\cdot(\mathbf{w}^t - \mathbf{w}^*) \geq \frac{Ca^2LR^4}{b}\|(\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}\|_2^2 - 2(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{w}^t - \mathbf{w}^*\|_2 - 2(\mathrm{OPT} + \epsilon)/b\,.$$

We now study the distance between $\mathbf{w}^{t+1}$ and $\mathbf{w}^*$, where $\mathbf{w}^{t+1}$ is updated from $\mathbf{w}^t$ according to Algorithm 10.

$$\begin{aligned}\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 &= \|\mathbf{w}^t - \eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) - \mathbf{w}^*\|_2^2 \\ &= \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \eta^2\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2^2 - 2\eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\cdot(\mathbf{w}^t - \mathbf{w}^*)\,.\end{aligned} \tag{6.33}$$

Applying Lemma 6.3.3 to (6.33), and plugging in Proposition 6.2.1, we get that under our choice of batch size $m$ it holds that with probability at least $1-\delta$,

$$\begin{aligned}\|\mathbf{w}^{t+1}-\mathbf{w}^*\|_2^2 &\leq \|\mathbf{w}^t-\mathbf{w}^*\|_2^2+\eta^2(10(\mathrm{OPT}+\epsilon)+4b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2)\\ &\quad+2\eta(2(\mathrm{OPT}+\epsilon)/b+2(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})\|\mathbf{w}^t-\mathbf{w}^*\|_2-\mu\|\mathbf{v}^t\|_2^2)\\ &\leq (1+4b^2\eta^2)\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+2\eta(2(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})\|\mathbf{w}^t-\mathbf{w}^*\|_2-\mu\|\mathbf{v}^t\|_2^2)\\ &\quad+5\eta(\mathrm{OPT}+\epsilon)\ , \end{aligned} \tag{6.34}$$

where $\mu = Ca^2LR^4/b$ and $C$ is an absolute constant. Note that in the last inequality we used that $\eta \leq 1/10$, hence $10\eta^2 \leq \eta$, and that $b \geq 1$.

When $t=0$, $\mathbf{v}^0=\mathbf{w}^*$, hence we have $\|\mathbf{v}^0\|_2 \geq \mu\|\mathbf{w}^*\|_2/(4b)$. Suppose that at iteration $t$, $\|\mathbf{v}^t\|_2 \geq \mu\|\mathbf{w}^*\|_2/(4b)$ is still valid. Then, (6.34) is transformed to:

$$\begin{aligned}\|\mathbf{w}^{t+1}-\mathbf{w}^*\|_2^2 &\leq (1+4b^2\eta^2)\|\mathbf{w}^t-\mathbf{w}^*\|_2^2+5\eta(\mathrm{OPT}+\epsilon)\\ &\quad+2\eta((\mu^3/(32b^2))\|\mathbf{w}^t-\mathbf{w}^*\|_2\|\mathbf{w}^*\|_2-(\mu^3/(16b^2))\|\mathbf{w}^*\|_2^2)\ . \end{aligned} \tag{6.35}$$

We use an inductive argument to show that at iteration $t$, $\|\mathbf{w}^t-\mathbf{w}^*\|_2 \leq \|\mathbf{w}^*\|_2$, which must eventually yield a contraction $\|\mathbf{w}^{t+1}-\mathbf{w}^*\|_2^2 \leq (1-c)\|\mathbf{w}^t-\mathbf{w}^*\|_2^2$ for some constant $c<1$. This condition $\|\mathbf{w}^t-\mathbf{w}^*\|_2 \leq \|\mathbf{w}^*\|_2$ certainly holds for the base case $t=0$ as $\mathbf{w}^0=0$, hence $\|\mathbf{w}^0-\mathbf{w}^*\|_2=\|\mathbf{w}^*\|_2$. Now, suppose $\|\mathbf{w}^t-\mathbf{w}^*\|_2 \leq \|\mathbf{w}^*\|_2$ holds for all

the iterations from 0 to $t$. Then, plugging $\eta = \mu^3/(2^7 b^4)$ into (6.35), we get:

$$
\begin{aligned}
&\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 \\
&\leq (1 + 4b^2\eta^2)\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 2\eta((\mu^3/(32b^2)) - (\mu^3/(16b^2)))\|\mathbf{w}^t - \mathbf{w}^*\|_2\|\mathbf{w}^*\|_2 + 5\eta(\mathrm{OPT} + \epsilon) \\
&\leq (1 + 4\eta^2 b^2 - 2\eta\mu^3/(32b^2))\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 5\mu^3/(2^7 b^4)(\mathrm{OPT} + \epsilon) \\
&\leq (1 - \mu^6/(2^{11} b^6))\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 5\mu^3/(2^7 b^4)(\mathrm{OPT} + \epsilon)\ .
\end{aligned}
$$

Since we have assumed $\sqrt{\mathrm{OPT}} + \sqrt{\epsilon} \leq \mu^3/(64b^2)\|\mathbf{w}^*\|_2$, it holds $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \geq \|\mathbf{v}^t\|_2 \geq \mu\|\mathbf{w}^*\|_2/(4b) \geq (16b/\mu^2)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, thus, we have (noting that $\mu \leq 1$):

$$
5\mu^3/(2^7 b^4)(\mathrm{OPT} + \epsilon) \leq 5\mu^3/(2^7 b^4)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})^2 \leq \mu^6/(2^{12} b^6)\|\mathbf{w}^t - \mathbf{w}^*\|_2\ .
$$

Therefore, combining the results above, we get:

$$
\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 \leq (1 - \mu^6/(2^{12} b^6))\|\mathbf{w}^t - \mathbf{w}^*\|_2^2\ ,
$$

for any iteration $t$ such that $\|\mathbf{v}^t\|_2 \geq \mu\|\mathbf{w}^*\|_2/(4b)$ holds. This validates the induction argument that $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^*\|_2$ for every $t = 0, \ldots, t_0$ and at the same time yields the desired contraction property of the sequence $\|\mathbf{w}^t - \mathbf{w}^*\|_2$, $t = 0, \ldots, t_0$. Now, since $\|\mathbf{w}^0 - \mathbf{w}^*\|_2 = \|\mathbf{w}^*\|_2$ and $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \geq \|\mathbf{v}^t\|_2$, we have

$$
\|\mathbf{v}^{t+1}\|_2^2 \leq (1 - \mu^6/(2^{12} b^6))^t\|\mathbf{w}^*\|_2^2 \leq \exp(-t\mu^6/(2^{12} b^6))\|\mathbf{w}^*\|_2^2\ .
$$

Thus, after at most $t_0 = 2^{12} b^6 \log(4b/\mu)/\mu^6$ iterations, it must hold that among all those vectors $\mathbf{v}^1, \ldots, \mathbf{v}^{t_0}$, there exists a vector $\mathbf{v}^{t_0^*}$ such that $\|\mathbf{v}^{t_0^*}\|_2 \leq \mu\|\mathbf{w}^*\|_2/(4b)$. Since

there are only a constant number of candidates, we can feed each one as the initialized input to the optimization subroutine Theorem 6.3.2. This will only result in a constant factor increase in the runtime and sample complexity.

Finally, recall that we need to draw

$$m \gtrsim \frac{W^{9/2} b^4 \log^4(d/(\epsilon\delta))}{L^4}\left(\frac{1}{\epsilon^{3/2}} + \frac{1}{\epsilon\delta}\right)$$

new samples at each iteration for (6.34) to hold with probability $1-\delta$, and the total number of iterations is $t_0$. Thus, applying a union bound, we know that the probability that (6.34) holds for all $t_0$ is $1 - t_0\delta$. Hence, choosing $\delta \leftarrow \delta t_0$, and noting that $t_0 \approx b^6/\mu^6 \log(b/\mu)$, it follows that setting the batch size to be

$$m_0 = \Theta\left(\frac{W^{9/2} b^4 \log^4(d/(\epsilon\delta))}{L^4}\left(\frac{1}{\epsilon^{3/2}} + \frac{b^6 \log(b/\mu)}{\mu^6 \epsilon\delta}\right)\right) = \Theta\left(\frac{W^{9/2} b^{10} d \log^4(d/(\epsilon\delta))}{L^4 \mu^6 \delta \epsilon^{3/2}}\right),$$

suffices and the total number of samples required for the initialization process is $t_0 m_0$. □

### 6.3.2 Optimization

Our main optimization algorithm is summarized in Algorithm 11. We now provide intuition for how guessing the value of $\|\mathbf{w}^*\|_2$ is used in the convergence analysis. Let $\mathbf{w}^t = \|\mathbf{w}^*\|_2 \bar{\mathbf{w}}^t / \|\bar{\mathbf{w}}^t\|_2$ so that $\|\mathbf{w}^t\|_2 = \|\mathbf{w}^*\|_2$ and let $\mathbf{v}^t := (\mathbf{w}^*)^{\perp_{\mathbf{w}^t}}$. Observe that $\|\mathbf{v}^t\|_2 = \|\mathbf{w}^t - \mathbf{w}^*\|_2 \cos(\theta(\mathbf{w}^t, \mathbf{w}^*)/2)$. Applying Proposition 6.2.1, it can be shown that $\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - C\|\mathbf{v}^t\|_2^2$ for some constant $C$. Thus, as long as the angle between $\mathbf{w}^t$ and $\mathbf{w}^*$ is not too large (ensured by initialization), $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \approx \|\mathbf{v}^t\|_2$. Hence, we can argue that $\|\mathbf{w}^t - \mathbf{w}^*\|_2$ contracts in each iteration, by observing that $\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 \approx \|\mathbf{v}^{t+1}\|_2^2 \leq \|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2$.

**Algorithm 11** Optimization

**Input:** initial $\mathbf{w}^{\text{ini}} = \mathbf{0}$; $\epsilon > 0$; parameters $a, b, L, R, W$; choose $\mu \lesssim \frac{a^2 LR^4}{b}$; step-size $\eta = \frac{\mu}{4b^2}$; iterations $T = O\Big((b/\mu)^2 \log(1/\epsilon)\Big)$.
$\{\mathbf{w}_0^{\text{ini}}, \ldots, \mathbf{w}_{t_0}^{\text{ini}}\} \leftarrow \text{Initialization}[\mathbf{w}^{\text{ini}}]$ (see Algorithm 10)
**Initialize:** $\mathcal{P} \leftarrow \{(\mathbf{w} = \mathbf{0}, u(z) = 0)\}$
**for** $k = 0$ **to** $t_0 \lesssim (b/\mu)^6 \log(b/\mu)$ **do**
  $\mathcal{P}_k \leftarrow \{\}$
  **for** $j = 1$ **to** $J = \frac{W}{\eta\sqrt{\epsilon}}$ **do**
    $\bar{\mathbf{w}}_{j,k}^0 \leftarrow \mathbf{w}_k^{\text{ini}}$
    $\beta_j \leftarrow j\,\eta\sqrt{\epsilon}$
    **for** $t = 0$ **to** $T - 1$ **do**
      $\widehat{\mathbf{w}}_{j,k}^t \leftarrow \beta_j \left(\bar{\mathbf{w}}_{j,k}^t / \|\bar{\mathbf{w}}_{j,k}^t\|_2\right)$
      Draw $m \gtrsim \frac{W^{11/2} b^{17}\, d\, \log^5(d/\epsilon)}{L^4 \mu^{12}\, \epsilon^{3/2}}$ fresh samples from $\mathcal{D}$
      $\hat{u}_{j,k}^t \leftarrow \arg\min_{u \in \mathcal{U}} \frac{1}{m} \sum_{i=1}^m \Big(u(\widehat{\mathbf{w}}_{j,k}^t \cdot \mathbf{x}^{(i)}) - y^{(i)}\Big)^2$
      $\nabla \widehat{\mathcal{L}}^{\text{sur}}(\widehat{\mathbf{w}}_{j,k}^t; \hat{u}_{j,k}^t) \leftarrow \frac{1}{m} \sum_{i=1}^m \Big(\hat{u}_{j,k}^t(\widehat{\mathbf{w}}_{j,k}^t \cdot \mathbf{x}^{(i)}) - y^{(i)}\Big)\, \mathbf{x}^{(i)}$
      $\bar{\mathbf{w}}_{j,k}^{t+1} \leftarrow \widehat{\mathbf{w}}_{j,k}^t - \eta\, \nabla \widehat{\mathcal{L}}^{\text{sur}}(\widehat{\mathbf{w}}_{j,k}^t; \hat{u}_{j,k}^t)$
    **end for**
    $\mathcal{P}_k \leftarrow \mathcal{P}_k \cup \{(\widehat{\mathbf{w}}_{j,k}^T,\ \hat{u}_{j,k}^T)\}$
  **end for**
  $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{P}_k$
**end for**
$(\widehat{\mathbf{w}}, \hat{u}) \leftarrow \text{Test}\Big[(\mathbf{w}, u) \in \mathcal{P}\Big]$ (see Algorithm 12)
**Return:** $(\widehat{\mathbf{w}}, \hat{u})$

**Theorem 6.3.2** (Main Result)**.** *Let $\mathcal{D}$ be a distribution in $\mathbb{R}^d \times \mathbb{R}$ and suppose that $\mathcal{D}_{\mathbf{x}}$ is $(L, R)$-well-behaved. Furthermore, let $\mathcal{U}_{(a,b)}$ be as in Definition 5.1.5, and $\epsilon > 0$. Let $\mu = Ca^2LR^4/b$, where $C$ is an absolute constant. Running Algorithm 11 with the following parameters: step size $\eta = \mu/(4b^2)$, batch size to be $m \gtrsim dW^{11/2}b^{17}\log^5(d/\epsilon)/(L^4\mu^{12}\epsilon^{3/2})$ and the total number of iterations to be $T' = t_0 JT = O(Wb^{11}/(\mu^{10}\sqrt{\epsilon})\log(1/\epsilon))$, where $T = O((b/\mu)^2\log(1/\epsilon))$, then with probability at least $2/3$, Algorithm 11 returns a*

*hypothesis* $(\hat{u}, \widehat{\mathbf{w}})$ *where* $\hat{u} \in \mathcal{U}_{(a,b)}$ *and* $\widehat{\mathbf{w}} \in \mathbb{B}(W)$ *such that*

$$\mathcal{L}_2(\widehat{\mathbf{w}}; \hat{u}) = O\left(\frac{b^4}{a^4 L^2 R^8}\right)\mathrm{OPT} + \epsilon \ ,$$

*using* $N = O(T'm) = \tilde{O}(dW^{13/2}b^{28}/(L^4\mu^{22}\epsilon^2))$ *samples.*

To prove Theorem 6.3.2, we make use of two technical results stated below. First, Lemma 6.3.3 provides an upper bound on the norm of the empirical gradient of the surrogate loss. The proof of the lemma relies on concentration properties of $(L, R)$-well behaved distributions $\mathcal{D}_{\mathbf{x}}$, and leverages the uniform convergence of the empirically-optimal activations $\hat{u}^t$.

**Lemma 6.3.3** (Upper Bound on Empirical Gradient Norm)**.** *Let* $S$ *be a set of i.i.d. samples of size* $m \gtrsim (dW^{9/2}b^4 \log^4(d/(\epsilon\delta))/L^4)(1/\epsilon^{3/2} + 1/(\epsilon\delta))$*. Given any* $\mathbf{w}^t \in \mathbb{B}(W)$*, let* $\hat{u}^t \in \mathcal{U}_{(a,b)}$ *be the solution of optimization problem* (P) *with respect to* $\mathbf{w}^t$ *and sample set* $S$*. Then, with probability at least* $1 - \delta$*, we have that* $\|\nabla \widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2^2 \leq 4b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 10(\mathrm{OPT} + \epsilon)$.

*Proof.* First we show the following:

**Lemma 6.3.4.** *Let* $S$ *be a sample set of* $m$ *i.i.d. samples of size at least* $m \gtrsim d\log^4(d/(\epsilon\delta))(b^2W^3/L^2\epsilon)^{3/2}$*. Furthermore, given* $\mathbf{w}^t \in \mathbb{B}(W)$*, let* $\hat{u}^t$ *be defined as in* (P)*. Then, it holds that with probability at least* $1 - \delta$*,*

$$\|\nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2^2 \leq 8(\mathrm{OPT} + \epsilon) + 2b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2.$$

*Proof.* By the definition of $\ell_2$ norms, we have:

$$\begin{aligned}
\|\nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 &= \max_{\|\mathbf{v}\|_2=1} \nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) \cdot \mathbf{v} \\
&= \max_{\|\mathbf{v}\|_2=1} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)\mathbf{v} \cdot \mathbf{x}] \\
&= \max_{\|\mathbf{v}\|_2=1} \Bigg\{ \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}) + u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))(\mathbf{v} \cdot \mathbf{x})] \\
&\quad + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}) + u^*(\mathbf{w}^* \cdot \mathbf{x}) - y)(\mathbf{v} \cdot \mathbf{x})] \Bigg\}.
\end{aligned}$$

By the Cauchy-Schwarz inequality, we further have:

$$\begin{aligned}
&\|\nabla \mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 \leq \\
&\max_{\|\mathbf{v}\|_2=1} \Bigg\{ \sqrt{\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2]\,\mathbf{E}[(\mathbf{v} \cdot \mathbf{x})^2]} + \sqrt{\mathbf{E}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2]\,\mathbf{E}[(\mathbf{v} \cdot \mathbf{x})^2]} \\
&\qquad + \sqrt{\mathbf{E}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]\,\mathbf{E}[(\mathbf{v} \cdot \mathbf{x})^2]} + \sqrt{\mathbf{E}[(u^*(\mathbf{w}^* \cdot \mathbf{x}) - y)^2]\,\mathbf{E}[(\mathbf{v} \cdot \mathbf{x})^2]} \Bigg\} \\
&\leq \underbrace{\sqrt{\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2]}}_{\mathcal{T}_1} + \underbrace{\sqrt{\mathbf{E}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2]}}_{\mathcal{T}_2} \\
&\quad + \underbrace{\sqrt{\mathbf{E}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]}}_{\mathcal{T}_3} + \underbrace{\sqrt{\mathbf{E}[(u^*(\mathbf{w}^* \cdot \mathbf{x}) - y)^2]}}_{\mathcal{T}_4},
\end{aligned}$$

where in the last inequality we used the assumption that $\mathbf{E}[\mathbf{x}\mathbf{x}^\intercal] \preccurlyeq \mathbf{I}$, hence $\mathbf{E}[(\mathbf{v}\cdot\mathbf{x})^2] \leq 1$. It remains to bound $\mathcal{T}_1 - \mathcal{T}_4$. Observe first that $\mathcal{T}_1 \leq \sqrt{\epsilon}$ for every $\mathbf{w}^t \in \mathbb{B}(W)$, with probability at least $1-\delta$, due to Lemma D.3.4. By definition, $\mathcal{T}_4 = \sqrt{\mathrm{OPT}}$. Recall that in Lemma 6.2.3 we showed the following $\mathcal{T}_2^2 = \mathbf{E}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \leq \mathrm{OPT}$. For $\mathcal{T}_3$, note that $u^{*t} \in \mathrm{argmin}_{u \in \mathcal{U}_{(a,b)}} \mathbf{E}[(u(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]$, therefore, since $u^* \in \mathcal{U}_{(a,b)}$,

we have

$$\mathcal{T}_3^2 = \mathbf{E}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq \mathbf{E}[(u^*(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq b^2 \|\mathbf{w}^t - \mathbf{w}^*\|_2^2,$$

after applying the assumption that $u^*$ is $b$-Lipschitz. Thus, in conclusion, we have

$$\|\nabla \mathcal{L}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 \leq 2\sqrt{\text{OPT}} + \sqrt{\epsilon} + b\|\mathbf{w}^t - \mathbf{w}^*\|_2.$$

Furthermore, since $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$, we get with probability at least $1 - \delta$:

$$\|\nabla \mathcal{L}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2^2 \leq 8\text{OPT} + 8\epsilon + 2b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2,$$

completing the proof of Lemma 6.3.4. □

We now prove that the distance between $\nabla \mathcal{L}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t)$ and $\nabla \widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t)$ is bounded by $b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \text{OPT} + \epsilon$ with high probability.

**Lemma 6.3.5.** *Let $S$ be a sample set of $m \gtrsim (dW^{9/2}b^4 \log^4(d/(\epsilon\delta))/L^4)(1/\epsilon^{3/2} + 1/(\epsilon\delta))$ i.i.d. samples. Given a vector $\mathbf{w}^t \in \mathbb{B}(W)$, it holds that with probability at least $1 - \delta$,*

$$\|\nabla \widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t) - \nabla \mathcal{L}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 \leq \sqrt{b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \text{OPT} + \epsilon}.$$

*Proof.* Since for any zero-mean independent random variable $\mathbf{z}_j$, we have $\mathbf{E}[||\sum_j \mathbf{z}_j||_2^2] = \sum_j \mathbf{E}[\|\mathbf{z}_j\|_2^2]$, by Chebyshev's inequality:

$$\mathbf{Pr}[\|\nabla \widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t) - \nabla \mathcal{L}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 \geq s] \leq \frac{1}{ms^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)\mathbf{x}\|_2^2] \,. \quad (6.36)$$

By linearity of expectation, we have:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)\mathbf{x}\|_2^2]=\sum_{k=1}^{d}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2(\mathbf{x}_k)^2],$$

where $\mathbf{x}_k=\mathbf{e}_k\cdot\mathbf{x}$ and $\mathbf{e}_k$ is the $k^{\text{th}}$ unit basis of $\mathbb{R}^d$. Let $r=O(W/L\log(1/(L\epsilon')))$, then it holds $\mathbf{Pr}[|\mathbf{x}_k|\geq r]\leq\epsilon'$. Then, the variance above can be decomposed into the following parts:

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2]=&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\geq r\}]\\&+\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\leq r\}].\end{aligned}$$

Since $|y|\leq M=O(bW/L\log(bW/\epsilon))$, and $\mathbf{E}[(\mathbf{w}^t\cdot\mathbf{x})^4\mathbf{x}_k^4]\leq W^4c^2/L^8$, $\mathbf{E}[\mathbf{x}_k^4]\leq c^2/L^4$ for $\mathcal{D}_{\mathbf{x}}$ is $L$-sub-exponential, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\geq r\}]&\leq 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}))^2+y^2)\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\geq r\}]\\&\leq 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(b(\mathbf{w}^t\cdot\mathbf{x}))^2+y^2)\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k|\geq r\}]\\&\leq 2b^2\sqrt{\mathbf{E}[((\mathbf{w}^t\cdot\mathbf{x})^4\mathbf{x}_k^4]\,\mathbf{Pr}[|\mathbf{x}_k|\geq r]}\\&\quad+2M^2\sqrt{\mathbf{E}[\mathbf{x}_k^4]\,\mathbf{Pr}[|\mathbf{x}_k|\geq r]}\\&\leq(2cb^2W^2/L^4)\sqrt{\epsilon'}+(2cM^2/L^2)\sqrt{\epsilon'}\\&\leq(4cM^2/L^2)\sqrt{\epsilon'}.\end{aligned}\tag{6.37}$$

In addition, $(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2$ can be decomposed as the following:

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2] \le 4\,\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2] + 4\,\mathbf{E}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \\ + 4\,\mathbf{E}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] + 4 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^*(\mathbf{w}^* \cdot \mathbf{x}) - y)^2].\end{aligned}$$

The first term is bounded above by $4\epsilon$ with probability at least $1-\delta$ for every $\mathbf{w}^t \in \mathbb{B}(W)$ whenever $m \gtrsim d\log^4(d/(\epsilon\delta))(b^2W^3/L^2\epsilon)^{3/2}$, as proved in Lemma D.3.4. The second term is smaller than 4OPT, which is shown in Lemma 6.2.3. The third term can be bounded above using again the definition of $u^{*t} = \operatorname{argmin}_{u\in\mathcal{U}_{(a,b)}} \mathbf{E}[(u(\mathbf{w}^t \cdot \mathbf{x}) - y^*)^2]$, as

$$4\,\mathbf{E}[(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \le 4\,\mathbf{E}[(u^*(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \le 4b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2,$$

using the fact that $u^*$ is $b$-Lipschitz and $\mathbf{E}[\mathbf{x}\mathbf{x}^\intercal] \preccurlyeq \mathbf{I}$. Lastly, the fourth term is bounded by 4OPT by the definition of $u^*(\mathbf{w}^* \cdot \mathbf{x})$. In summary, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2\mathbf{x}_k^2\mathbb{1}\{|\mathbf{x}_k| \le r\}] &\le r^2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2] \\ &\le 4r^2(b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 2\mathrm{OPT} + \epsilon),\end{aligned}$$

which, combining with (6.37), implies that the expectation $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2\mathbf{x}_k^2]$ is bounded by:

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2\mathbf{x}_k^2] &\le 4r^2b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + 4r^2(2\mathrm{OPT} + 2\epsilon) \\ &\le \frac{CW^2}{L^2}\log^2\left(\frac{b}{L\epsilon}\right)(b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \mathrm{OPT} + \epsilon),\end{aligned}$$

where $C$ is a large absolute constant. Note to get the inequality above we chose

$\epsilon' = C\epsilon^2(L/b)^4$, which then indicates that $4c(M/L)^2\sqrt{\epsilon'} \leq r^2\epsilon$. Summing the inequality above from $k = 1$ to $d$ delivers the final upper bound on the variance:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)\mathbf{x}\|_2^2] \leq \frac{dCW^2}{L^2} \log^2\left(\frac{b}{L\epsilon}\right)(b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \mathrm{OPT} + \epsilon).$$

Thus, plugging the upper bound on the variance above back to (6.36), as long as $m \gtrsim (dW^2/L^2)\log^2(b/(L\epsilon))/\delta$, we get with probability at least $1 - \delta$,

$$\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t) - \nabla\mathcal{L}_{\mathrm{sur}}(\mathbf{w}^t; \hat{u}^t)\|_2 \leq \sqrt{b^2\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \mathrm{OPT} + \epsilon}.$$

Noting that $m \gtrsim (dW^{9/2}b^4\log^4(d/(\epsilon\delta))/L^4)(1/\epsilon^{3/2} + 1/(\epsilon\delta))$ certainly satisfies the condition on $m$ above as $m \gtrsim (dW^2/L^2)\log^2(b/(L\epsilon))/\delta$, thus, we completed the proof of Lemma 6.3.5 □

Lemma 6.3.3 follows directly by combining Lemma 6.3.4, Lemma 6.3.5 and the triangle inequality. □

The following claim bounds the $L_2^2$ error of a hypothesis $\hat{u}_{\mathbf{w}}(\mathbf{w} \cdot \mathbf{x})$ by the distance between $\mathbf{w}$ and $\mathbf{w}^*$.

**Claim 6.3.6.** *Let* $\mathbf{w}$ *be any vector from* $\mathbb{B}$. *Let* $\hat{u}_{\mathbf{w}}$ *be a solution to* (P) *for a fixed parameter vector* $\mathbf{w} \in \mathbb{R}^d$ *with sample size* $m \gtrsim d\log^4(d/(\epsilon\delta))(b^2W^3/(L^2\epsilon))^{3/2}$. *Then*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}_{\mathbf{w}}(\mathbf{w} \cdot \mathbf{x}) - y)^2] \leq 8(\mathrm{OPT} + \epsilon) + 4b^2\|\mathbf{w} - \mathbf{w}^*\|_2^2.$$

*Proof.* Let $u_{\mathbf{w}}^*$, $u_{\mathbf{w}}$ be the optimal activations for problems (EP*) and (EP) under

parameter $\mathbf{w}$, respectively. Then, a direct calculation gives:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}_{\mathbf{w}}-y)^2] = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})+u_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u^*_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})\\
&+u^*_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x})+u^*(\mathbf{w}^*\cdot\mathbf{x})-y)^2]\\
&\le 4\,\mathbf{E}[(\hat{u}_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x}))^2]+4\,\mathbf{E}[(u_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u^*_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x}))^2]\\
&\quad+4\,\mathbf{E}[(u^*_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]+4\mathrm{OPT}\\
&\le 8(\mathrm{OPT}+\epsilon)+4b^2\|\mathbf{w}-\mathbf{w}^*\|_2^2, \qquad (6.38)
\end{aligned}$$

where in the second inequality we used the results from Lemma 6.2.3, Lemma D.3.4 and we applied the observation that:

$$\mathbf{E}[(u^*_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\le\mathbf{E}[(u^*(\mathbf{w}\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]\le b^2\|\mathbf{w}-\mathbf{w}^*\|_2^2,$$

by the definition of $u^*_{\mathbf{w}}$. □

*Proof.* As proved in Lemma 6.3.1, the initialization subroutine Algorithm 10 outputs a list of points $\{\mathbf{w}_k^{\mathrm{ini}}\}_{k=1}^{t_0}$ that contains a point $\mathbf{w}_{k^*}^{\mathrm{ini}}$ such that

$$\|(\mathbf{w}^*)^{\perp_{\mathbf{w}_{k^*}^{\mathrm{ini}}}}\|_2\le\max\{\mu\|\mathbf{w}^*\|_2/(4b),64b^2/\mu^3(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})\}.$$

Suppose first that $\mu\|\mathbf{w}^*\|_2/(4b)\le 64b^2/\mu^3(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})$. Then this implies that $\|\mathbf{w}^*\|_2\le 256b^3/\mu^4(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})$. Therefore, applying Claim 6.3.6 we immediately get that the trivial hypothesis $(\mathbf{w}=0,u(z)=0)$ works as a constant approximate solution, as in this case

$$\mathcal{L}_2(\mathbf{w};u)\le 8(\mathrm{OPT}+\epsilon)+4b^2\|\mathbf{w}^*\|_2=O((b/\mu)^8)\mathrm{OPT}+\epsilon.$$

This hypothesis ($\mathbf{w} = 0, u(z) = 0$) is contained in our solution set $\mathcal{P}$ and tested in Algorithm 12.

Thus, in the rest of the proof we assume that $\mathbf{w}_{k^*}^{\text{ini}}$ satisfies

$$\|(\mathbf{w}^*)^{\perp_{\mathbf{w}_{k^*}^{\text{ini}}}}\|_2 \leq \mu\|\mathbf{w}^*\|_2/(4b).$$

Let us consider this initialized parameter at $k^*$ step in the outer loop, $\bar{\mathbf{w}}_{j,k^*}^0 = \mathbf{w}_{k^*}^{\text{ini}}$. In the rest of the proof we drop the subscript $k^*$ since the context is clear.

Since we constructed a grid with grid width $\eta\sqrt{\epsilon}$ from 0 to $W$ to find the (approximate) value of $\|\mathbf{w}^*\|_2$, there must exist an index $j^*$ such that the value of $\beta_{j^*}$ is $\eta\sqrt{\epsilon}$ close to $\|\mathbf{w}^*\|_2$, i.e., $|\beta_{j^*} - \|\mathbf{w}^*\|_2| \leq \eta\sqrt{\epsilon}$. We now consider this $j^{*\text{th}}$ outer loop and ignore the subscript $j^*$ for simplicity. Let $\mathbf{w}^t = \|\mathbf{w}^*\|_2(\bar{\mathbf{w}}^t/\|\bar{\mathbf{w}}^t\|_2)$, which is the true normalized vector of $\bar{\mathbf{w}}^t$ that has no error.

We study the squared distance between $\bar{\mathbf{w}}^{t+1}$ and $\mathbf{w}^*$:

$$\begin{aligned}\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 &= \|\widehat{\mathbf{w}}^t - \eta\nabla\widehat{\mathcal{L}}_{\text{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t) - \mathbf{w}^*\|_2^2 \\ &= \|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2^2 + \eta^2\|\nabla\widehat{\mathcal{L}}_{\text{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t)\|_2^2 - 2\eta\nabla\widehat{\mathcal{L}}_{\text{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t) \cdot (\widehat{\mathbf{w}}^t - \mathbf{w}^*).\end{aligned} \tag{6.39}$$

Applying Lemma 6.3.3 to (6.39), and plugging in Proposition 6.2.1, we get that when drawing

$$m \gtrsim \frac{dW^{9/2}b^4\log^4(d/(\epsilon\delta))}{L^4}\left(\frac{1}{\epsilon^{3/2}} + \frac{1}{\epsilon\delta}\right), \tag{6.40}$$

samples from the distribution, it holds with probability at least $1-\delta$ that:

$$\begin{aligned}\|\bar{\mathbf{w}}^{t+1}-\mathbf{w}^*\|_2^2 \le \|\widehat{\mathbf{w}}^t-\mathbf{w}^*\|_2^2 + \eta^2(10(\mathrm{OPT}+\epsilon)+4b^2\|\widehat{\mathbf{w}}^t-\mathbf{w}^*\|_2^2) \\ + 2\eta(2(\mathrm{OPT}+\epsilon)/b + 2(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})\|\widehat{\mathbf{w}}^t-\mathbf{w}^*\|_2 - \mu\|\mathbf{v}^t\|_2^2),\end{aligned} \tag{6.41}$$

where $\mu = Ca^2LR^4/b$ with $C$ being an absolute constant, and where $\mathbf{v}^t$ is the component of $\mathbf{w}^*$ that is orthogonal to $\widehat{\mathbf{w}}^t$, i.e.,

$$\mathbf{v}^t = \mathbf{w}^* - (\mathbf{w}^*\cdot\widehat{\mathbf{w}}^t)\widehat{\mathbf{w}}^t/\|\widehat{\mathbf{w}}^t\|_2^2 = (\mathbf{w}^*)^{\perp_{\widehat{\mathbf{w}}^t}}.$$

Note that $\|\mathbf{v}^t\|_2$ is invariant to the rescaling of $\widehat{\mathbf{w}}^t$, in other words, $\mathbf{w}^*$ has the same orthogonal component $\mathbf{v}^t$ for all $\bar{\mathbf{w}}^t$, $\mathbf{w}^t$ and $\widehat{\mathbf{w}}^t$.

Since $\|\widehat{\mathbf{w}}^t-\mathbf{w}^t\|_2 \le \eta\sqrt{\epsilon}$, we have

$$\|\widehat{\mathbf{w}}^t-\mathbf{w}^*\|_2^2 = \|\widehat{\mathbf{w}}^t-\mathbf{w}^t+\mathbf{w}^t-\mathbf{w}^*\|_2^2 \le \|\mathbf{w}^t-\mathbf{w}^*\|_2^2 + \eta^2\epsilon + 2\eta\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2. \tag{6.42}$$

In addition, by triangle inequality we have $\|\widehat{\mathbf{w}}^t-\mathbf{w}^*\|_2 \le \|\mathbf{w}^t-\mathbf{w}^*\|_2 + \eta\sqrt{\epsilon}$. Therefore, substituting $\mathbf{w}^t$ with $\widehat{\mathbf{w}}^t$ in (6.41), we get:

$$\begin{aligned}\|\bar{\mathbf{w}}^{t+1}-\mathbf{w}^*\|_2^2 &\le \|\mathbf{w}^t-\mathbf{w}^*\|_2^2 + \eta^2\epsilon + 2\eta\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2 \\ &\quad + \eta^2(10(\mathrm{OPT}+\epsilon)+4b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2 + 4b^2\eta^2\epsilon + 8b^2\eta\sqrt{\epsilon}\|\mathbf{w}^t-\mathbf{w}^*\|_2) \\ &\quad + 2\eta(2(\mathrm{OPT}+\epsilon)/b + 2(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})(\|\mathbf{w}^t-\mathbf{w}^*\|_2+\eta\sqrt{\epsilon}) - \mu\|\mathbf{v}^t\|_2^2) \\ &\le \|\mathbf{w}^t-\mathbf{w}^*\|_2^2 + \eta^2(24(\mathrm{OPT}+\epsilon)+4b^2\|\mathbf{w}^t-\mathbf{w}^*\|_2^2) \\ &\quad + 2\eta(2(\mathrm{OPT}+\epsilon)/b + 4(\sqrt{\mathrm{OPT}}+\sqrt{\epsilon})\|\mathbf{w}^t-\mathbf{w}^*\|_2 - \mu\|\mathbf{v}^t\|_2^2),\end{aligned} \tag{6.43}$$

where we used $4b^2\eta^2 \leq 1$, which holds because $\eta = \mu/(4b^2)$.

Our goal is to show that $\|\mathbf{v}^{t+1}\|_2^2 \leq \|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \leq (1-c)\|\mathbf{v}^t\|_2^2 + \epsilon$, where $c \in (0,1)$ is a constant and $\epsilon$ is a small error parameter. However, this linear contraction can only be obtained when $\|\mathbf{v}^t\|_2$ is relatively small compared to $\|\mathbf{w}^*\|_2$. Specifically, as will be manifested in Claim 6.3.7 and the proceeding proof, the linear contraction is achieved only when $\|\mathbf{v}^t\|_2 \leq \mu\|\mathbf{w}^*\|_2/(4b)$. Luckily, we can start with a $\mathbf{v}^0$ such that this condition is satisfied, due to the initialization subroutine Algorithm 10, as proved in Lemma 6.3.1. We prove the following claim.

**Claim 6.3.7.** *Let $\eta = \mu/(4b^2)$. Then, under the assumptions of Theorem 6.3.2, with probability at least $1-\delta$, we have*

$$\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \leq \left(1 - \frac{\mu^2}{32b^2}\right)\|\mathbf{v}^t\|_2^2,$$

*whenever $\|\mathbf{v}^t\|_2 \geq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$.*

*Proof of Claim 6.3.7.* Since the norm of $\mathbf{w}^t$ is normalized to $\mathbf{w}^*$, the quantity $\|\mathbf{w}^t - \mathbf{w}^*\|^2$ is controlled by $\|\mathbf{v}^t\|_2^2$. In particular, let $\mathbf{w}^* = \alpha_t \mathbf{w}^t + \mathbf{v}^t$. Then, since $\mathbf{v}^t \perp \mathbf{w}^t$, we have $\|\mathbf{w}^*\|_2^2 = \alpha_t^2\|\mathbf{w}^t\|_2^2 + \|\mathbf{v}^t\|_2^2 = \alpha_t^2\|\mathbf{w}^*\|_2^2 + \|\mathbf{v}^t\|_2^2$, thus, $\alpha_t^2 = 1 - \|\mathbf{v}^t\|_2^2/\|\mathbf{w}^*\|_2^2$, and $\|\mathbf{v}^t\|_2^2 = (1-\alpha_t^2)\|\mathbf{w}^*\|_2^2$. In addition, $\|\mathbf{w}^t - \mathbf{w}^*\|_2^2$ can be expressed as a function of $\alpha_t$ and $\mathbf{w}^*$, as

$$\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 = (1-\alpha_t)^2\|\mathbf{w}^*\|_2^2 + \|\mathbf{v}^t\|_2^2 = 2(1-\alpha_t)\|\mathbf{w}^*\|_2^2. \tag{6.44}$$

Note that since $\alpha_t = \sqrt{1 - \|\mathbf{v}^t\|_2^2/\|\mathbf{w}^*\|_2^2}$, denoting $\rho_t = \|\mathbf{v}^t\|_2/\|\mathbf{w}^*\|_2$, we further have:

$$1 - \alpha_t = 1 - \sqrt{1 - \|\mathbf{v}^t\|_2^2/\|\mathbf{w}^*\|_2^2} = 1 - \sqrt{1-\rho_t^2} \leq \frac{1}{2}\rho_t^2 + \frac{1}{2}\rho_t^4 \leq \rho_t^2, \ \forall \rho_t \in [0,1]. \tag{6.45}$$

Therefore, plugging (6.44) and (6.45) back into (6.43), we get:

$$
\begin{aligned}
&\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \\
&\leq 2(1-\alpha_t)\|\mathbf{w}^*\|_2^2 + 4b^2\eta^2(2(1-\alpha_t)\|\mathbf{w}^*\|_2^2) + 8\eta(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\sqrt{2(1-\alpha_t)}\|\mathbf{w}^*\|_2 \\
&- 2\eta\mu\|\mathbf{v}^t\|_2^2 + 24\eta^2(\mathrm{OPT} + \epsilon) + 4\eta(\mathrm{OPT} + \epsilon)/b \\
&\leq (\rho_t^2 + \rho_t^4)\|\mathbf{w}^*\|_2^2 + 4b^2\eta^2(\rho_t^2 + \rho_t^4)\|\mathbf{w}^*\|_2^2 + 8\sqrt{2}\eta(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\rho_t\|\mathbf{w}^*\|_2 \\
&- 2\eta\mu\|\mathbf{v}^t\|_2^2 + 24\eta^2(\mathrm{OPT} + \epsilon) + 4\eta(\mathrm{OPT} + \epsilon)/b \\
&= (1 + \rho_t^2 + 4b^2\eta^2(1+\rho_t^2))\|\mathbf{v}^t\|_2^2 + 12\eta(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{v}^t\|_2 - 2\eta\mu\|\mathbf{v}^t\|_2^2 \\
&+ 4(6\eta^2 + \eta/b)(\mathrm{OPT} + \epsilon) \\
&\leq (1 + \rho_t^2 + 4b^2\eta^2(1+\rho_t^2))\|\mathbf{v}^t\|_2^2 + 12\eta(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{v}^t\|_2 - 2\eta\mu\|\mathbf{v}^t\|_2^2 + 5\eta(\mathrm{OPT} + \epsilon),
\end{aligned}
\tag{6.46}
$$

where in the last inequality we observed that since $\eta = \frac{\mu}{4b^2}$, it holds that $24\eta \leq 1$, as $\mu$ is small and $b \geq 1$.

Note that we have assumed that $\|\mathbf{v}^t\|_2 \geq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, which indicates

$$
12\eta(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})\|\mathbf{v}^t\|_2 \leq \frac{1}{8}\eta\mu\|\mathbf{v}^t\|_2^2,
$$

since $b \geq 1$ was assumed without loss of generality. Furthermore, when $\|\mathbf{v}^t\|_2 \geq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, it also holds that

$$
\frac{1}{8}\eta\mu\|\mathbf{v}^t\|_2^2 \geq \frac{(96)^2}{8\mu^2}\eta\mu(\mathrm{OPT} + \epsilon) \geq 5\eta(\mathrm{OPT} + \epsilon),
$$

since we have assumed $\mu = Ca^2LR^4/b \leq 1$ without loss of generality. Finally, as we will show in the rest of the proof, it holds that $\|\mathbf{v}^{t+1}\|_2 \leq \|\mathbf{v}^t\|_2$ for $t = 0, 1, \ldots, T$, thus as

$\eta = \mu/(4b^2)$, we have $\|\mathbf{v}^t\|_2 \leq \sqrt{\eta\mu}\|\mathbf{w}^*\|_2/2 = \mu\|\mathbf{w}^*\|_2/(4b)$, since $\|\mathbf{v}^0\|_2 \leq \sqrt{\eta\mu}\|\mathbf{w}^*\|_2/2$. This condition guarantees that

$$\rho_t^2 = \|\mathbf{v}^t\|_2^2/\|\mathbf{w}^*\|_2^2 \leq \frac{1}{4}\eta\mu.$$

Plugging these conditions back into (6.46), it is then simplified as (note that $1 + \rho_t^2 \leq 1 + (1/4)\eta\mu \leq 9/8$ for $\eta\mu \leq 1/2$):

$$\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \leq \left(1 + \frac{9}{2}b^2\eta^2 - \frac{3}{2}\eta\mu\right)\|\mathbf{v}^t\|_2^2.$$

Therefore, when $\eta = \mu/(4b^2)$ we have

$$\|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2 \leq \left(1 - \frac{\mu^2}{32b^2}\right)\|\mathbf{v}^t\|_2^2,$$

completing the proof. □

We proceed first under the condition that $\|\mathbf{v}^t\|_2 \geq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$ holds for $t = 0, \ldots, T$ and show that after some certain number of iterations $T$ this condition must be violated. Observe that if $\|\mathbf{v}^t\|_2 \leq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, then it holds $\|\mathbf{w}^t - \mathbf{w}^*\|_2^2 \lesssim (1/\mu^2)(\mathrm{OPT} + \epsilon)$, implying that $\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x})$ is a hypothesis achieving constant approximation error according to Claim 6.3.6, hence the algorithm can be terminated. However, note that $T$ only works as an upper bound for the iteration complexity of our algorithm, and it is possible that the condition $\|\mathbf{v}^t\|_2 \geq (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$ is violated at some step $t^* < T$. However, as we show later, the value of $\|\mathbf{v}^T\|_2$ cannot be larger

than $c\|\mathbf{v}^{t^*}\|_2$, where $c$ is an absolute constant. We observe that:

$$\mathbf{v}^{t+1} = \mathbf{w}^* - (\mathbf{w}^* \cdot \mathbf{w}^{t+1})\mathbf{w}^{t+1}/\|\mathbf{w}^{t+1}\|_2^2 = \mathbf{w}^* - (\mathbf{w}^* \cdot \bar{\mathbf{w}}^{t+1})\bar{\mathbf{w}}^{t+1}/\|\bar{\mathbf{w}}^{t+1}\|_2^2 = (\mathbf{w}^*)^{\perp_{\bar{\mathbf{w}}^{t+1}}},$$

therefore, $\|\mathbf{v}^{t+1}\|_2^2 \le \|\bar{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2^2$, which, combined with Claim 6.3.7, yields

$$\|\mathbf{v}^{t+1}\|_2^2 \le \left(1 - \frac{\mu^2}{32b^2}\right)\|\mathbf{v}^t\|_2^2 \le \left(1 - \frac{\mu^2}{32b^2}\right)^t \|\mathbf{v}^0\|_2^2 \le \exp\left(-\frac{\mu^2 t}{32b^2}\right)2W^2.$$

The above contraction only holds when $\|\mathbf{v}^t\|_2 \ge (96/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$. Hence, after at most

$$T = O\left(\frac{b^2}{\mu^2}\log\left(\frac{\mu W}{\epsilon}\right)\right)$$

inner iterations, the algorithm outputs a vector $\mathbf{w}^{t^*}$ with $\|\mathbf{v}^{t^*}\|_2 \le \frac{96}{\mu}(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, where $t^* \in [T]$.

Now suppose that at step $t^* < T$ it holds that $\|\mathbf{v}^{t^*}\|_2 \le 96(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})/\mu$ but at the next iteration $\|\mathbf{v}^{t^*+1}\|_2 \ge 96(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})/\mu$. Recall first that in Lemma 6.3.3 we showed that $\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^t; \hat{u}^t)\|_2^2 \le 4b^2\|\widehat{\mathbf{w}}^t - \mathbf{w}^*\|_2^2 + 10(\mathrm{OPT} + \epsilon)$. Therefore, revisiting the updating scheme of the algorithm we have

$$\begin{aligned}
\|\mathbf{v}^{t^*+1}\|_2^2 \le \|\bar{\mathbf{w}}^{t^*+1} - \mathbf{w}^*\|_2^2 &= \|\widehat{\mathbf{w}}^{t^*} - \eta\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^{t^*}; \hat{u}^{t^*}) - \mathbf{w}^*\|_2^2 \\
&\le 2\|\widehat{\mathbf{w}}^{t^*} - \mathbf{w}^*\|_2^2 + 2\eta^2\|\nabla\widehat{\mathcal{L}}_{\mathrm{sur}}(\widehat{\mathbf{w}}^{t^*}; \hat{u}^{t^*})\|_2^2 \\
&\le (2 + 8b^2\eta^2)\|\widehat{\mathbf{w}}^{t^*} - \mathbf{w}^*\|_2^2 + 20\eta^2(\mathrm{OPT} + \epsilon) \\
&\le 3\|\widehat{\mathbf{w}}^{t^*} - \mathbf{w}^*\|_2^2 + (\mathrm{OPT} + \epsilon),
\end{aligned}$$

where in the last inequality we plugged in the value of $\eta = \mu/(4b^2)$, and used the assumption that $\mu \le 1$ and $b \ge 1$, hence $20\eta^2 \le 1$ and $8b^2\eta^2 \le 1$. Furthermore, recall

that by the construction of the grid, $\|\widehat{\mathbf{w}}^{t^*} - \mathbf{w}^t\|_2 \leq \eta\sqrt{\epsilon}$, implying that $\|\widehat{\mathbf{w}}^{t^*} - \mathbf{w}^*\|_2^2 \leq 2\|\mathbf{w}^{t^*} - \mathbf{w}^*\|_2^2 + 2\eta^2\epsilon$ by triangle inequality. Therefore, going back to the inequality of $\|\mathbf{v}^{t^*+1}\|_2^2$ above, we get

$$\|\mathbf{v}^{t^*+1}\|_2^2 \leq 6\|\mathbf{w}^{t^*} - \mathbf{w}^*\|_2^2 + 6\eta^2\epsilon + \mathrm{OPT} + \epsilon \leq 6\|\mathbf{w}^{t^*} - \mathbf{w}^*\|_2^2 + 2(\mathrm{OPT} + \epsilon).$$

Finally, observe that since $\|\mathbf{w}^{t^*}\|_2 = \|\mathbf{w}^*\|_2$, it holds $\|\mathbf{w}^{t^*} - \mathbf{w}^*\|_2 \leq \sqrt{2}\|\mathbf{v}^{t^*}\|_2$, hence, we get

$$\|\mathbf{v}^{t^*+1}\|_2^2 \leq 12\|\mathbf{v}^{t^*}\|_2^2 + 2(\mathrm{OPT} + \epsilon).$$

Now since $\|\mathbf{v}^{t^*+1}\|_2 \geq 96(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})/\mu$, the value of $\|\mathbf{v}^t\|_2^2$ will start to decrease again for $t \geq t^* + 1$. This implies that the value of $\|\mathbf{v}^T\|_2$ satisfies

$$\|\mathbf{v}^T\|_2 \leq \sqrt{12}\|\mathbf{v}^{t^*}\|_2 + \sqrt{2}(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon}) \leq \frac{384}{\mu}(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon}).$$

Combining Claim 6.3.6 and Lemma D.3.4, as we have guaranteed that $\|\mathbf{v}^T\|_2 \leq (384/\mu)(\sqrt{\mathrm{OPT}} + \sqrt{\epsilon})$, the hypothesis $\hat{u}^T(\widehat{\mathbf{w}}^T \cdot \mathbf{x})$ has the $L_2^2$ error that can be bounded as:

$$\mathcal{L}_2(\widehat{\mathbf{w}}^T; \hat{u}^T) \leq 6\mathrm{OPT} + 3b^2(4\|\mathbf{v}^T\|_2^2 + \eta^2\epsilon) + \epsilon = O\left(\frac{b^2}{\mu^2}(\mathrm{OPT} + \epsilon)\right).$$

For any $\epsilon_1 > 0$, setting $\epsilon = C'(\mu^2/b^2)\epsilon_1$ with $C'$ being some small universal absolute constant, we finally get $\mathcal{L}_2(\widehat{\mathbf{w}}^T; \hat{u}^T) \leq O((b^2/\mu^2)\mathrm{OPT}) + \epsilon_1$.

It still remains to determine the batch size as drawing a sample set of size $m$ as displayed in (6.40) only guarantees that the contraction of $\|\mathbf{v}^t\|_2$ at step $t$ holds with probability $1 - \delta$. Applying a union bound on all $t_0 JT = O(\frac{Wb^{10}}{\mu^9\sqrt{\epsilon}}\log(1/\epsilon)) = O(\frac{Wb^{11}}{\mu^{10}\sqrt{\epsilon_1}}\log(1/\epsilon_1))$ iterations yields that the contraction holds at every step with prob-

ability at least $1 - t_0 JT\delta$. Therefore, setting $\delta \leftarrow \delta(t_0 JT)$ and bringing the value of $\delta$ back to (6.40), we get that it suffices to choose the batch size as:

$$m = \Theta\left(\frac{dW^{9/2}b^4 \log^4(d/(\epsilon\delta))}{L^4}\left(\frac{1}{\epsilon^{3/2}} + \frac{Wb^{10}}{\mu^9\epsilon^{3/2}\delta}\right)\right) = \Theta\left(\frac{dW^{11/2}b^{17}\log^5(d/(\epsilon_1\delta))}{L^4\mu^{12}\delta\epsilon_1^{3/2}}\right),$$

to guarantee that we get an $O(\mathrm{OPT}) + \epsilon_1$-solution with probability at least $1 - \delta$. Note that we have set $\epsilon = C'(\mu^2/b^2)\epsilon_1$ in the last equality above.

The argument above justifies the claim that among all $t_0 J = Wb^7 \log(b/\mu)/(\eta\mu^7\sqrt{\epsilon_1})$ hypotheses in $\mathcal{P} = \{(\widehat{\mathbf{w}}_j^T; \hat{u}_j^T)\}_{j=1}^{t_0 J}$, there exists at least one hypothesis that achieves $L_2^2$ error $O(\mathrm{OPT}) + \epsilon_1$. To select the correct hypothesis from the set $\mathcal{P}$, one only needs to draw a new batch of $m' = \tilde{\Theta}(b^4 W^4 \log(1/\delta)/(L^4\epsilon_1^2))$ i.i.d. samples from $\mathcal{D}$, and choose the hypothesis from $\mathcal{P}$ that achieves the minimal empirical error defined in Algorithm 11. As discussed in Section (6.3.3), this procedure introduces an error at most $\epsilon_1$.

In conclusion, it holds by a union bound that Algorithm 11 delivers a solution with $O(\mathrm{OPT}) + \epsilon_1$ error with probability at least $1 - 2\delta$. The total sample complexity of our algorithm is

$$\begin{aligned} N = t_0 JTm + m' &= \Theta\left(\frac{W^{13/2}b^{28}d\log^5(d/(\epsilon_1\delta))}{L^4\mu^{22}\delta\epsilon_1^2} + \frac{b^4W^4\log(1/\delta)\log^5(1/\epsilon_1)}{L^4\epsilon_1^2}\right) \\ &= \Theta\left(\frac{W^{13/2}b^{28}d\log^6(d/(\epsilon_1\delta))}{L^4\mu^{22}\delta\epsilon_1^2}\right). \end{aligned}$$

Choosing $\delta = 1/6$ above we get that the Algorithm 11 succeeds to generate an $O(\mathrm{OPT}) + \epsilon_1$-solution for any $\epsilon_1 > 0$ with probability at least $1 - 2\delta = 2/3$, hence replacing $\epsilon_1$ with $\epsilon$ completes the proof of Theorem 6.3.2. □

### 6.3.3 Testing

We now briefly discuss the testing procedure, which allows our algorithm to select a hypothesis with minimum empirical error while maintaining validity of the claims. This part relies on standard arguments and is provided for completeness. Concretely, we rely on the following claim.

**Algorithm 12** Testing

1: **Input:** $\epsilon > 0$; parameters $a, b, L, R, W$; candidate list $\mathcal{P}$; radius $r \gtrsim \frac{1}{L}\log\Big(\frac{bW}{L\epsilon}\log^2\frac{1}{\epsilon}\Big)$

2: Draw $m' \gtrsim \frac{(bW/L)^4\log^5(1/\epsilon)}{\epsilon^2}$ i.i.d. samples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{m'} \sim \mathcal{D}$

3: $(\widehat{\mathbf{w}}, \hat{u}) \leftarrow \arg\min_{(\mathbf{w},u)\in\mathcal{P}} \frac{1}{m'}\sum_{i=1}^{m'}\Big(u(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)}\Big)^2 \mathbf{1}\{|\mathbf{w}\cdot\mathbf{x}^{(i)}| \le W\, r\}$

4: **Return:** $(\widehat{\mathbf{w}}, \hat{u})$

**Claim 6.3.8.** *Let $\mu$, $\epsilon_1$, $\delta \in (0,1)$ be fixed. Let $r = \frac{1}{L}\log(\frac{Cb^4W^4}{L^6\epsilon_1^2}\log^2(\frac{bW}{\epsilon_1}))$, where $C$ is a sufficiently large absolute constant. Given a set of parameter-activation pairs $\mathcal{P} = \{(\mathbf{w}_j; u_j)\}_{j=1}^{t_0 J}$ such that $\mathbf{w}_j \in \mathbb{B}(W)$ and $u_j \in \mathcal{U}_{(a,b)}$ for $j \in [t_0 J]$, where $t_0 J = 4b^9W/(\mu^8\sqrt{\epsilon_1})$, we have that using*

$$m' = \Theta\bigg(\frac{b^4W^4\log(1/\delta)}{L^4\epsilon_1^2}\log^5\left(\frac{bW}{L\mu\epsilon_1}\right)\bigg),$$

*i.i.d. samples from $\mathcal{D}$, for any $(\mathbf{w}_j; u_j) \in \mathcal{P}$ it holds with probability at least $1-\delta$,*

$$\bigg|\frac{1}{m'}\sum_{i=1}^{m'}(u_j(\mathbf{w}_j\cdot\mathbf{x}^{(i)}) - y^{(i)})^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}^{(i)}| \le Wr\} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x}) - y)^2]\bigg| \le 2\epsilon_1.$$

*Proof.* Fix $(\mathbf{w}_j, u_j) \in \mathcal{P}$. Since $\mathcal{D}_{\mathbf{x}}$ is sub-exponential, we have $\mathbf{Pr}[|\mathbf{w}_j\cdot\mathbf{x}| \ge \|\mathbf{w}_j\|_2 r] \le \frac{1}{L^2}\exp(-Lr)$. Consider random variables $Z_{i,j} = (u_j(\mathbf{w}_j\cdot\mathbf{x}^{(i)}) - y^{(i)})^2\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}^{(i)}| \le r\}$,

$i = 1, \cdots, m$, $j = 1, \cdots, t_0 J$, where $(\mathbf{x}^{(i)}, y^{(i)})$ are independent random variables drawn from $\mathcal{D}$. Using Fact D.3.3 (Section D.3), we can truncate the labels $y$ such that $|y| \leq M$, where $M = C(bW/L)\log(bW/\epsilon_1)$ for some large absolute constant $C$. Hence, $|Z_{i,j}| \leq 2(u_j^2(\mathbf{w}_j \cdot \mathbf{x}^{(i)}) + (y^{(i)})^2)\mathbb{1}\{|\mathbf{w}_j \cdot \mathbf{x}^{(i)}| \leq Wr\} \leq 2(b^2W^2r^2 + M^2)$, where we used the assumption that $u$ is $b$-Lipschitz in the last inequality. Therefore, applying Hoeffding's inequality to $Z_{i,j}$ we get:

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{m'}(Z_{i,j} - E[Z_{i,j}])\right| \geq m't\right] \leq 2\exp\left(-\frac{m't^2}{8(b^2W^2r^2 + M^2)^2}\right).$$

Since there are $t_0 J = Wb^7/(\mu^7\eta\sqrt{\epsilon_1}) = 4b^9W/(\mu^8\sqrt{\epsilon_1})$ elements in the set $\mathcal{P}$, applying a union bound leads to:

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{m'} Z_{i,j} - E[Z_{i,j}]\right| \geq m't, \forall j \in [J]\right] \leq 2\exp\left(-\frac{m't^2}{8(b^2W^2r^2 + M^2)^2} + \log(4b^9W/(\mu^8\sqrt{\epsilon_1}))\right).$$

Therefore, when

$$m' \geq \frac{8(b^2W^2r^2 + M^2)^2}{\epsilon_1^2}\left(\log\left(\frac{4b^9W}{\mu^8\sqrt{\epsilon_1}}\right) + \log(2/\delta)\right), \tag{6.47}$$

we have that with probability at least $1 - \delta$:

$$\left|\frac{1}{m'}\sum_{i=1}^{m'}(u_j(\mathbf{w}_j \cdot \mathbf{x}^{(i)}) - y^{(i)})^2\mathbb{1}\{|\mathbf{w}_j \cdot \mathbf{x}^{(i)}| \leq Wr\} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j \cdot \mathbf{x}) - y)^2\mathbb{1}\{|\mathbf{w}_j \cdot \mathbf{x}| \leq Wr\}]\right| \leq \epsilon_1, \tag{6.48}$$

for any $(\mathbf{w}_j, u_j) \in \mathcal{P}$. In addition, as $\mathbf{Pr}[|\mathbf{w}_j \cdot \mathbf{x}| \geq Wr] \leq \mathbf{Pr}[|\mathbf{w}_j \cdot \mathbf{x}| \geq \|\mathbf{w}_j\|_2 r] \leq$

$\frac{2}{L^2}\exp(-Lr)$, letting $\epsilon' = \frac{2}{L^2}\exp(-Lr)$, we further have:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\geq Wr\}]\\
&\leq 2\,\mathbf{E}[((u_j(\mathbf{w}_j\cdot\mathbf{x}))^2+M^2)\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\geq Wr\}]\\
&\leq 2b^2\sqrt{\mathbf{E}[(\mathbf{w}_j\cdot\mathbf{x})^4]\,\mathbf{Pr}[|\mathbf{w}_j\cdot\mathbf{x}|\geq Wr]}+M^2\,\mathbf{Pr}[|\mathbf{w}_j\cdot\mathbf{x}|\geq Wr]\\
&\leq 2cb^2(W/L)^2\sqrt{\epsilon'}+M^2\epsilon' \leq (2cb^2(W/L)^2+M^2)\sqrt{\epsilon'},
\end{aligned}$$

where in the second inequality we used Cauchy-Schwarz inequality and in the last inequality we used the property that for any unit vector $\mathbf{a}$ it holds $\mathbf{E}[(\mathbf{a}\cdot\mathbf{x})^4]\leq c^2/L^4$ for some absolute constant $c$ as $\mathbf{x}$ possesses a $\frac{1}{L}$-sub-exponential tail. Therefore, choosing $r=\frac{1}{L}\log(\frac{C^2b^4W^4}{L^6\epsilon_1^2}\log^2(\frac{bW}{\epsilon_1}))=\tilde{O}(\frac{1}{L}\log(\frac{bW}{L\epsilon_1}))$ for some large absolute constant $C$ renders $\sqrt{\epsilon'}\leq\epsilon_1/(2Cb^2(W/L)^2\log^2(bW/\epsilon_1))$, and we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\geq Wr\}]\leq\epsilon_1.$$

Observe that as $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2]$ is the sum of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\geq Wr\}]$ and $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\leq Wr\}]$, we have

$$\begin{aligned}
0&\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2]-\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\leq Wr\}]\\
&\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\geq Wr\}]\leq\epsilon_1.
\end{aligned}$$

Plugging the choice of $r$ back into (6.47), we get that it is sufficient to choose $m'$ as

$$\begin{aligned} m' &= \frac{C\log(\log(1/\epsilon_1))}{\epsilon_1^2}\left(b^2\left(\frac{W}{L}\right)^2\log^2\left(\frac{bW}{L\epsilon_1^2}\right)\right)^2\left(\log\left(\frac{4b^9W}{\mu^8\sqrt{\epsilon_1}}\right)+\log(1/\delta)\right) \\ &= \tilde{\Theta}\left(\frac{b^4W^4\log(1/\delta)}{L^4\epsilon_1^2}\log^5\left(\frac{bW}{L\mu\epsilon_1}\right)\right). \end{aligned}$$

Therefore, using $m' = \tilde{\Omega}(b^4W^4/(L^4\epsilon_1^2))$ samples, (6.48) indicates that with probability at least $1-\delta$, for any $(\mathbf{w}_j, u_j) \in \mathcal{P}$ it holds

$$\begin{aligned} &|\frac{1}{m'}\sum_{i=1}^{m'}(u_j(\mathbf{w}_j\cdot\mathbf{x}^{(i)})-y^{(i)})^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}^{(i)}|\le Wr\} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2]| \le \\ &|\frac{1}{m'}\sum_{i=1}^{m'}(u_j(\mathbf{w}_j\cdot\mathbf{x}^{(i)})-y^{(i)})^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}^{(i)}|\le Wr\} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\le Wr\}]| \\ &+|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_j(\mathbf{w}_j\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}_j\cdot\mathbf{x}|\le Wr\}]| \\ &\le 2\epsilon_1, \end{aligned}$$

thus completing the proof of Claim 6.3.8. □

Therefore, Claim 6.3.8 guarantees that selecting a hypothesis using the provided testing procedure introduces an error at most $2\epsilon_1$, with high probability.

# Chapter 7

# Robustly Learning Monotone GLMs under Gaussian Distribution

## 7.1 Introduction

This chapter provides a definitive positive answer to a core question in robust learning: whether all monotone GLMs can be learned efficiently under advcersarial label noise. We develop the first efficient, constant-factor approximate learner for GLMs with *any* known monotone activation function as long as the marginals are distributed according to a standard normal. Our approach bypasses the limitations of prior work, which required specific structural properties on the activation such as being sigmoidal (Definition 5.1.2) or having a strictly positive derivative (Definition 5.1.5). We achieve this by introducing a novel framework based on data augmentation, where we prove that injecting Gaussian noise into the data provably improves the non-convex loss landscape.

Our primary result (Theorem 7.1.1) is an algorithm that, for any monotone and $b$-Lipschitz activation $\sigma$, returns a hypothesis with error at most $C \cdot \mathrm{OPT} + \epsilon$ in polynomial

time. The approximation constant $C > 0$ is universal and, notably, independent of the problem parameters like the dimension $d$ and the Lipschitz constant $b$. Crucially, this framework extends beyond the Lipschitz case to encompass nearly all monotone activations in $L_2(\mathcal{N})$, where the sample complexity gracefully increases depending on the function's tail behavior (TODO:Add general-thm). This generality is near-optimal, as the tail behavior is required for the statistical error, i.e., it is information-theoretically necessary even for realizable learning (Theorem 7.5.13).

### Main Results

**Theorem 7.1.1** (Robustly Learning Monotone & Lipschitz GLMs)**.** *There exists an algorithm with the following performance guarantee: For any known monotone and $b$-Lipschitz activation $\sigma$, given $\epsilon > 0$, $W > 0$, and $N = \tilde{\Theta}(d(bW)^2/\epsilon + d/\epsilon^2)$ samples, the algorithm runs in* $\mathrm{poly}(d, N)$ *time and returns a vector $\widehat{\mathbf{w}}$ such that with high probability, $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}) - y)^2] \leq C\mathrm{OPT} + \epsilon$, where $C$ is an absolute constant independent of $\epsilon, d, b, W$.*

The above theorem is a special case of our main algorithmic result, which applies more generally, encompassing all monotone activations with bounded $(2 + \zeta)$ moment, for any $\zeta > 0$ (Corollary 7.4.4).

**Theorem 7.1.2** (Learning Monotone Activations With Bounded $2 + \zeta$ Moments)**.** *Let $\epsilon > 0$, and let $\sigma$ be a monotone activation that satisfies $\mathbf{E}_{z\sim\mathcal{N}}[\sigma^{2+\zeta}(z)] \leq B_\sigma$. Then, Algorithm 14 draws $N = \tilde{\Theta}(d(B_\sigma/\epsilon)^{2/\zeta} \log(B_\sigma/\epsilon)/\epsilon + d/\epsilon^2)$ samples, runs in* $\mathrm{poly}(d, N)$ *time, and returns a vector $\widehat{\mathbf{w}}$ such that with probability at least $2/3$, it holds that $\mathcal{L}(\widehat{\mathbf{w}}) \leq C\mathrm{OPT} + \epsilon$, where $C$ is an absolute constant independent of $\epsilon, d, B_\sigma, L$.*

In particular, we show that all monotone activations in $L_2(\mathcal{N})$ are $\epsilon$-approximable by an activation with bounded $(2+\zeta)$ moments (Lemma 8.2.2), and thus, our algorithm can be applied to learn GLMs with *any* monotone activation in $L^2$.

**Remark** In the sequel, we assume that the scale of the target vector $\mathbf{w}^*$, $\|\mathbf{w}^*\|_2$, is known and, by, rescaling the space, we optimize $\mathbf{w}$ on the unit sphere. The unknown scale of $\mathbf{w}^*$ can be resolved by a simple grid search. For our approach, this rescaling is w.l.o.g. because—unlike in prior work [WZDD23, WZDD24, ZWDD24]—the approximation ratio of our algorithm is **independent** of any problem parameters. For a formal justification, see Remark 7.5.3 and Lemma 7.5.4.

**Organization** In Section 7.2, we analyze the landscape of the augmented loss. Our main algorithm and its analysis for learning Gaussian GLMs of general activations is presented in Section 7.3. In Section 7.4, we focus on monotone activations and show that our algorithm achieves error $O(\mathrm{OPT}) + \epsilon$ under very mild assumptions. Due to space limitations, several proofs have been deferred to the Appendix.

## Technical Preliminaries

We collect some tools for this chapter. In the following fact, we gather some useful facts about Hermite polynomials that are used throughout the section.

**Fact 7.1.3** (See, e.g., [Bog98])**.** *The following statements hold for Hermite polynomials as defined above.*

1. *(Parseval's identity) For any $f \in L_2(\mathcal{N})$, we have $\mathbf{E}_{z\sim\mathcal{N}(0,1)}[(f(z) - \mathrm{P}_k f(z))^2] = \sum_{i=k+1}^{\infty} \hat{f}(i)^2$.*

*2. (Mehler's Identity) For any real number $|\rho| < 1$ and $x, y \in \mathbb{R}$, it holds*

$$\sum_{k \geq 0} \rho^k \mathrm{He}_k(x)\mathrm{He}_k(y) = \frac{1}{\sqrt{1-\rho^2}} \exp\left( -\frac{(\rho x - y)^2}{2(1-\rho^2)} + \frac{y^2}{2} \right). \tag{7.1}$$

*3. (Differentiation)* $(\mathrm{He}_i(z))' = \sqrt{i}\mathrm{He}_{i-1}(z)$.

Finally, the following facts about Gaussian distribution are useful to our work:

**Fact 7.1.4** (Stein's Lemma [Ste81])**.** *Suppose that* $\mathbf{x}$ *is distributed as* $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ *for some* $\boldsymbol{\mu} \in \mathbb{R}^d, \sigma \in \mathbb{R}_+$ *and let* $g : \mathbb{R}^d \to \mathbb{R}$ *be an almost everywhere differentiable function such that both* $\mathbf{E}[g(\mathbf{x})\mathbf{x}]$ *and* $\mathbf{E}[\nabla g(\mathbf{x})]$ *exist. Then, it holds*

$$\mathbf{E}[g(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu})] = \sigma^2 \, \mathbf{E}[\nabla g(\mathbf{x})] \, .$$

**Fact 7.1.5** (Komatsu's Inequality)**.** *For any* $t \geq 0$ *it holds:*

$$C\frac{\exp(-t^2/2)}{t + \sqrt{t^2+4}} < \mathop{\mathbf{Pr}}_{x \sim \mathcal{N}(0,1)}[x \geq t] < C\frac{\exp(-t^2/2)}{t + \sqrt{t^2+2}} \, ,$$

*where* $C > 0$ *is a universal constant.*

An important tool for our work is the Ornstein–Uhlenbeck semigroup. The Ornstein–Uhlenbeck semigroup and operators are broadly used in stochastic analysis and control theory (see, e.g., [Bog98]). Within learning theory, they have found applications in bounding the sensitivity of a Boolean function [KOS08]. A formal definition of the Ornstein–Uhlenbeck semigroup is provided below.

**Definition 7.1.6** (Ornstein–Uhlenbeck Semigroup)**.** *Let* $\rho \in (0, 1)$*. The Ornstein–Uhlenbeck semigroup, denoted by* $\mathrm{T}_\rho$*, is a linear operator that maps a function* $g \in L_2(\mathcal{N})$

*to the function* $\mathrm{T}_\rho g$ *defined as:* $(\mathrm{T}_\rho g)(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[g(\rho\mathbf{x} + \sqrt{1-\rho^2}\mathbf{z})]$.

The following fact summarizes useful properties of the Ornstein–Uhlenbeck semigroup.

**Fact 7.1.7** (see, e.g., [Bog98], [O'D14](Chapter 11))**.** *Let* $f, g \in L_2(\mathcal{N})$.

1. *For any* $f, g \in L_2$ *and any* $t > 0$, $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{T}_t f(\mathbf{x}))g(\mathbf{x})] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{T}_t g(\mathbf{x}))f(\mathbf{x})]$ .
2. *For any* $g : \mathbb{R}^d \to \mathbb{R}$, $g \in L_2$, *all of the following statements hold.*
   (a) *For any* $t, s > 0$, $\mathrm{T}_t \mathrm{T}_s g = \mathrm{T}_{ts} g$.
   (b) *For any* $\rho \in (0,1)$, $\mathrm{T}_\rho g(\mathbf{x})$ *is differentiable at every point* $\mathbf{x} \in \mathbb{R}^d$.
   (c) *For any* $\rho \in (0,1)$, $\mathrm{T}_\rho g(\mathbf{x})$ *is* $\|g\|_{L_\infty}/(1-\rho^2)^{1/2}$*-Lipschitz, i.e.,* $\|\nabla \mathrm{T}_\rho g(\mathbf{x})\|_{L_\infty} \le \|g\|_{L_\infty}/(1-\rho^2)^{1/2}$, $\forall \mathbf{x} \in \mathbb{R}^d$.
   (d) *For any* $\rho \in (0,1)$, $\mathrm{T}_\rho g(\mathbf{x}) \in \mathcal{C}^\infty$.
   (e) *For any* $p \ge 1$, $\mathrm{T}_\rho$ *is nonexpansive with respect to the norm* $\|\cdot\|_{L_p}$*, i.e.,* $\|\mathrm{T}_\rho g\|_{L_p} \le \|g\|_{L_p}$.
   (f) $\|\mathrm{T}_\rho g(\mathbf{x})\|_{L_2}$ *is non-decreasing w.r.t.* $\rho$.
   (g) *If* $g$ *is, in addition, a differentiable function, then for all* $\rho \in (0,1)$*, it holds that:* $\nabla_{\mathbf{x}} \mathrm{T}_\rho g(\mathbf{x}) = \rho \mathrm{T}_\rho \nabla_{\mathbf{x}} g(\mathbf{x})$*, for any* $\mathbf{x} \in \mathbb{R}^d$.
3. *For all* $\rho \in (0,1)$ *and* $i \in \mathbb{Z}_+$, $\mathrm{T}_\rho \mathrm{He}_i(z) = \rho^i \mathrm{He}_i(z)$.

The Ornstein–Uhlenbeck semigroup induces an operator L applying to functions $f \in L_2(\mathcal{N})$, defined below.

**Definition 7.1.8** (Definition 11.24 in [O'D14]). *The Ornstein–Uhlenbeck operator is a linear operator applied that applies to functions $f \in L_2(\mathcal{N})$ and is defined by $\mathrm{L}f = \frac{\mathrm{dT}_\rho f}{\mathrm{d}\rho} \mid_{\rho=1}$, provided that $\mathrm{L}f$ exists.*

**Fact 7.1.9.** *Let $f, g \in L_2(\mathcal{N})$, $\rho \in (0,1)$. Then:*

1. *([O'D14, Proposition 11.27]) $\frac{\mathrm{dT}_\rho f}{\mathrm{d}\rho} = \frac{1}{\rho}\mathrm{LT}_\rho f = \frac{1}{\rho}\mathrm{T}_\rho \mathrm{L}f$.*
2. *([O'D14, Proposition 11.28]) $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\left[f(\mathbf{x})\mathrm{LT}_\rho g(\mathbf{x})\right] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\left[\nabla f(\mathbf{x})\nabla \mathrm{T}_\rho g(\mathbf{x})\right]$ .*

We use Fact 7.1.9 to prove the following Lemma 7.1.10:

**Lemma 7.1.10.** *Let $f \in L_2(\mathcal{N})$ be a continuous and (almost everywhere) differentiable function. Then $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{T}_\rho f(\mathbf{x}) - f(\mathbf{x}))^2] \leq 3(1-\rho)\,\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2]$.*

*Proof.* Observe that $(\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{T}_\rho f(\mathbf{x}) - f(\mathbf{x}))^2])^{1/2} = \sup_{g\in\mathcal{C}^\infty, \|g\|_{L_2}\leq 1} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[g(\mathbf{x})(\mathrm{T}_\rho f(\mathbf{x}) - f(\mathbf{x}))]$. Consider any $g \in \mathcal{C}^\infty$ with $\|g\|_{L_2} \leq 1$. We have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[g(\mathbf{x})(\mathrm{T}_\rho f(\mathbf{x}) - f(\mathbf{x}))] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})(\mathrm{T}_\rho g(\mathbf{x}) - g(\mathbf{x}))] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[f(\mathbf{x})\int_\rho^1 \frac{\mathrm{dT}_t g(\mathbf{x})}{\mathrm{d}t}\mathrm{d}t\right].$$

As $g \in \mathcal{C}^\infty$, we can use Fact 7.1.9 to conclude that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[f(\mathbf{x})\int_\rho^1 \frac{\mathrm{dT}_t g(\mathbf{x})}{\mathrm{d}t}\mathrm{d}t\right] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[f(\mathbf{x})\int_\rho^1 (1/t)\mathrm{LT}_t g(\mathbf{x})\mathrm{d}t\right],$$

where L is the Ornstein–Uhlenbeck operator. Using the identity that for $f$ such that $\nabla f \in L_2(\mathcal{N})$ and $g \in \mathcal{C}^\infty$ it holds that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\left[f(\mathbf{x})\mathrm{LT}_t g(\mathbf{x})\right] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\left[\nabla f(\mathbf{x})\nabla \mathrm{T}_t g(\mathbf{x})\right]$ (Fact 7.1.9), we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[g(\mathbf{x})(\mathrm{T}_\rho f(\mathbf{x}) - f(\mathbf{x}))] = \int_\rho^1 (1/t) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\nabla f(\mathbf{x})\nabla \mathrm{T}_t g(\mathbf{x})]\,\mathrm{d}t.$$

Note that using Fact 7.1.7 (f) and Stein's lemma Fact 7.1.4, we have:

$$\begin{aligned}\nabla \mathrm{T}_\rho g(\mathbf{x}) &= \rho \mathrm{T}_\rho \nabla g(\mathbf{x}) = \rho \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\nabla g(\rho\mathbf{x} + \sqrt{1-\rho^2}\mathbf{z})] \\ &= (\rho/(\sqrt{1-\rho^2})) \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}[g(\rho\mathbf{x} + \sqrt{1-\rho^2}\mathbf{z})\mathbf{z}].\end{aligned}$$

Therefore, since $\mathbf{z}, \mathbf{x}$ are independent standard Gaussian random vectors, we have that

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[g(\mathbf{x})(\mathrm{T}_\rho f(\mathbf{x}) - f(\mathbf{x}))] \\ &= \int_\rho^1 \frac{1}{\sqrt{1-t^2}} \mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}}\left[g(t\mathbf{x} + \sqrt{1-t^2}\mathbf{z})\mathbf{z}\cdot\nabla f(\mathbf{x})\right] \mathrm{d}t \\ &\le \int_\rho^1 \frac{1}{\sqrt{1-t^2}} \left(\mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}}\left[g(t\mathbf{x} + \sqrt{1-t^2}\mathbf{z})^2\right] \mathop{\mathbf{E}}_{\mathbf{x},\mathbf{z}\sim\mathcal{N}}\left[(\mathbf{z}\cdot\nabla f(\mathbf{x}))^2\right]\right)^{1/2} \mathrm{d}t \\ &= \int_\rho^1 \frac{1}{\sqrt{1-t^2}} \left(\mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{N}}\left[g(\mathbf{u})^2\right] \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\|\nabla f(\mathbf{x})\|_2^2\right]\right)^{1/2} \mathrm{d}t \\ &\le \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\|\nabla f(\mathbf{x})\|_2^2\right]\right)^{1/2} \int_\rho^1 \frac{1}{\sqrt{1-t^2}} \mathrm{d}t = \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\|\nabla f(\mathbf{x})\|_2^2\right]\right)^{1/2} \arccos\rho\,,\end{aligned}$$

where we used the Cauchy-Schwarz inequality and the fact that $\|g\|_{L_2} \le 1$. Using the inequality $\arccos\rho \le \sqrt{3}\sqrt{1-\rho}$, we complete the proof of Lemma 7.1.10. □

## 7.2 Data Augmentation and Its Effect on the $L_2^2$ Loss Landscape

### 7.2.1 Augmenting the Data: Connection to Ornstein–Uhlenbeck Semigroup

Our algorithm relies on the data augmentation technique, i.e., in each iteration, the algorithm injects Gaussian noise (see Algorithm 13), which has the effect of improving the regularity properties of the loss landscape, as shown in this section.

**Algorithm 13** Augment Dataset with Injected White Noise

**Input:** Parameters $\rho$, $m$; sample data $\mathfrak{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$
$S \leftarrow \emptyset$
**for** each $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathfrak{D}$ **do**
    **for** $j = 1$ **to** $m$ **do**
        Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
        Let $\tilde{\mathbf{x}}^{(j)} \leftarrow \rho\, \mathbf{x}^{(i)} + \sqrt{1-\rho^2}\, \mathbf{z}$
        $S \leftarrow S \cup \{(\tilde{\mathbf{x}}^{(j)}, y^{(i)})\}$
    **end for**
**end for**
**Return:** $S$

**Remark 7.2.1.** Note that Algorithm 13 does not require new samples from the distribution $\mathcal{D}$. As we will see in Lemma 7.2.4, the purpose of Algorithm 13 is to estimate $\mathrm{T}_\rho \sigma'(\mathbf{w} \cdot \mathbf{x}^{(i)})$ for each sample $\mathbf{x}^{(i)}$ we received. By standard concentration bounds using at most $\tilde{O}(L/\epsilon^2)$ independent (unlabeled) Gaussian samples suffices for all $\{\mathbf{x}^{(i)}, y^{(i)}\} \in \mathfrak{D}$. Since this affects only the runtime in polynomially, for simplicity of analysis we assume Algorithm 13 can be executed efficiently and have access to the population one.

The augmentation can be viewed as a transformation of the distribution $\mathcal{D}$ to $\mathcal{D}_\rho$, where for any $(\tilde{\mathbf{x}}, y) \sim \mathcal{D}_\rho$, it holds $\tilde{\mathbf{x}} \sim \rho\mathcal{D}_{\mathbf{x}} + (1-\rho^2)^{1/2}\mathcal{N}(\mathbf{0}, \mathbf{I})$. The data augmentation introduced in Algorithm 13 in fact simulates the Ornstein–Uhlenbeck semigroup, which we formalize in the following lemma.

**Lemma 7.2.2.** *Let $\mathcal{D}$ be a distribution of labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such with $\mathbf{x}$-marginal $\mathcal{D}_{\mathbf{x}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Furthermore, let $\mathcal{D}_\rho$ be the distribution constructed by applying Algorithm 13 to samples from $\mathcal{D}$. Then for any $f : \mathbb{R} \to \mathbb{R}$ and any unit vector $\mathbf{w} \in \mathbb{R}^d$ with $|\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[f(\mathbf{w}\cdot\mathbf{x})]| < \infty$, we have $\mathbf{E}_{\tilde{\mathbf{x}}\sim(\mathcal{D}_\rho)_{\tilde{\mathbf{x}}}}[f(\mathbf{w}\cdot\tilde{\mathbf{x}})] = \mathbf{E}[\mathrm{T}_\rho f(\mathbf{w}\cdot\mathbf{x})]$ .*

*Proof.* Using the definition of $\mathcal{D}_\rho$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\tilde{\mathbf{x}}\sim(\mathcal{D}_\rho)_{\tilde{\mathbf{x}}}}[f(\mathbf{w}\cdot\tilde{\mathbf{x}})] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\rho\mathbf{w}\cdot\mathbf{x} + \sqrt{1-\rho^2}\mathbf{w}\cdot\mathbf{z})]] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathop{\mathbf{E}}_{\zeta\sim\mathcal{N}(0,1)}[f(\rho\mathbf{w}\cdot\mathbf{x} + \sqrt{1-\rho^2}\zeta)]] = \mathbf{E}[\mathrm{T}_\rho f(\mathbf{w}\cdot\mathbf{x})] \ ,\end{aligned}$$

where we have used that $\mathbf{w}\cdot\mathbf{z}$ is distributed according to the standard normal distribution.

□

Lemma 7.2.2 shows that our data augmentation technique is equivalent to applying the Ornstein–Uhlenbeck semigroup to our dataset. This application of the Ornstein–Uhlenbeck semigroup to the dataset has the effect of smoothing the landscape of the square loss, which in turn allows us to prove that the gradient of the smoothed/augmented loss carries information about the direction of the target vector. This is the main structural result obtained in the next subsection.

### 7.2.2 Alignment of the Gradients of the Augmented Loss

In this section, we provide the main structural result of this work, showing that the gradients of the square loss applied to the augmented data correlate with a target parameter vector $\mathbf{w}^*$. For notational convenience, we use

$$\mathcal{L}_\rho(\mathbf{w}) = \mathop{\mathbf{E}}_{(\tilde{\mathbf{x}},y)\sim\mathcal{D}_\rho}[(\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}}) - y)^2] \tag{7.2}$$

to denote the square loss on the augmented data and refer to it as the "augmented loss."

**Proposition 7.2.3** (Main Structural Result)**.** *Fix an activation $\sigma : \mathbb{R} \to \mathbb{R}$. Let $\mathcal{D}$ be a distribution of labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such that its $\mathbf{x}$-marginal $\mathcal{D}_\mathbf{x}$ is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Moreover, let $\mathcal{D}_\rho$ be the distribution constructed by applying Algorithm 13 with parameter $\rho \in (0, 1)$ to the distribution $\mathcal{D}$. Fix unit vectors $\mathbf{w}^*, \mathbf{w} \in \mathbb{R}^d$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2] = \mathrm{OPT}$ and let $\theta = \theta(\mathbf{w}^*, \mathbf{w})$. Let $\mathbf{g}(\mathbf{w})$ be the gradient of the loss $\mathcal{L}_\rho(\mathbf{w}) = \mathbf{E}_{(\tilde{\mathbf{x}},y)\sim\mathcal{D}_\rho}[(\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}}) - y)^2]$ projected on the subspace $\mathbf{w}^\perp$ and scaled by $1/(2\rho)$, i.e., $\mathbf{g}(\mathbf{w}) = (1/(2\rho))(\nabla_\mathbf{w}\mathcal{L}_\rho(\mathbf{w}))^{\perp\mathbf{w}}$. Then,*

$$\mathbf{g}(\mathbf{w})\cdot\mathbf{w}^* \le -\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin^2\theta + \sqrt{\mathrm{OPT}}\|\mathrm{T}_\rho\sigma'\|_{L_2}\sin\theta.$$

*In particular, if $0 < \rho \le \cos\theta < 1$ and $\sin\theta \ge 3\sqrt{\mathrm{OPT}}/\|\mathrm{T}_\rho\sigma'\|_{L_2}$, then*

$$\mathbf{g}(\mathbf{w})\cdot\mathbf{w}^* \le -(2/3)\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin^2\theta\ .$$

Before proving the proposition, we first prove the following auxiliary lemma, which establishes a connection between the Riemannian gradient of Equation (7.2) and the Ornstein–Uhlenbeck semigroup applied to the derivative of the activation.

**Lemma 7.2.4.** *Let* $\mathbf{g}(\mathbf{w}) = (1/(2\rho))(\nabla_{\mathbf{w}}\mathcal{L}_\rho(\mathbf{w}))^{\perp\mathbf{w}}$. *Then,* $\mathbf{g}(\mathbf{w}) = -\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp\mathbf{w}}]$.

*Proof.* By definition, the projected gradient vector $\mathbf{g}(\mathbf{w})$ equals

$$\mathbf{g}(\mathbf{w}) = \frac{1}{\rho}\left(\mathop{\mathbf{E}}_{(\tilde{\mathbf{x}},y)\sim\mathcal{D}_\rho}[\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})\sigma'(\mathbf{w}\cdot\tilde{\mathbf{x}})\tilde{\mathbf{x}}^{\perp\mathbf{w}}] - \mathop{\mathbf{E}}_{(\tilde{\mathbf{x}},y)\sim\mathcal{D}_\rho}[y\sigma'(\mathbf{w}\cdot\tilde{\mathbf{x}})\tilde{\mathbf{x}}^{\perp\mathbf{w}}]\right).$$

Note that since $(\mathcal{D}_\rho)_{\tilde{\mathbf{x}}}$ is also a standard Gaussian distribution, we have $\mathbf{E}_{\tilde{\mathbf{x}}\sim(\mathcal{D}_\rho)_{\tilde{\mathbf{x}}}}[\tilde{\mathbf{x}}^{\perp\mathbf{w}}] = \mathbf{0}$, hence $\mathbf{E}_{(\tilde{\mathbf{x}},y)\sim\mathcal{D}_\rho}[\sigma(\mathbf{w}\cdot\tilde{\mathbf{x}})\sigma'(\mathbf{w}\cdot\tilde{\mathbf{x}})\tilde{\mathbf{x}}^{\perp\mathbf{w}}] = \mathbf{0}$, as $\mathbf{w}\cdot\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}^{\perp\mathbf{w}}$ are independent. Thus, $\rho\mathbf{g}(\mathbf{w}) = -\mathbf{E}_{(\tilde{\mathbf{x}},y)\sim\mathcal{D}_\rho}[y\sigma'(\mathbf{w}\cdot\tilde{\mathbf{x}})\tilde{\mathbf{x}}^{\perp\mathbf{w}}]$. Now, since $(\mathcal{D}_\rho)_{\tilde{\mathbf{x}}} = \rho\mathcal{D}_{\mathbf{x}} + (1-\rho^2)^{1/2}\mathcal{N}(\mathbf{0},\mathbf{I})$, we have

$$\mathbf{g}(\mathbf{w}) = -\frac{1}{\rho}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[y\sigma'(\mathbf{w}\cdot(\rho\mathbf{x}+\sqrt{1-\rho^2}\mathbf{z}))(\rho\mathbf{x}+\sqrt{1-\rho^2}\mathbf{z})^{\perp\mathbf{w}}].$$

As $\mathbf{z}$ is independent of $\mathbf{x}, y$ and follows the standard Gaussian distribution, it must be $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[y\sigma'(\mathbf{w}\cdot\tilde{\mathbf{x}})\mathbf{z}^{\perp\mathbf{w}}] = 0$, and thus we further have

$$\begin{aligned}
\mathbf{g}(\mathbf{w}) &= -\frac{1}{\rho}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[y\sigma'(\mathbf{w}\cdot(\rho\mathbf{x}+\sqrt{1-\rho^2}\mathbf{z}))\rho\mathbf{x}^{\perp\mathbf{w}}] \\
&= -\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\sigma'(\mathbf{w}\cdot(\rho\mathbf{x}+\sqrt{1-\rho^2}\mathbf{z}))]\mathbf{x}^{\perp\mathbf{w}}] \\
&= -\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp\mathbf{w}}],
\end{aligned}$$

completing the proof. □

We are now ready to prove our main structural result.

*Proof of Proposition 7.2.3.* When $\mathbf{w}^*$ is parallel to $\mathbf{w}$, then $\mathbf{g}(\mathbf{w})\cdot\mathbf{w}^* = 0$ since $\mathbf{g}(\mathbf{w})$ is orthogonal to $\mathbf{w}$, and $\sin\theta = 0$, hence the statements hold trivially. Thus in the

rest of the proof we assume that $(\mathbf{w}^*)^{\perp_\mathbf{w}} \neq \mathbf{0}$. Denote $\mathbf{v} \stackrel{\text{def}}{=} (\mathbf{w}^*)^{\perp_\mathbf{w}}/\|(\mathbf{w}^*)^{\perp_\mathbf{w}}\|_2$. Then, $\mathbf{w}^* = \mathbf{w}\cos\theta + \mathbf{v}\sin\theta$, where $\theta \stackrel{\text{def}}{=} \theta(\mathbf{w}, \mathbf{w}^*)$. Using Lemma 7.2.4, the inner product between $\mathbf{w}^*$ and $-\mathbf{g}(\mathbf{w})$ equals

$$-\mathbf{g}(\mathbf{w}) \cdot \mathbf{w}^* = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{v}\cdot\mathbf{x}]\sin\theta\ .$$

By adding and subtracting $\sigma(\mathbf{w}^*\cdot\mathbf{x})$ on the right-hand side, we get that

$$\begin{aligned}-\mathbf{g}(\mathbf{w}) \cdot \mathbf{w}^* = \mathbf{E}[\sigma(\mathbf{w}^*\cdot\mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{v}\cdot\mathbf{x}]\sin\theta \\ + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-\sigma(\mathbf{w}^*\cdot\mathbf{x}))\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{v}\cdot\mathbf{x}]\sin\theta. \qquad (7.3)\end{aligned}$$

Observe that since $\mathbf{w}^* = \cos\theta\mathbf{w} + \sin\theta\mathbf{v}$ and $\mathbf{x}$ is a standard Gaussian random vector, we have $\mathbf{w}\cdot\mathbf{x}$ and $\mathbf{v}\cdot\mathbf{x}$ are independent standard Gaussian random variables. By applying Cauchy-Schwarz inequality to the expectation in the last term, we obtain

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-\sigma(\mathbf{w}^*\cdot\mathbf{x}))\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{v}\cdot\mathbf{x}] \\ &\geq -\left(\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2]\,\mathbf{E}[(\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{v}\cdot\mathbf{x})^2]\right)^{1/2} \\ &= -\sqrt{\mathrm{OPT}\,\mathbf{E}[(\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x}))^2]} = -\sqrt{\mathrm{OPT}}\|\mathrm{T}_\rho\sigma'\|_{L_2},\end{aligned}$$

where in the first equality we used the definition of OPT and that $\mathbf{w}\cdot\mathbf{x}$ and $\mathbf{v}\cdot\mathbf{x}$ are independent standard Gaussian random variables noted above.

To bound the first term on the right-hand side of (7.3), we again use that $\mathbf{w}\cdot\mathbf{x}$ and $\mathbf{v}\cdot\mathbf{x}$ are independent standard Gaussian random variables and apply Stein's lemma

(Fact 7.1.4) to obtain

$$\begin{aligned}\mathbf{E}[\sigma(\mathbf{w}^* \cdot \mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x})\mathbf{v} \cdot \mathbf{x}] &= \mathbf{E}[\sigma(\cos\theta\mathbf{w} \cdot \mathbf{x} + \sin\theta\mathbf{v} \cdot \mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x})\mathbf{v} \cdot \mathbf{x}] \\ &= \mathbf{E}[\sigma'(\cos\theta\mathbf{w} \cdot \mathbf{x} + \sin\theta\mathbf{v} \cdot \mathbf{x})\sin\theta\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x})] \\ &= \mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w} \cdot \mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x})]\sin\theta = \|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin\theta \ ,\end{aligned}$$

where in the last equality we used the identity $\mathrm{T}_a\mathrm{T}_b = \mathrm{T}_{ab} = \mathrm{T}_{\sqrt{ab}}\mathrm{T}_{\sqrt{ab}}$. Therefore, we have that

$$-\mathbf{g}(\mathbf{w}) \cdot \mathbf{w}^* \geq \|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin^2\theta - \sqrt{\mathrm{OPT}}\|\mathrm{T}_\rho\sigma'\|_{L_2}\sin\theta.$$

To argue the last part of Proposition 7.2.3, recall (by Fact 7.1.7, Part 2(e)) that the function $g(\lambda) \stackrel{\mathrm{def}}{=} \|\mathrm{T}_\lambda f\|_{L_2}$ is non-decreasing in $\lambda \in (0,1)$ for any function $f \in L_2(\mathcal{N})$, therefore $\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2} \geq \|\mathrm{T}_\rho\sigma'\|_{L_2}$ if $\cos\theta \geq \rho$. By using the assumption that $\sin\theta \geq 3\sqrt{\mathrm{OPT}}/\|\mathrm{T}_\rho\sigma'\|_{L_2}$, we obtain that

$$-\mathbf{g}(\mathbf{w}) \cdot \mathbf{w}^* \geq (2/3)\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin^2\theta \ .$$

This completes the proof of Proposition 7.2.3. □

### 7.2.3 Critical Points and Their Connection to the $L_2^2$ Loss

Proposition 7.2.3 provides sufficient conditions ensuring that the vector $-\mathbf{g}(\mathbf{w})$ directs $\mathbf{w}$ towards the direction of $\mathbf{w}^*$ whenever we are in a region around approximate solutions. Specifically, if the parameter $\rho$ is chosen appropriately and the following alignment condition holds: $\sin\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2} \geq 3\sqrt{\mathrm{OPT}}$, then $-\mathbf{g}(\mathbf{w})$ has a nontrivial correlation

with $\mathbf{w}^*$. Otherwise, we can guarantee that the angle between $\mathbf{w}$ and $\mathbf{w}^*$ is already sufficiently small. This implies that the region of convergence of an algorithm that relies on $-\mathbf{g}(\mathbf{w})$ depends on the quantity:

$$\psi_\sigma(\theta) \coloneqq \sin\theta \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}.$$

Motivated by this observation, we define the *Convergence Region*, which characterizes the region of $\theta$ (and, equivalently, the region of $\mathbf{w}$) for which the algorithm makes progress towards $\mathbf{w}^*$.

**Definition 7.2.5** (Critical Point and Convergence Region of $\sigma$)**.** *Given $\sigma : \mathbb{R} \to \mathbb{R}$, $\sigma \in L_2(\mathcal{N})$, and $\theta_0 \in [0, \pi/2]$, we define the error alignment function $\psi_\sigma : [0, \pi/2] \to \mathbb{R}_+$ with respect to $\sigma$ as follows: $\psi_\sigma(\theta) \stackrel{\text{def}}{=} \sin\theta \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$. For any $\epsilon > 0$, we define the Convergence Region $\mathcal{R}_{\sigma,\theta_0}(\epsilon) = \{\theta : \psi_\sigma(\theta) \leq \sqrt{\epsilon}\} \cap \{\theta : 0 \leq \theta \leq \theta_0\}$. We say that $\theta^*$ is $(\sigma, \theta_0, \epsilon)$-Critical Point if $\theta^*$ is the maximum $\theta$ in $\mathcal{R}_{\sigma,\theta_0}(\epsilon)$.*

Definition 7.2.5 defines the Convergence Region using an upper bound $\theta_0$. This upper bound is necessary because $\psi_\sigma(\theta)$ is not necessarily monotonic. Specifically, it can be shown that $\psi_\sigma(\theta)$ is non-decreasing up to a critical point $\theta'$ and then non-increasing (see Figure 7.1 for illustrative examples). Consequently, the region $\mathcal{R}_{\sigma,\theta_0}(\epsilon)$ may consist of two disjoint intervals. The role of (an appropriately selected) $\theta_0$ is to ensure that this does not happen.

**Claim 7.2.6.** *Let $\sigma \in L_2(\mathcal{N})$. Then there exists a real number $\bar{\theta} \in (0, \pi/2)$, such that for any $\theta \leq \bar{\theta}$, $\psi_\sigma(\theta)$ is non-decreasing. If $\|\sigma''\|_{L_2} \leq L'$, then $\bar{\theta} \geq \min(\pi/3, \|\mathrm{T}_{1/2}\sigma'\|_{L_2}^2/(L')^2)$.*

*Proof.* Since $\psi_\sigma(\theta) \geq 0$, to show that $\psi_\sigma(\theta)$ is non-decreasing is equivalent to show that

$\psi_\sigma^2(\theta)$ is non-decreasing. Let us calculate the derivative of $\psi_\sigma^2(\theta)$:

$$\begin{aligned}(\psi_\sigma^2(\theta))' &= \frac{\mathrm{d}}{\mathrm{d}\theta}(\sin^2\theta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\cos\theta}\sigma')^2]) \\ &= 2\sin\theta\cos\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 + 2\sin^2\theta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\mathrm{T}_{\cos\theta}\sigma'\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{T}_{\cos\theta}\sigma'\right]\end{aligned}$$

Using Fact 7.1.9, since $\frac{\mathrm{d}}{\mathrm{d}\rho}\mathrm{T}_\rho f = (\mathrm{LT}_\rho f)/\rho$, we further have

$$\begin{aligned}(\psi_\sigma^2(\theta))' &= 2\sin\theta\cos\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 + 2\sin^2\theta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\mathrm{T}_{\cos\theta}\sigma'\frac{1}{\cos\theta}\mathrm{LT}_{\cos\theta}\sigma'(-\sin\theta)\right] \\ &= 2\sin\theta\cos\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 - 2\frac{\sin^3\theta}{\cos\theta}\mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\frac{\mathrm{d}}{\mathrm{d}z}\mathrm{T}_{\cos\theta}\sigma'\frac{\mathrm{d}}{\mathrm{d}z}\mathrm{T}_{\cos\theta}\sigma'\right] \\ &= 2\sin\theta\cos\theta\left(\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 - \tan^2\theta\|(\mathrm{T}_{\cos\theta}\sigma')'\|_{L_2}^2\right),\end{aligned}$$

where in the second equality we used Fact 7.1.9 $\mathbf{E}_{z\sim\mathcal{N}}[f(z)\mathrm{LT}_\rho g(z)] = \mathbf{E}_{z\sim\mathcal{N}}[f'(z)(\mathrm{T}_\rho g(z))']$. Therefore, we only need to prove that $h(\theta) \overset{\text{def}}{=} \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 - \tan^2\theta\|(\mathrm{T}_{\cos\theta}\sigma')'\|_{L_2}^2 \geq 0$ in a region $(0, \bar{\theta})$. Note that $h(0) = \|\sigma'\|_{L_2}^2 > 0$. Furthermore, since $\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$ and $\|(\mathrm{T}_{\cos\theta}\sigma')'\|_{L_2}^2$ are continuous functions of $\theta$ (as we can see by Hermite expansion), we know that there exists a threshold $\bar{\theta}$ such that for any $\theta \leq \bar{\theta}$, it holds $\psi_\sigma'(z) \geq 0$. Furthermore, if $\sigma''$ is in $L_2(\mathcal{N})$ and $\|\sigma''\|_{L_2} \leq L'$, then since $(\mathrm{T}_\rho f(z))' = \rho\mathrm{T}_\rho f'(z)$ and $\mathrm{T}_\rho$ is a non-expansive operator (Fact 7.1.7), we have

$$h(\theta) = \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 - \sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma''\|_{L_2}^2 \geq \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 - \sin^2\theta(L')^2.$$

Assuming $\theta \leq \pi/3$, we have $h(\theta) \geq 0$ as long as $\theta \leq \bar{\theta} = \min(\pi/3, \|\mathrm{T}_{1/2}\sigma'\|_{L_2}^2/(L')^2)$. □

The significance of the Critical Point and the Convergence Region comes from the following proposition, which bounds the $L_2^2$ error for points within the Convergence

Region.

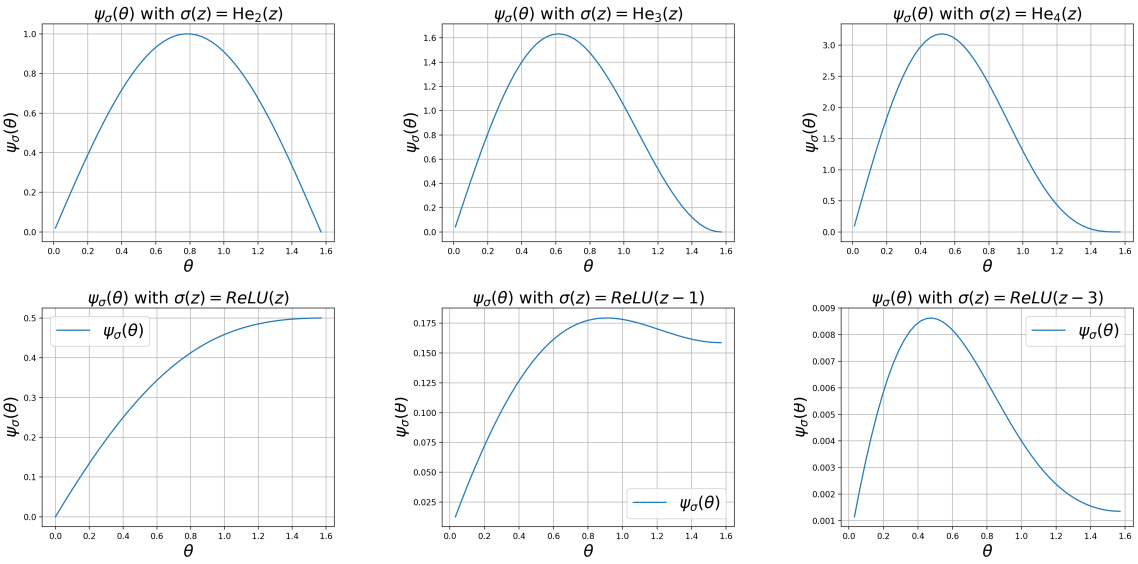



Figure 7.1: *(Up)*: The plot of $\psi_\sigma(\theta)$, where $\sigma(z) = \mathrm{He}_i(z)$, $i = 2, 3, 4$. *(Down)*: The plot of $\psi_\sigma(\theta)$, where $\sigma(z) = \mathrm{ReLU}(z - t)$, $t = 0, 1, 3$.

**Proposition 7.2.7** (Critical Points and $L_2^2$ Error)**.** *Given $\sigma : \mathbb{R} \to \mathbb{R}$, $\sigma \in L_2(\mathcal{N})$, and a distribution $\mathcal{D}$ of labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such that $\mathcal{D}_\mathbf{x} = \mathcal{N}(\mathbf{0}, \mathbf{I})$, let $\mathbf{w}^* \in \mathbb{R}^d$ be such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2] = \mathrm{OPT}$. Then, for any unit vector $\mathbf{w} \in \mathbb{R}^d$ with $\theta = \theta(\mathbf{w}, \mathbf{w}^*)$ such that $\theta \leq \theta^*$, where $\theta^*$ is the $(\sigma, \theta_0, C\mathrm{OPT})$-Critical Point for some $\theta_0$ and $C > 1$ an absolute constant, it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2] \leq O(\mathrm{OPT}) + 4\|\mathrm{P}_{>(1/\theta^*)^2}\sigma\|_{L_2}^2$.*

Proposition 7.2.7 provides a sufficient condition for proving that our algorithm converges to a region with the target approximation error. In particular, if we argue that the iterates of the algorithm we use land in the Critical Region, then Proposition 7.2.7 gives us the target $L_2^2$ error.

To prove Proposition 7.2.7, we first prove the following technical lemma, which decomposes the error into $O(\mathrm{OPT})$ and error terms that depend on the properties of the activation $\sigma$.

**Lemma 7.2.8** (Error Decomposition)**.** *Given $\sigma : \mathbb{R} \to \mathbb{R}$, $\sigma \in L_2(\mathcal{N})$, and a distribution $\mathcal{D}$ of labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such that $\mathcal{D}_{\mathbf{x}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$, let $\mathbf{w}^* \in \mathbb{R}^d$ be such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2] = \mathrm{OPT}$. Then, for any unit vector $\mathbf{w} \in \mathbb{R}^d$ with $\theta = \theta(\mathbf{w}, \mathbf{w}^*)$ and any $k \in \mathbb{Z}_+$, it holds that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2] \leq 2\mathrm{OPT} + 4\theta^2 \|\mathrm{P}_k \sigma'\|_{L_2}^2 + 4\|\mathrm{P}_{>k}\sigma\|_{L_2}^2 \,. \tag{7.4}$$

*Furthermore, if $k \geq 2$, then for any $c \in [1, (k/2)^{1/4}]$,*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2] \leq 2\mathrm{OPT} + 8e^{c^2}\theta^2 \|\mathrm{T}_{\sqrt{1-c^2/k}}\sigma'\|_{L_2}^2 + 4\|\mathrm{P}_{>k}\sigma\|_{L_2}^2 \,. \tag{7.5}$$

*Finally, if $k = 0, 1$, then for any $\rho \in (0, 1)$ it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2] \leq 2\mathrm{OPT} + \theta^2 \|\mathrm{T}_{\rho}\sigma'\|_{L_2}^2 + 4\|\mathrm{P}_{>k}\sigma\|_{L_2}^2.$$

*Proof.* First, we decompose the error into the minimum error (the one achieved by $\mathbf{w}^*$) and the alignment error (the one from the misalignment of $\mathbf{w}$ and $\mathbf{w}^*$). By Young's inequality, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2)] &\leq 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2)] + 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - \sigma(\mathbf{w} \cdot \mathbf{x}))^2)] \\ &= 2\mathrm{OPT} + 2 \underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - \sigma(\mathbf{w} \cdot \mathbf{x}))^2]}_{\text{Alignment Error}} \,. \end{aligned} \tag{7.6}$$

In the rest of the proof, we bound above the alignment error.

**Claim 7.2.9** (Angle and Alignment Error)**.** *Let* $\mathbf{w}, \mathbf{w}^* \in \mathbb{R}^d$ *be unit vectors and let* $\theta \stackrel{\text{def}}{=} \theta(\mathbf{w}, \mathbf{w}^*)$*. Then, for any* $k \in \mathbb{Z}_+$,

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - \sigma(\mathbf{w}\cdot\mathbf{x}))^2] \le 2\theta^2\|\mathrm{P}_k\sigma'\|_{L_2}^2 + 2\mathop{\mathbf{E}}_{t\sim\mathcal{N}}[(\mathrm{P}_{>k}\sigma(t))^2].$$

*Proof.* By expanding the square, since $\mathbf{w}, \mathbf{w}^*$ are fixed vectors independent of $\mathbf{x}$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - \sigma(\mathbf{w}\cdot\mathbf{x}))^2] &= 2\left(\mathop{\mathbf{E}}_{t\sim\mathcal{N}}[(\sigma(t))^2] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\sigma(\mathbf{w}^*\cdot\mathbf{x})\sigma(\mathbf{w}\cdot\mathbf{x})]\right)\\ &= 2\left(\mathop{\mathbf{E}}_{t\sim\mathcal{N}}[(\sigma(t))^2] - \mathop{\mathbf{E}}_{t\sim\mathcal{N}}[\sigma(t)\mathrm{T}_{\cos\theta}\sigma(t)]\right), \qquad (7.7)\end{aligned}$$

where in the second inequality we used that $t = \mathbf{w}^*\cdot\mathbf{x} \sim \mathcal{N}(0,1)$, decomposed $\mathbf{w}$ into components parallel and orthogonal to $\mathbf{w}^*$, and applied the definition of $\mathrm{T}_{\cos\theta}$.

Let $\sigma(t) \doteq \sum_{i=0} a_i \mathrm{He}_i(t)$ with $a_i = \mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)\mathrm{He}_i(z)]$ be the Hermite expansion of $\sigma$. Using the property that $\mathrm{T}_\rho \mathrm{He}_i(t) = \rho^i \mathrm{He}_i(t)$, for any integer $k \ge 1$ (see Fact 7.1.7, Part 3), we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{t\sim\mathcal{N}}[(\sigma(t))^2] - \mathop{\mathbf{E}}_{t\sim\mathcal{N}}[\sigma(t)\mathrm{T}_{\cos\theta}\sigma(t)] &= \sum_{i=1}^{+\infty} a_i^2(1-\cos^i\theta)\\ &= a_1^2(1-\cos\theta) + \sum_{i=2}^{+\infty} a_i^2(1-(1-\sin^2\theta)^{i/2})\\ &\le (1/2)a_1^2\theta^2 + \theta^2\sum_{i=2}^{k}(i/2)a_i^2 + \sum_{i=k+1}^{+\infty} a_i^2, \qquad (7.8)\end{aligned}$$

where for the first term we used that $(1-\cos\theta) = 2\sin^2(\theta/2) \le \theta^2/2$, and for the terms from $i = 2, \ldots, k$, we used the Bernoulli inequality and that $\sin\theta \le \theta$ for $\theta \ge 0$.

Furthermore, note that if $\sigma(t) \doteq \sum_{i=0} a_i \mathrm{He}_i(t)$ is the Hermite expansion of $\sigma$, then $\sigma'(t) \doteq \sum_{i=1} a_i \sqrt{i} \mathrm{He}_{i-1}(t)$ is the Hermite expansion of $\sigma'$. Therefore, we have that

$$\mathop{\mathbf{E}}_{t\sim\mathcal{N}}[(\sigma(t))^2] - \mathop{\mathbf{E}}_{t\sim\mathcal{N}}[\sigma(t)\mathrm{T}_{\cos\theta}\sigma(t)] \le \theta^2 \|\mathrm{P}_k\sigma'\|_{L_2}^2 + \|\mathrm{P}_{>k}\sigma\|_{L_2}^2,$$

which, combined with Equation (7.7), completes the proof. □

To complete the proof, it remains to bound $\|\mathrm{P}_k\sigma'\|_{L_2}^2$ above, which is done in the following claim.

**Claim 7.2.10.** *When $k \ge 2$, for any $c \in [1, (k/2)^{1/4}]$,*

$$\|\mathrm{P}_k\sigma'(t)\|_{L_2}^2 \le 2e^{c^2} \|\mathrm{T}_{\sqrt{1-c^2/k}}\sigma'(t)\|_{L_2}^2.$$

*Proof.* Note that the $\mathrm{T}_{\sqrt{1-c^2/k}}\sigma'(t) \doteq \sum_{i=1}(1-c^2/k)^{(i-1)/2}\sqrt{i}a_i\mathrm{He}_{i-1}(t)$. Therefore, we have that

$$\|\mathrm{T}_{\sqrt{1-c^2/k}}\sigma'\|_{L_2}^2 = \sum_{i=1}^{+\infty} i(1-c^2/k)^{i-1}a_i^2 \ge \sum_{i=1}^{k} i(1-c^2/k)^i a_i^2 \ .$$

By assumption, we have $c^4/k \le 1/2$. Therefore, for any $i \le k$, using the inequality $(1-c^2/k)^i \ge (1-c^2/k)^k \ge e^{-c^2}(1-c^4/k) \ge e^{-c^2}/2$, we have that

$$\|\mathrm{T}_{\sqrt{1-c^2/k}}\sigma'\|_{L_2}^2 \ge \sum_{i=1}^{k} ie^{-c^2}(1/2)a_i^2 = e^{-c^2}(1/2)\|\mathrm{P}_k\sigma'\|_{L_2}^2 \ .$$

This completes the proof. □

Combining Claim 7.2.10 with Claim 7.2.9 and bring back the bounds to Equa-

tion (7.6), we have that when $k \geq 2$ and for any $c \in [1, (k/2)^{1/4}]$, it holds

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] \leq 2\mathrm{OPT} + 8e^{c^2}\theta^2\|\mathrm{T}_{\sqrt{1-c^2/k}}\sigma'\|_{L_2}^2 + 4\|\mathrm{P}_{>k}\sigma\|_{L_2}^2 \ .$$

When $k = 0$, then according to Equation (7.8), for any $\rho \in (0, 1)$, we have:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - \sigma(\mathbf{w}\cdot\mathbf{x}))^2] \leq 2\|\mathrm{P}_{>0}\sigma\|_{L_2}^2 \leq \frac{1}{2}\theta^2\|\mathrm{T}_\rho\sigma'\|_{L_2}^2 + 2\|\mathrm{P}_{>0}\sigma\|_{L_2}^2,$$

since $\frac{1}{2}\theta^2\|\mathrm{T}_\rho\sigma'\|_{L_2}^2 \geq 0$ for any $\rho \in (0, 1)$. When $k = 1$, similarly according to Equation (7.8), we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - \sigma(\mathbf{w}\cdot\mathbf{x}))^2] \leq \frac{1}{2}\theta^2 a_1^2 + 2\|\mathrm{P}_{>1}\sigma\|_{L_2}^2.$$

Observe that for any $\rho \in (0, 1)$, we have $\mathrm{T}_\rho\sigma'(z) \doteq \sum_{i\geq 1}\rho^{i-1}\sqrt{i}a_i\mathrm{He}_{i-1}(z)$, therefore $\|\mathrm{T}_\rho\sigma'\|_{L_2}^2 = \sum_{i\geq 1}\rho^{2(i-1)}ia_i^2 \geq a_1^2$. Thus, we further obtain

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - \sigma(\mathbf{w}\cdot\mathbf{x}))^2] \leq \frac{1}{2}\theta^2\|\mathrm{T}_\rho\sigma'\|_{L_2}^2 + 2\|\mathrm{P}_{>1}\sigma\|_{L_2}^2.$$

Plugging the above bounds for the cases $k = 0, 1$ back into Equation (7.6) completes the proof. □

Having proved Lemma 7.2.8, it is not hard to see that Proposition 7.2.7 follows as a direct corollary:

*Proof of Proposition 7.2.7.* Since $\theta^*$ is the $(\sigma, \theta_0, C\mathrm{OPT})$-Critical point, we have

$$(\theta^*)^2\|\mathrm{T}_{\cos(c\theta^*)}\sigma'\|_{L_2}^2 \leq C\mathrm{OPT} \ .$$

We apply Lemma 7.2.8 with $k = \lfloor 1/(\theta^*)^2 \rfloor$. Consider first $\theta^* \le 1/\sqrt{2}$, which implies that $k \ge 2$. Then for any $c \in [1, (k/2)^{1/4}]$, since $c\theta^* \ge \sin(c\theta^*)$ and $c \ge 1$, it holds

$$\sqrt{1 - \frac{c^2}{k}} \le \sqrt{1 - (c\theta^*)^2} \le \sqrt{1 - \sin^2(c\theta^*)} = \cos(c\theta^*) \le \cos\theta^*.$$

Thus as $\|\mathrm{T}_\rho \sigma'\|_{L_2}$ is non-decreasing with respect to $\rho$ (Fact 7.1.7), we further have $\|\mathrm{T}_{\sqrt{1-c^2/k}}\sigma'\|_{L_2}^2 \le \|\mathrm{T}_{\cos(\theta^*)}\sigma'\|_{L_2}^2$. Therefore, applying Lemma 7.2.8, for any $\theta \le \theta^*$, we obtain

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] &\le 2\mathrm{OPT} + 8e^{c^2}\theta^2 \|\mathrm{T}_{\sqrt{1-c^2/k}}\sigma'\|_{L_2}^2 + 4\|\mathrm{P}_{>k}\sigma\|_{L_2}^2 \\
&\le 2\mathrm{OPT} + 8e^{c^2}(\theta^*)^2 \|\mathrm{T}_{\cos(\theta^*)}\sigma'\|_{L_2}^2 + 4\|\mathrm{P}_{>(1/\theta^*)^2}\sigma\|_{L_2}^2 \\
&\le C'\mathrm{OPT} + 4\|\mathrm{P}_{>(1/\theta^*)^2}\sigma\|_{L_2}^2 ,
\end{aligned}$$

where $C'$ is an absolute constant. In particular, when $c = 1$, we have $C' = 2 + 8eC$. When $\theta^* > 1/\sqrt{2}$, then $k = 0, 1$. Choose $\rho = \cos(\theta^*) \in (0, 1)$ in Lemma 7.2.8, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] &\le 2\mathrm{OPT} + \theta^2 \|\mathrm{T}_{\cos(\theta^*)}\sigma'\|_{L_2}^2 + 4\|\mathrm{P}_{>k}\sigma\|_{L_2}^2 \\
&\le (2 + C)\mathrm{OPT} + 4\|\mathrm{P}_{>(1/\theta^*)^2}\sigma\|_{L_2}^2
\end{aligned}$$

In summary, for all $\theta^* \in (0, \pi/2)$, we have $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] \le O(\mathrm{OPT}) + 4\|\mathrm{P}_{>(1/\theta^*)^2}\sigma\|_{L_2}^2$. □

## 7.3 Learning GLMs via Variable Augmentation

In this section, we present our main algorithm (Algorithm 14) for learning GLMs under Gaussian marginals with adversarial corruptions, as stated in Definition 1.3.15. Our algorithm applies to the following large class of activations:

**Definition 7.3.1** ($(B, L)$-Regular Activations)**.** *Given parameters $B, L > 0$, we define the class of $(B, L)$-Regular activations, denoted by $\mathcal{H}(B, L)$, as the class containing all functions $\sigma : \mathbb{R} \to \mathbb{R}$ such that 1) $\|\sigma\|_{L_\infty} \leq B$ and 2) $\|\sigma'\|_{L_2} \leq L$. Given $\epsilon > 0$, we define the class of $\epsilon$-Extended $(B, L)$-Regular activations, denoted by $\mathcal{H}_\epsilon(B, L)$, as the class containing all activations $\sigma_1 : \mathbb{R} \to \mathbb{R}$ for which there exists $\sigma_2 \in \mathcal{H}(B, L)$ such that $\|\sigma_1 - \sigma_2\|_{L_2}^2 \leq \epsilon$.*

Algorithm 14 uses the main structural result of Section 7.2 (Proposition 7.2.3) to update its iterates $\mathbf{w}^{(t)}$. In particular, for $\theta_t = \theta(\mathbf{w}^{(t)}, \mathbf{w}^*)$, we show that after one gradient descent-style update, the angle $\theta_{t+1}$ shrinks by a factor $1-c$, i.e., $\theta_{t+1} \leq (1-c)\theta_t$, where $0 < c < 1$ is an absolute constant. A crucial feature of Algorithm 14 is that in each iteration it carefully chooses a new value of $\rho_t$. This *variable* update of $\rho_t$ ensures the 'signal' of the gradient is present until $\mathbf{w}^{(t)}$ reaches a small region centered at $\mathbf{w}^*$. Within this region, the agnostic noise corrupts the signal of the augmented gradient and convergence to $\mathbf{w}^*$ is no longer be guaranteed. However, the region that $\mathbf{w}^{(t)}$ reaches is in fact the Convergence Region $\mathcal{R}_{\sigma,\theta_0}(O(\mathrm{OPT}))$, within which all points are solutions with the target approximation error. We will show in Section 7.4 that for any *monotone* $(B, L)$-Regular activations, any point $\widehat{\mathbf{w}}$ in $\mathcal{R}_{\sigma,\theta_0}(O(\mathrm{OPT}))$ is a solution with error $C\mathrm{OPT} + \epsilon$, provided that the initialized angle $\theta_0 = \theta(\mathbf{w}^{(0)}, \mathbf{w}^*)$ is suitably small.

We now present our main algorithm.

**Algorithm 14** SGD-VA: SGD with Variable Augmentation

**Input:** Parameters $\epsilon, T$; sample access to $\mathcal{D}$.
$[\mathbf{w}^{(0)}, \bar{\theta}] \leftarrow$ **Initialization**$[\sigma]$ (see Section (7.4.3)); set $\rho_0 \leftarrow \cos\bar{\theta}$.
**for** $t = 0$ **to** $T$ **do**
  Draw $n$ samples $\{(\tilde{\mathbf{x}}^{(i)}, y^{(i)})\}_{i=1}^{n}$ from $\mathcal{D}_{\rho_t}$ using Algorithm 13; form empirical $\widehat{\mathcal{D}}_{\rho_t}$.
  $\hat{\mathbf{g}}(\mathbf{w}^{(t)}) \leftarrow -\frac{1}{\rho_t}\,\mathbf{E}_{(\tilde{\mathbf{x}},y)\sim\widehat{\mathcal{D}}_{\rho_t}}\left[y\,\sigma'(\mathbf{w}^{(t)}\cdot\tilde{\mathbf{x}})\,(\tilde{\mathbf{x}})^{\perp\mathbf{w}^{(t)}}\right]$
  $\eta_t \leftarrow \dfrac{\sqrt{(1-\rho_t)/2}}{4\,\|\hat{\mathbf{g}}(\mathbf{w}^{(t)})\|_2}$
  $\mathbf{w}^{(t+1)} \leftarrow \dfrac{\mathbf{w}^{(t)} - \eta_t\,\hat{\mathbf{g}}(\mathbf{w}^{(t)})}{\|\mathbf{w}^{(t)} - \eta_t\,\hat{\mathbf{g}}(\mathbf{w}^{(t)})\|_2}$
  $\rho_{t+1} \leftarrow 1 - (1 - \frac{1}{256})^2\left(1 - \rho_t\right)$
**end for**
$\hat{\mathbf{w}} \leftarrow$ **Test**$[\mathbf{w}^{(0)}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)}]$ (see Algorithm 15)
**Return:** $\hat{\mathbf{w}}$

Our main result concerning the general setting of $(B, L)$-Regular activations is summarized in the following theorem.

**Theorem 7.3.2.** *Let $\epsilon > 0$. Let $\sigma$ be a $(B, L)$-Regular activation. Algorithm 14 given initialization $\mathbf{w}^{(0)}$ with $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \le \bar{\theta}$, runs at most $T = O(\log(L/\epsilon))$ iterations, draws $\tilde{\Theta}(dB^2\log(L/\epsilon)/\epsilon + B^4\log(L/\epsilon)/\epsilon^2)$ samples, and returns a vector $\hat{\mathbf{w}}$ such that with probability at least $2/3$, $\hat{\mathbf{w}}$ lies in the target region $\mathcal{R}_{\sigma,\theta_0}(O(\mathrm{OPT}))$. Moreover, $\mathcal{L}(\hat{\mathbf{w}}) = O(\mathrm{OPT}) + \epsilon + 4\|\mathrm{P}_{>1/(\theta^*)^2}\sigma\|_{L_2}^2$.*

Let us provide a roadmap of the proof. Suppose for simplicity that we have taken enough samples so that we have access to the population gradient $\mathbf{g}(\mathbf{w}^{(t)})$. Furthermore, for the convenience of notation, let us use $\zeta(\rho)$ to denote the value $\sqrt{\mathrm{OPT}}/\|\mathrm{T}_\rho\sigma'\|_{L_2}$.

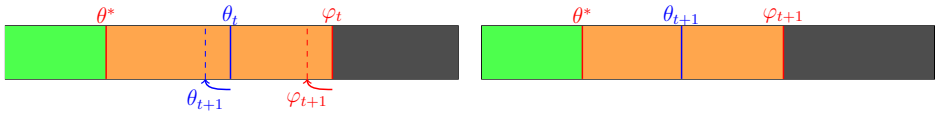



Figure 7.2: Successfull Update

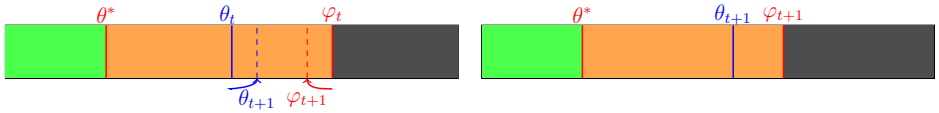



Figure 7.3: Wrong Update

Figure 7.4: Illustration of $\theta^*$, $\theta_t$, and $\varphi_t$ at different stages. The green region represents the Convergence Region, while the black region denotes the area that $\theta_t$ will never enter. Notably, the black region consistently expands, irrespective of whether the update is successful. The parameter $\theta_t$ is always guaranteed to never reach the black region.

Our main tool is the structural result in Proposition 7.2.3, which shows that when

$$\text{conditions for fast convergence: } \sin\theta_t \geq 3\zeta(\rho_t),\ \zeta(\rho_t) \stackrel{\text{def}}{=} \sqrt{\text{OPT}}/\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2},\ \rho_t \leq \cos\theta_t \tag{7.9}$$

are satisfied, the gradient $\mathbf{g}(\mathbf{w}^{(t)})$ correlates strongly with the target vector $\mathbf{w}^*$, hence providing sufficient information about the direction of $\mathbf{w}^*$. This structural result enables us to decrease the angle $\theta_{t+1}$ efficiently so that $\theta_{t+1} \leq (1-c)\theta_t$. However, two critical problems arise:

1. If $\sin\theta_t \lesssim \zeta(\rho_t)$, then since the conditions in Equation (7.9) are not valid, we cannot guarantee that the angle $\theta_t$ contracts. On the other hand, since $\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2} \leq \|\mathrm{T}_{\cos\theta_t}\sigma'\|_{L_2}$, it is not necessarily the case that $\sin\theta_t \lesssim \zeta(\cos\theta_t)$, therefore we also cannot assert that $\mathbf{w}^{(t)}$ has reached the target region $\mathcal{R}_{\sigma,\theta_0}(C^2\text{OPT})$.

2. Suppose that the conditions in Equation (7.9) are satisfied, and we have contraction of angle $\theta_{t+1} \leq (1-c)\theta_t$. Assume that $\mathbf{w}^{(t+1)}$ is still far away from $\mathbf{w}^*$ and $\theta_{t+1} \gtrsim$

$\zeta(\cos\theta_{t+1})$, meaning that we still need to further decrease the angle between $\mathbf{w}^{(t+1)}$ and $\mathbf{w}^*$. However, it is possible that $\zeta(\cos\theta_{t+1}) \lesssim \theta_{t+1} \lesssim \zeta(\rho_t)$, because $\|\mathrm{T}_{\cos\theta_{t+1}}\sigma'\|_{L_2} \geq \|\mathrm{T}_{\rho_t}\sigma'\|_{L_2}$, as we have $\rho_t \leq \cos\theta_t \leq \cos\theta_{t+1}$ and $\|\mathrm{T}_{\rho}\sigma'\|_{L_2}$ is an increasing function of $\rho$ (see Fact 7.1.7). This implies that the fast convergence conditions (Equation (7.9)) might be invalid if we continue using $\rho_t$. Thus, we need to carefully increase $\rho_t$ to $\rho_{t+1}$ so that $\sin\theta_{t+1} \gtrsim \zeta(\rho_{t+1})$, while maintaining the other condition $\rho_{t+1} \leq \cos\theta_{t+1}$. This seems impossible since we do not have any lower bound on $\theta_{t+1}$.

To overcome these hurdles, let us study the event $\mathcal{E}_t \stackrel{\text{def}}{=} \{|\cos\theta_t - \rho_t| \leq \sin^2\theta_t, \sin\theta_t \leq C\zeta(\rho_t)\}$. We first observe that when $\mathcal{E}_t$ is satisfied, then, since $\sin\theta_t \leq C\zeta(\rho_t)$ the algorithm may not be converging anymore, as discussed in Case 1 above. However, since $\mathcal{E}_t$ also satisfies $|\cos\theta_t - \rho_t| \leq \sin^2\theta_t$, one can show that $\zeta(\rho_t) \approx \zeta(\cos\theta_t)$, therefore, we have that $\sin\theta_t \leq C\zeta(\cos\theta_t)$. In other words, we can certify that $\mathbf{w}^{(t)}$ lies in the target region $\mathcal{R}_{\sigma,\theta_0}(C^2\mathrm{OPT})$. This solves the first problem above.

Now suppose $\mathcal{E}_t$ is not satisfied. We use induction to show that updating $\rho_t$ by Line 8, it always holds $\rho_{t+1} \leq \cos\theta_{t+1}$. To see this, suppose $\rho_t \leq \cos\theta_t$ holds at iteration $t$ and $\mathcal{E}_t$ is not satisfied. Then if we have $\sin\theta_t \geq C\zeta(\rho_t)$, the conditions in Equation (7.9) are satisfied hence we have control of $\theta_{t+1}$. We can then show that $\rho_{t+1} \leq \cos\theta_t$ and $\sin\theta_{t+1} \gtrsim \zeta(\rho_{t+1})$ with $\rho_{t+1}$ defined by Line 8. On the other hand, if $|\cos\theta_t - \rho_t| \geq \sin^2\theta_t$, then since $\rho_t \leq \cos\theta_t$ we know that $\rho_t$ is much smaller compared to $\cos\theta_t$. Thus, since we are taking small gradient steps and making very small increments to $\rho_t$, we have that $\rho_{t+1} \leq \cos\theta_{t+1}$ and $\sin\theta_{t+1} \gtrsim \zeta(\rho_{t+1})$ continue to hold. This resolves the second problem. See Figure 7.4 for a visual illustration of the mechanism of Algorithm 14.

We can now proceed to the proof of Theorem 7.3.2.

*Proof of Theorem 7.3.2.* In the proof, we denote the angle between $\mathbf{w}^{(t)}$ and $\mathbf{w}^*$ by $\theta_t =$

$\theta(\mathbf{w}^{(t)}, \mathbf{w}^*)$ and denote $\widehat{\mathbf{g}}(\mathbf{w}^{(t)})$ by $\widehat{\mathbf{g}}^{(t)}$. After initialization, it holds $\theta_0 = \theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \leq \bar{\theta}$. Furthermore, the algorithm uses the following parameters: $\varphi_t = (1/\sqrt{2})(1-\beta)^t \sin\bar{\theta}$, $\beta = 1/256$; $\rho_t \stackrel{\text{def}}{=} 1 - 2\varphi_t^2$, $\rho_0 = \cos\bar{\theta}$; $\eta_t = \varphi_t/(4\|\widehat{\mathbf{g}}^{(t)}\|_2)$. Note that if $\epsilon \geq C\text{OPT}$, then we can run the algorithm with $\epsilon' = \epsilon/(2C)$ and assume that we have more noise of order $\text{OPT}' = 2\epsilon'$. In this case, the final error bound will be $C\text{OPT}' \leq \epsilon/2 \leq \text{OPT} + \epsilon$. So, without loss of generality, we can assume that $\epsilon \leq \text{OPT}$. The goal of the algorithm is to converge to a vector in the region $\mathcal{R}_{\sigma,\theta_0}(C\text{OPT})$ where $C > 0$ is an absolute constant. For this reason, we consider the following event

$$\mathcal{E}_t \stackrel{\text{def}}{=} \left\{ |\cos\theta_t - \rho_t| \leq \sin^2\theta_t,\, \sin\theta_t \leq \frac{C\sqrt{\text{OPT}}}{\|\text{T}_{\rho_t}\sigma'\|_{L_2}} \right\},$$

where $C > 0$ is an absolute constant.

We argue that if $\mathcal{E}_t$ is satisfied at some iteration $t$, then the algorithm converges to a vector that lies in the $\mathcal{R}_{\sigma,\theta_0}(C\text{OPT})$ region. We consider two cases, the first case is if $\rho_t \geq \cos\theta_t$ and the second one if $\rho_t \leq \cos\theta_t$. Assume first that $1 > \rho_t \geq \cos\theta_t$. Then, since $\|\text{T}_\rho\sigma'\|_{L_2}$ is an increasing function with respect to the variable $\rho$ (see Fact 7.1.7), and $\mathcal{E}_t$ implies $\sin\theta_t \leq C\sqrt{\text{OPT}}/\|\text{T}_{\rho_t}\sigma'\|_{L_2}$, we also have that $\sin\theta_t \leq C\sqrt{\text{OPT}}/\|\text{T}_{\cos\theta_t}\sigma'\|_{L_2}$ and therefore that means that $\mathbf{w}^{(t)}$ is inside the region $\mathcal{R}_{\sigma,\theta_0}(C^2\text{OPT})$. Next, we consider the case where $\rho_t \leq \cos\theta_t$. Since $\mathcal{E}_t$ implies that $|\cos\theta_t - \rho_t| \leq \sin^2\theta_t$, we further have

$$\rho_t \geq \cos\theta_t - \sin^2\theta_t \geq \cos^2\theta_t - \sin^2\theta_t = \cos(2\theta_t).$$

Therefore, we have $\sin\theta_t \leq C\sqrt{\text{OPT}}/\|\text{T}_{\rho_t}\sigma'\|_{L_2} \leq C\sqrt{\text{OPT}}/\|\text{T}_{\cos(2\theta_t)}\sigma'\|_{L_2}$, i.e., $\psi_\sigma(2\theta_t) \leq 2C\sqrt{\text{OPT}}$. Let $\theta^*$ be the $(\sigma, \theta_0, 4C^2\text{OPT})$-Critical Point, we thus have $2\theta_t \leq \theta^*$ and $\mathbf{w}^{(t)}$ is inside the region $\mathcal{R}_{\sigma,\theta_0}(4C^2\text{OPT})$. Now since $\theta_t \leq \theta^*$, applying Proposition 7.2.7

yields

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x})-y)^2] \le O(\mathrm{OPT}) + 4\|\mathrm{P}_{>1/(\theta^*)^2}\sigma\|_{L_2}^2.$$

indicating that $\mathbf{w}^{(t)}$ is a solution that achieves the target error.

We proceed to show that the algorithm is guaranteed to generate a solution $\mathbf{w}^{(t^*)}$ that satisfies the event $\mathcal{E}_{t^*}$ at some iteration $t^* \le T = O(\log(\|\sigma'\|_{L_2}^2/\mathrm{OPT}))$. Our strategy is to prove that in every iteration $t \le t^*$, it holds that $\rho_t \le \cos\theta_t$, due to the careful design of the algorithm. Furthermore, we guarantee that $\rho_t$ can grow geometrically; therefore, we obtain an exponentially growing lower bound on $\cos\theta_t$, which implies that $\sin\theta_t$ shrinks at a linear rate and hence the event $\mathcal{E}_t$ will eventually be satisfied at some iteration $t^*$.

**Claim 7.3.3.** *Let $t'$ be the maximum $t \in [0,T]$ such that for all $t = 0,\dots,t'$, $\mathcal{E}_t$ is not satisfied. Then, for all $t \le t'$, it holds that $\rho_t \le \cos\theta_t$.*

*Proof of Claim 7.3.3.* We use induction to show the claim that $\rho_t \le \cos\theta_t$ for all the iterations $t = 0,\dots,t'$ where the event $\mathcal{E}_t$ is not satisfied.

**Base Case $t = 0$.** Recall that:

$$\varphi_t = (1/\sqrt{2})(1-\beta)^t \sin\bar{\theta},\ \beta = 1/256;\ \rho_t \stackrel{\mathrm{def}}{=} 1 - 2\varphi_t^2,\ \rho_0 = \cos\bar{\theta};\ \eta_t = \varphi_t/(4\|\widehat{\mathbf{g}}^{(t)}\|_2).$$

Therefore, since $\theta_0 \le \bar{\theta}$, $\rho_0 = \cos(\bar{\theta}) \le \cos\theta_0$ is satisfied in the base case.

**Induction Step.** For the induction step, suppose that $\mathcal{E}_t$ is not satisfied, in other words, we have either $|\cos\theta_t - \rho_t| \ge \sin^2\theta_t$ or $\sin\theta_t \ge C\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2}$. Assume that $\rho_t \le \cos\theta_t$ for the iterations $0,\dots,t$. We argue that $\rho_{t+1} \le \cos\theta_{t+1}$ continues to

hold after one iteration.

**Case I.** Consider first the case where $\sin\theta_t \geq C\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2}$. We study the distance between $\mathbf{w}^{(t)}$ and $\mathbf{w}^*$ after one iteration from $t$ to $t+1$. Since $\widehat{\mathbf{g}}^{(t)}$ is orthogonal to $\mathbf{w}^{(t)}$, it must be $\|\mathbf{w}^{(t)} - \eta_t\widehat{\mathbf{g}}^{(t)}\|_2 \geq 1$, therefore, $\mathbf{w}^{(t+1)} = \mathrm{proj}_{\mathbb{B}}(\mathbf{w}^{(t)} - \eta_t\widehat{\mathbf{g}}^{(t)})$. By the non-expansiveness of the projection operator, we have

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 &= \|\mathrm{proj}_{\mathbb{B}}(\mathbf{w}^{(t)} - \eta_t\widehat{\mathbf{g}}^{(t)}) - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}^{(t)} - \eta_t\widehat{\mathbf{g}}^{(t)} - \mathbf{w}^*\|_2^2 \\ &= \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \eta_t^2\|\widehat{\mathbf{g}}^{(t)}\|_2^2 - 2\eta_t\widehat{\mathbf{g}}^{(t)}(\mathbf{w}^{(t)} - \mathbf{w}^*) \\ &= \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \eta_t^2\|\widehat{\mathbf{g}}^{(t)}\|_2^2 + 2\eta_t\widehat{\mathbf{g}}^{(t)} \cdot \mathbf{w}^*. \end{aligned} \tag{7.10}$$

Next, we use the following lemma about the concentration of $\widehat{\mathbf{g}}$.

**Lemma 7.3.4.** *Suppose $\sigma$ is a $(B, L)$-Regular activation. If $0 < \rho \leq \cos\theta < 1$ and $\sin\theta \geq 4(\sqrt{\mathrm{OPT}})/\|\mathrm{T}_\rho\sigma'\|_{L_2}$, then using*

$$n = \Theta\left(\frac{dB^2}{\sin^2\theta\|\mathrm{T}_\rho\sigma'\|_{L_2}^2\delta}\right)$$

*samples, with probability at least $1-\delta$, we have*

$$\begin{aligned}\|\widehat{\mathbf{g}}(\mathbf{w})\|_2 &\leq (3/2)\,\mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})]\sin\theta\,, \\ \widehat{\mathbf{g}}(\mathbf{w})\cdot\mathbf{w}^* &\leq -\frac{1}{2}\,\mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})]\sin^2\theta.\end{aligned}$$

We defer the proof of Lemma 7.3.4 to Section (7.3.2). Using Lemma 7.3.4, we know that with a batch size of

$$n = \Theta\left(\frac{dB^2}{\sin^2\theta\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2}^2\delta}\right) \leq \Theta\left(\frac{dB^2}{\epsilon\delta}\right)$$

and if the following conditions are satisfied

$$\rho_t \leq \cos\theta_t, \tag{7.11}$$

$$\text{and } \sin\theta_t \geq C\sqrt{\text{OPT}}/\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2}, \tag{7.12}$$

then, with probability at least $1-\delta$, we have that

$$\widehat{\mathbf{g}}^{(t)} \cdot \mathbf{w}^* \leq -(1/2)\,\mathbf{E}[\mathrm{T}_{\cos\theta_t}\sigma'(\mathbf{w}^{(t)}\cdot\mathbf{x})\mathrm{T}_{\rho_t}\sigma'(\mathbf{w}^{(t)}\cdot\mathbf{x})]\sin^2\theta_t, \tag{7.13}$$

$$\|\widehat{\mathbf{g}}^{(t)}\|_2 \leq 2\,\mathbf{E}[\mathrm{T}_{\cos\theta_t}\sigma'(\mathbf{w}^{(t)}\cdot\mathbf{x})\mathrm{T}_{\rho_t}\sigma'(\mathbf{w}^{(t)}\cdot\mathbf{x})]\sin\theta_t. \tag{7.14}$$

Combining Equation (7.13) with Equation (7.14)we get $\widehat{\mathbf{g}}^{(t)}\cdot\mathbf{w}^* \geq (1/4)\|\widehat{\mathbf{g}}^{(t)}\|_2\sin\theta_t$. Therefore, bringing in our choice of stepsize $\eta_t = \varphi_t/(4\|\widehat{\mathbf{g}}^{(t)}\|_2)$, and noticing $\sin\theta_t \geq \sin(\theta_t/2)$ we obtain:

$$\begin{aligned} 4\sin^2(\theta_{t+1}/2) = \|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|_2^2 &\leq \|\mathbf{w}^{(t)}-\mathbf{w}^*\|_2^2 + \frac{\varphi_t^2}{16} - \frac{\varphi_t\sin\theta_t}{8} \\ &\leq 4\sin^2(\theta_t/2) + \frac{\varphi_t}{16}(\varphi_t - 2\sin(\theta_t/2)). \end{aligned} \tag{7.15}$$

Since, by assumption, we have $\rho_t \leq \cos\theta_t$, it holds $2\sin^2(\theta_t/2) = 1-\cos\theta_t \leq 1-\rho_t = 2\varphi_t^2$, in other words, $\sin(\theta_t/2) \leq \varphi_t$. Consider first the case that $\varphi_t \geq \sin(\theta_t/2) \geq (3/4)\varphi_t$. Then, according to Equation (7.15) we get

$$4\sin^2(\theta_{t+1}/2) \leq 4\sin^2(\theta_t/2) - \frac{1}{32}\varphi_t^2 \leq 4(1-1/128)\varphi_t^2.$$

Hence, since $1-1/128 \leq (1-1/256)^2$, we get $\sin(\theta_{t+1}/2) \leq (1-1/256)\varphi_t = \varphi_{t+1}$.

On the other hand, if $\sin(\theta_t/2) \leq (3/4)\varphi_t$, then by the triangle inequality and the

non-expansiveness of the projection operator, we have

$$\begin{aligned}
2\sin(\theta_{t+1}/2) &= \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 = \|\mathrm{proj}_{\mathbb{B}}(\mathbf{w}^{(t)} - \eta_t \widehat{\mathbf{g}}(\mathbf{w}^{(t)})) - \mathbf{w}^*\|_2 \\
&\leq \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta_t \widehat{\mathbf{g}}^{(t)}\|_2 \\
&\leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 + (1/4)\varphi_t = 2\sin(\theta_t/2) + (1/4)\varphi_t \\
&\leq (3/2)(\varphi_t/2) + (1/4)\varphi_t \leq (7/4)\varphi_t \ .
\end{aligned}$$

Therefore, it holds that $\sin(\theta_{t+1}/2) \leq (7/8)\varphi_t \leq \varphi_{t+1}$.

To conclude, we proved that if $\sin\theta_t \geq C\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2}$ and $\rho_t \leq \cos\theta_t$, we have $\sin(\theta_{t+1}/2) \leq \varphi_{t+1}$ after one step of the algorithm, which immediately implies that $\cos\theta_{t+1} = 1 - 2\sin^2(\theta_{t+1}/2) \geq 1 - 2\varphi_{t+1}^2 = \rho_{t+1}$. Note that when $\sin\theta_t \geq C\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2}$, our argument indicates that $\cos\theta_{t+1} \geq \rho_{t+1}$ holds regardless of whether $|\cos\theta_t - \rho_t| \leq \sin^2\theta_t$ or not.

**Case II.** It remains to consider the case where $|\cos\theta_t - \rho_t| \geq \sin^2\theta_t$. In fact, we consider the setting where $\cos\theta_t - \rho_t \geq \sin^2\theta_t$, because from the induction argument we have $\rho_t \leq \cos\theta_t$. Observe that this case only requires discussing the setting where $\sin\theta_t \leq C\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2}$, because if $\sin\theta_t \geq C\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2}$ then our previous argument already implies that $\rho_{t+1} \leq \cos\theta_{t+1}$ after one iteration. Therefore, assuming that $\sin\theta_t \leq C\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\rho_t}\sigma'\|_{L_2}$, applying triangle inequality and the non-expansiveness of projection operator, it holds

$$\begin{aligned}
2\sin(\theta_{t+1}/2) &= \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 = \|\mathrm{proj}_{\mathbb{B}}(\mathbf{w}^{(t)} - \eta_t \widehat{\mathbf{g}}(\mathbf{w}^{(t)})) - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(t)} - \eta_t \widehat{\mathbf{g}}^{(t)} - \mathbf{w}^*\|_2 \\
&\leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 + \eta_2 \|\widehat{\mathbf{g}}^{(t)}\|_2 = 2\sin(\theta_t/2) + \varphi_t/4.
\end{aligned}$$

Using the assumption $\cos\theta_t - \rho_t \geq \sin^2\theta_t$, we observe that

$$1 - \sin^2(\theta_t/2) - (1 - 2\varphi_t^2) \geq \sin^2\theta_t \geq 2\sin^2(\theta_t/2),$$

in other words, we have $\sin(\theta_t/2) \leq \sqrt{2/3}\varphi_t$. Hence, it holds

$$\sin(\theta_{t+1}/2) \leq \sin(\theta_t/2) + \varphi_t/8 \leq (\sqrt{2/3} + 1/8)\varphi_t \leq (1 - 1/256)\varphi_t = \varphi_{t+1}.$$

Since $\sin(\theta_{t+1}/2) \leq \varphi_{t+1}$, using similar argument we have $\cos\theta_{t+1} \geq 1 - 2\varphi_{t+1}^2 = \rho_{t+1}$, therefore the induction argument continues to hold at step $t+1$. □

In conclusion, from Claim 7.3.3, we have $\cos\theta_t \geq \rho_t$ holds for all the iterations $t = 0, \ldots, t^* - 1$. It remains to show that the event $\mathcal{E}_{t^*}$ is satisfied at some iteration $t^*$.

**Claim 7.3.5.** *If $T = c\log(\|\sigma'\|_{L_2}^2/\mathrm{OPT})$, where $c > 0$ is a sufficiently large absolute constant, then there exists $t^* \leq T$, so that the event $\mathcal{E}_{t^*}$ is satisfied.*

*Proof of Claim 7.3.5.* Since $\sin\theta_t \leq \varphi_t \leq (1-\beta)^t$ for all the iterations $0, 1, \ldots, t$ where the event $\mathcal{E}_t$ is not satisfied, we have $\theta_t \to 0$. After at most $T = (1/\beta)\log(\|\sigma'\|_{L_2}/\sqrt{\mathrm{OPT}}) = O(\log(\|\sigma'\|_{L_2}^2/\mathrm{OPT}))$ iterations, it must hold that $\sin\theta_t \leq \sqrt{\mathrm{OPT}}/\|\sigma'\|_{L_2}$. Note that $\sqrt{\mathrm{OPT}}/\|\sigma'\|_{L_2} \leq \sqrt{\mathrm{OPT}}/\|\mathrm{T}_\rho\sigma'\|_{L_2}$ for any $\rho \in (0,1)$ therefore there exists an iteration $t^*$ for which $\mathcal{E}_{t^*}$ is satisfied. □

Therefore, we guarantee that $\mathcal{E}_t$ will be satisfied in at most $T = O(\log(\|\sigma'\|_{L_2}^2/\mathrm{OPT})) = O(\log(L/\epsilon))$ iterations. Setting $\delta = 2/(3T)$ and using a union bound, we have that at most

$$N_1 = nT = \Theta\left(\frac{dB^2T^2}{\epsilon}\right) = \tilde{\Theta}\left(\frac{dB^2\log(L/\epsilon)}{\epsilon}\right)$$

samples suffices to guarantee that the algorithm generates a target solution with probability at least 2/3. To pick out the target vector $\widehat{\mathbf{w}}$ from the list, we can apply a testing procedure with $N_2 = \tilde{\Theta}(B^4/\epsilon^2)$ samples (see Algorithm 15 and Lemma 7.3.6 in Section (7.3.1)). Thus, in summary, the sample complexity is $N_1 + N_2 = \tilde{\Theta}(dB^2 \log(L/\epsilon)/\epsilon + B^4 \log(L/\epsilon)/\epsilon^2)$. □

### 7.3.1 Finding the Best Parameter

As we showed in Theorem 7.3.2, Algorithm 14 returns a list of vectors that contains at least one vector $\widehat{\mathbf{w}}$ in the target region $\mathcal{R}_{\sigma,\theta_0}(O(\mathrm{OPT}) + \epsilon)$. We now present a simple testing algorithm to identify one of such target vectors.

**Algorithm 15** Testing

1: **Input:** Vectors $\{\mathbf{w}^{(0)}, \ldots, \mathbf{w}^{(T)}\}$; number of samples $m$
2: Draw samples $\{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\} \sim \mathcal{D}$ and form empirical $\widehat{\mathcal{D}}$
3: **for** $t = 0$ **to** $T$ **do**
4: $\quad \widehat{\mathcal{L}}\left(\mathbf{w}^{(t)}\right) \leftarrow \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}}\left[\left(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}) - y\right)^2\right]$
5: **end for**
6: $\widehat{\mathbf{w}} \leftarrow \arg\min_{t\in\{0,\ldots,T\}} \widehat{\mathcal{L}}\left(\mathbf{w}^{(t)}\right)$
7: **Output:** $\widehat{\mathbf{w}}$

**Lemma 7.3.6** (Testing). *Let $\sigma$ be a $(B, L)$-Regular activation. Let $\{\mathbf{w}^{(t)}\}_{t\in[T]}$ be the list of vectors generated by Algorithm 14 with $T = O(\log(L/\epsilon))$. Let $t^* \in [T]$ be the index such that $\mathcal{L}(\mathbf{w}^{(t^*)}) \in \mathrm{argmin}_{t\in[T]} \mathcal{L}(\mathbf{w}^{(t)})$. We have*

*1. If $\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2 \le (\theta_{t^*})^2 \|\mathrm{T}_{\cos(\theta_{t^*})}\sigma'\|_{L_2}^2$ for an absolute constant $C$, then using*

$$m \le \Theta\left(\frac{B^2 \log(L/\epsilon)}{\epsilon}\right)$$

*samples, Algorithm 15 finds a vector* $\widehat{\mathbf{w}} \in \{\mathbf{w}^{(t)}\}_{t\in[T]}$ *such that* $\widehat{\mathbf{w}} \in \mathcal{R}_{\sigma,\theta_0}(O(\mathrm{OPT})+\epsilon)$ *and* $\mathcal{L}(\widehat{\mathbf{w}}) \le O(\mathrm{OPT}) + \epsilon$.

2. *Otherwise, using*

$$m \le \Theta\left(\frac{B^4 \log\log(L/\epsilon)}{\epsilon^2}\right)$$

*samples, Algorithm 15 outputs a vector* $\widehat{\mathbf{w}} \in \{\mathbf{w}^{(t)}\}_{t\in[T]}$ *such that* $\widehat{\mathbf{w}} \in \mathcal{R}_{\sigma,\theta_0}(O(\mathrm{OPT})+\epsilon)$ *and* $\mathcal{L}(\widehat{\mathbf{w}}) \le O(\mathrm{OPT}) + \epsilon + 4\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2$.

*Proof.* Let $t^* \in \mathrm{argmin}_{t\in[T]} \mathcal{L}(\mathbf{w}^{(t)})$. Let $\ell(\mathbf{w};\mathbf{x},y) \stackrel{\mathrm{def}}{=} (\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2$, and $\Delta(\mathbf{w}^1,\mathbf{w}^2) \stackrel{\mathrm{def}}{=} \ell(\mathbf{w}^1;\mathbf{x},y) - \ell(\mathbf{w}^2;\mathbf{x},y)$. Given a data set $\{(\mathbf{x}^{(i)},y^{(i)})\}_{i=1}^m$, we denote by $\widehat{\mathcal{L}}(\mathbf{w})$ the empirical version of $\mathcal{L}(\mathbf{w})$, i.e., $\widehat{\mathcal{L}}(\mathbf{w}) = (1/m)\sum_{i=1}^m \ell(\mathbf{w};\mathbf{x}^{(i)},y^{(i)})$.

Consider first the case when $\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2 \le C(\theta_{t^*})^2\|\mathrm{T}_{\cos(\theta_{t^*})}\sigma'\|_{L_2}^2$ for an absolute constant $C$. Our goal is to show that $\widehat{\mathcal{L}}(\mathbf{w}^{(t^*)})$ can be separated from all $\widehat{\mathcal{L}}(\mathbf{w}^{(t)})$ for all $\mathbf{w}^{(t)}$ with large $L_2^2$ error. As shown in Claim 7.5.6, when $\sigma$ is a $(B,L)$-Regular activation, labels $y$ can be assumed to be bounded above by $B$ without loss of generality. Therefore, the variance of $\Delta(\mathbf{w}^1,\mathbf{w}^2)$ is bounded by

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\Delta^2(\mathbf{w}^1,\mathbf{w}^2)] &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^1\cdot\mathbf{x}) - \sigma(\mathbf{w}^2\cdot\mathbf{x}))^2(\sigma(\mathbf{w}^1\cdot\mathbf{x}) + \sigma(\mathbf{w}^2\cdot\mathbf{x}) - 2y)^2] \\ &\le 16B^2\,\mathbf{E}[(\sigma(\mathbf{w}^1\cdot\mathbf{x}) - \sigma(\mathbf{w}^2\cdot\mathbf{x}))^2]\,.\end{aligned}$$

On the other hand, suppose without loss of generality that $\mathcal{L}(\mathbf{w}^1) \ge \mathcal{L}(\mathbf{w}^2)$; then, the

expectation of $\Delta(\mathbf{w}^1, \mathbf{w}^2)$ can be bounded below by

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\Delta(\mathbf{w}^1, \mathbf{w}^2)] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^1 \cdot \mathbf{x}) - \sigma(\mathbf{w}^2 \cdot \mathbf{x}))(\sigma(\mathbf{w}^1 \cdot \mathbf{x}) - \sigma(\mathbf{w}^2 \cdot \mathbf{x}) + 2\sigma(\mathbf{w}^2 \cdot \mathbf{x}) - 2y)] \\
&= \mathbf{E}[(\sigma(\mathbf{w}^1 \cdot \mathbf{x}) - \sigma(\mathbf{w}^2 \cdot \mathbf{x}))^2] + 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^1 \cdot \mathbf{x}) - \sigma(\mathbf{w}^2 \cdot \mathbf{x}))(\sigma(\mathbf{w}^2 \cdot \mathbf{x}) - y)] \\
&\geq \mathbf{E}[(\sigma(\mathbf{w}^1 \cdot \mathbf{x}) - \sigma(\mathbf{w}^2 \cdot \mathbf{x}))^2] - \frac{1}{2}\mathbf{E}[(\sigma(\mathbf{w}^1 \cdot \mathbf{x}) - \sigma(\mathbf{w}^2 \cdot \mathbf{x}))^2] - 2\mathcal{L}(\mathbf{w}^2) \\
&= \frac{1}{2}\mathbf{E}[(\sigma(\mathbf{w}^1 \cdot \mathbf{x}) - \sigma(\mathbf{w}^2 \cdot \mathbf{x}))^2] - 2\mathcal{L}(\mathbf{w}^2).
\end{aligned}$$

where in the last inequality we used Young's inequality $ab \geq -((1/4)a^2 + b^2)$.

Now consider first $\mathcal{L}(\mathbf{w}^2) \leq (1/8)\,\mathbf{E}[(\sigma(\mathbf{w}^1 \cdot \mathbf{x}) - \sigma(\mathbf{w}^2 \cdot \mathbf{x}))^2]$. Then, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\Delta(\mathbf{w}^1, \mathbf{w}^2)] \geq \frac{1}{4}\mathbf{E}[(\sigma(\mathbf{w}^1 \cdot \mathbf{x}) - \sigma(\mathbf{w}^2 \cdot \mathbf{x}))^2].$$

Therefore, using Markov's inequality, we have

$$\begin{aligned}
&\mathbf{Pr}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\left(\ell(\mathbf{w}^1; \mathbf{x}^{(i)}, y^{(i)}) - \ell(\mathbf{w}^2; \mathbf{x}^{(i)}, y^{(i)})\right) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\Delta(\mathbf{w}^1, \mathbf{w}^2)]\right| \geq \frac{1}{3}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\Delta(\mathbf{w}^1, \mathbf{w}^2)]\right] \\
&\leq \frac{\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\Delta^2(\mathbf{w}^1, \mathbf{w}^2)]}{m((1/3)\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\Delta(\mathbf{w}^1, \mathbf{w}^2)])^2} \leq \frac{cB^2}{m\,\mathbf{E}[(\sigma(\mathbf{w}^1 \cdot \mathbf{x}) - \sigma(\mathbf{w}^2 \cdot \mathbf{x}))^2]} \leq \frac{cB^2}{m\mathcal{L}(\mathbf{w}^2)}\,.
\end{aligned}$$

Let $\mathbf{w}^2 = \mathbf{w}^{(t^*)}$. With $m \geq (cB^2/(\mathcal{L}(\mathbf{w}^{(t^*)})\delta))$, the inequality above implies that when $\mathbf{w}^1 = \mathbf{w}^{(t)}$, $t \in [T]$, such that $\mathcal{L}(\mathbf{w}^{(t)}) > \mathcal{L}(\mathbf{w}^{(t^*)})$ and $\mathcal{L}(\mathbf{w}^{(t^*)}) \leq (1/8)\,\mathbf{E}[(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}) - \sigma(\mathbf{w}^{(t^*)} \cdot \mathbf{x}))^2]$, it holds with probability at least $1 - \delta$:

$$|\widehat{\mathcal{L}}(\mathbf{w}^{(t)}) - \widehat{\mathcal{L}}(\mathbf{w}^{(t^*)}) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\Delta(\mathbf{w}^{(t)}, \mathbf{w}^{(t^*)})]| \leq \frac{1}{3}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\Delta(\mathbf{w}^{(t)}, \mathbf{w}^{(t^*)})],$$

in other words, we have

$$\widehat{\mathcal{L}}(\mathbf{w}^{(t^*)}) \leq \widehat{\mathcal{L}}(\mathbf{w}^{(t)}) - \frac{2}{3} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\Delta(\mathbf{w}^{(t)}, \mathbf{w}^{(t^*)})] \,.$$

Let $\delta = 1/(3T)$. Applying a union bound to all $\mathbf{w}^{(t)}$, $t \in [T]$, we have that with probability at least $2/3$, using $m \geq (cB^2T/\mathcal{L}(\mathbf{w}^{(t^*)}))$ samples suffices to distinguish $\mathbf{w}^{(t^*)}$ from other vectors $\mathbf{w}^{(t)}$ that have large $L_2^2$ error.

But what if $\mathcal{L}(\mathbf{w}^{(t^*)}) \geq (1/8)\,\mathbf{E}[(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}) - \sigma(\mathbf{w}^{(t^*)} \cdot \mathbf{x}))^2]$? In this case, note that

$$\begin{aligned}
\mathcal{L}(\mathbf{w}^{(t^*)}) &\geq \frac{1}{8}\,\mathbf{E}[(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}) - \sigma(\mathbf{w}^{(t^*)} \cdot \mathbf{x}))^2] \\
&= \frac{1}{8}\Bigg( \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}) - y)^2] + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^{(t^*)} \cdot \mathbf{x}) - y)^2] \\
&\quad - 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}) - y)(\sigma(\mathbf{w}^{(t^*)} \cdot \mathbf{x}) - y)]\Bigg) \\
&\geq \frac{1}{8}\Bigg( \mathcal{L}(\mathbf{w}^{(t)}) + \mathcal{L}(\mathbf{w}^{(t^*)}) - \frac{1}{2}\mathcal{L}(\mathbf{w}^{(t)}) - 2\mathcal{L}(\mathbf{w}^{(t^*)})\Bigg),
\end{aligned}$$

where in the last inequality we used Young's inequality $ab \leq (1/4)a^2 + b^2$. The inequality above indicates that $\mathcal{L}(\mathbf{w}^{(t)}) \leq 18\mathcal{L}(\mathbf{w}^{(t^*)})$. Note that according to Theorem 7.3.2, $\mathbf{w}^{(t^*)}$ is guaranteed to reside in the $\mathcal{R}_{\sigma,\theta_0}(C\mathrm{OPT} + \epsilon)$ region, therefore, by definition of the region $\mathcal{R}_{\sigma,\theta_0}(C\mathrm{OPT} + \epsilon)$,

$$(\theta_{t^*})^2 \|\mathrm{T}_{\cos(\theta_{t^*})}\sigma'\|_{L_2}^2 \leq C\mathrm{OPT} + \epsilon.$$

When $\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2 \leq C(\theta_{t^*})^2\|\mathrm{T}_{\cos(\theta_{t^*})}\sigma'\|_{L_2}^2$, according to the error bound displayed in Proposition 7.2.7, we have $\mathcal{L}(\mathbf{w}^{(t^*)}) \leq O(\mathrm{OPT}) + 4C(\theta_{t^*})^2\|\mathrm{T}_{\cos(\theta_{t^*})}\sigma'\|_{L_2}^2 \leq O(\mathrm{OPT}) + 4C\epsilon$. Therefore, any vector $\mathbf{w}^{(t)}$ that satisfies $\mathcal{L}(\mathbf{w}^{(t^*)}) \geq (1/8)\,\mathbf{E}[(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}) - \sigma(\mathbf{w}^{(t^*)} \cdot$

$\mathbf{x}))^2]$ is in the $\mathcal{R}_{\sigma,\theta_0}(C_1\mathrm{OPT}+\epsilon)$ region and can be output as a constant factor solution.

In summary, when $\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2 \lesssim (\theta_{t^*})^2\|\mathrm{T}_{\cos(\theta_{t^*})}\sigma'\|_{L_2}^2$, using

$$m \le \Theta\bigg(\frac{B^2\log(L/\epsilon)}{\epsilon}\bigg)$$

samples suffices for the testing algorithm.

Now consider the general case where $\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2 \gtrsim (\theta_{t^*})^2\|\mathrm{T}_{\cos(\theta_{t^*})}\sigma'\|_{L_2}^2$. We still have $\mathcal{L}(\mathbf{w}^{(t^*)}) \le O(\mathrm{OPT}) + 4\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2$ but it is no longer acceptable to output a vector $\widehat{\mathbf{w}}$ such that $\mathcal{L}(\widehat{\mathbf{w}}) = C\mathcal{L}(\mathbf{w}^{(t^*)})$ for constant $C > 1$. This is because in the worst case the $L_2^2$ error of $\widehat{\mathbf{w}}$ can be as large as $\mathcal{L}(\widehat{\mathbf{w}}) = O(\mathrm{OPT}) + 4C\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2$. When $\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2$ is large, this error bound does not imply that $\widehat{\mathbf{w}}$ lies in the target region $\mathcal{R}_{\sigma,\theta_0}(O(\mathrm{OPT}))$. Therefore, we need a different analysis.

To find a vector $\widehat{\mathbf{w}}$ from the set $\{\mathbf{w}^{(t)}\}_{t\in[T]}$ such that $\widehat{\mathbf{w}} \in \mathcal{R}_{\sigma,\theta_0}(O(\mathrm{OPT})+\epsilon)$, we need to approximate $\mathcal{L}(\mathbf{w}^{(t)})$ to error at most $O(\mathrm{OPT})+\epsilon$ for each $\mathbf{w}^{(t)}$, $t \in [T]$. Since $|\ell(\mathbf{w};\mathbf{x},y)| \le 4B^2$, we know that $\ell(\mathbf{w};\mathbf{x},y)$ is a sub-Gaussian random variable with $\|\ell(\mathbf{w};\mathbf{x},y)\|_{\psi_2} \le cB^2$ for some absolute constant $c$. Then using Hoeffding's inequality, we have

$$\mathbf{Pr}\bigg[\bigg|\sum_{i=1}^m \frac{1}{m}\ell(\mathbf{w};\mathbf{x}^{(i)},y^{(i)}) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{w};\mathbf{x},y)]\bigg| \ge (\mathrm{OPT}+\epsilon)\bigg] \le \exp\bigg(-\frac{c'(\mathrm{OPT}+\epsilon)^2}{mB^4}\bigg).$$

Therefore, using $m \le B^4\log(1/\delta)/\epsilon^2$ samples suffices to approximate $\mathcal{L}(\mathbf{w})$ to error $\mathrm{OPT}+\epsilon$. Using a union bound on all $\mathbf{w}^{(t)}$, $t \in [T]$, and set $\delta = 1/(3T)$, we obtain that with probability at least $2/3$, using

$$m \le \Theta\bigg(\frac{B^4\log\log(L/\epsilon)}{\epsilon^2}\bigg)$$

samples, it holds

$$|\widehat{\mathcal{L}}(\mathbf{w}^{(t)}) - \mathcal{L}(\mathbf{w}^{(t)})| \leq \mathrm{OPT} + \epsilon$$

for all $\mathbf{w}^{(t)}$, $t \in [T]$. Therefore, since $\mathcal{L}(\mathbf{w}^{(t^*)}) \leq O(\mathrm{OPT}) + \epsilon + 4\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2$, by outputting $\widehat{\mathbf{w}} = \min\{\mathbf{w}^{(t)}, t \in [T] : \widehat{\mathcal{L}}(\mathbf{w}^{(t)})\}$ we guarantee that $\mathcal{L}(\widehat{\mathbf{w}}) \leq O(\mathrm{OPT}) + \epsilon + 4\|\mathrm{P}_{>1/(\theta_{t^*})^2}\sigma\|_{L_2}^2$, hence $\widehat{\mathbf{w}} \in \mathcal{R}_{\sigma,\theta_0}(O(\mathrm{OPT}) + \epsilon)$. □

### 7.3.2 Proof of Lemma 7.3.4

In this subsection, we provide technical lemmas that determine the number of samples required for each iteration. We start with Lemma 7.3.7 that bounds the population gradient $\|\mathbf{g}(\mathbf{w})\|_2$. Then in Lemma 7.3.4 we provide the sufficient batch size of samples per iteration, utilizing the bounds on $\|\mathbf{g}(\mathbf{w})\|_2$ and the truncated upper bounds on the activation $\sigma(z)$ and labels $y$.

**Lemma 7.3.7.** *Let* $\mathbf{g}(\mathbf{w}) \stackrel{\mathrm{def}}{=} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp\mathbf{w}}]$ *and* $\theta = \theta(\mathbf{w}, \mathbf{w}^*)$. *Then, we have*

$$\|\mathbf{g}(\mathbf{w})\|_2 \leq \sqrt{\mathrm{OPT}}\|\mathrm{T}_\rho\sigma'\|_{L_2} + \|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin\theta.$$

*If, in addition,* $0 < \rho \leq \cos\theta < 1$ *and* $\sin\theta \geq 4\sqrt{\mathrm{OPT}}/\|\mathrm{T}_\rho\sigma'\|_{L_2}$, *then*

$$\|\mathbf{g}(\mathbf{w})\|_2 \leq (5/4)\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin\theta$$

*Proof.* By the variational definition of vector norms, we have

$$\begin{aligned}\|\mathbf{g}(\mathbf{w})\|_2 &= \max_{\|\mathbf{u}\|_2=1} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp_\mathbf{w}}\cdot\mathbf{u}] \\ &= \max_{\|\mathbf{u}\|_2=1} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-\sigma(\mathbf{w}^*\cdot\mathbf{x}))\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp_\mathbf{w}}\cdot\mathbf{u}] + \mathbf{E}[\sigma(\mathbf{w}^*\cdot\mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp_\mathbf{w}}\cdot\mathbf{u}].\end{aligned} \tag{7.16}$$

Observe here that the maximizing $\mathbf{u}$ depends on the expectation defining $\mathbf{g}(\mathbf{w})$ and is thus deterministic. To bound the right-hand side in Equation (7.16), we fix an arbitrary unit vector $\mathbf{u}$ and bound the two summands. Using the Cauchy-Schwarz inequality, the first term in Equation (7.16) above can be bounded by:

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-\sigma(\mathbf{w}^*\cdot\mathbf{x}))\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp_\mathbf{w}}\cdot\mathbf{u}] \\ &\le \sqrt{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2]\,\mathbf{E}[(\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x}))^2(\mathbf{x}^{\perp_\mathbf{w}}\cdot\mathbf{u})^2]} \\ &\le \sqrt{\mathrm{OPT}}\sqrt{\mathbf{E}[(\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x}))^2]\,\mathbf{E}[(\mathbf{x}^{\perp_\mathbf{w}}\cdot\mathbf{u})^2]} \\ &= \sqrt{\mathrm{OPT}}\sqrt{\mathbf{E}[(\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x}))^2]},\end{aligned}$$

where in the second inequality we used the fact that $\mathbf{w}\cdot\mathbf{x}$ is independent of $\mathbf{x}^{\perp_\mathbf{w}}\cdot\mathbf{u}$, due to the Gaussianity. The last equality uses $\mathbf{x}^{\perp_\mathbf{w}}\cdot\mathbf{u}\sim\mathcal{N}(0,1)$, as $\mathbf{u}$ is independent of $\mathbf{x}$ and $\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$.

For the second term in Equation (7.16), observe that if $\mathbf{u}\perp\mathbf{w},\mathbf{w}^*$, then the expectation takes value zero due to the independence between Gaussian random variables $\mathbf{x}^{\perp_\mathbf{w}}\cdot\mathbf{u}$ and $\mathbf{w}\cdot\mathbf{x},\mathbf{w}^*\cdot\mathbf{x}$. Therefore, we only need to consider $\mathbf{u}$ in the span of $\mathbf{w},\mathbf{w}^*$, which can be expressed as $\mathbf{u}=\cos\alpha\mathbf{w}+\sin\alpha(\mathbf{w}^*)^{\perp_\mathbf{w}}/\|(\mathbf{w}^*)^{\perp_\mathbf{w}}\|_2$, for some $\alpha\in[0,2\pi]$. Thus, plugging this $\mathbf{u}$ back into the second term in Equation (7.16), and setting $z_1=\mathbf{w}\cdot\mathbf{x}$,

$z_2 = (\mathbf{w}^*)^{\perp_{\mathbf{w}}}/\|(\mathbf{w}^*)^{\perp_{\mathbf{w}}}\|_2 \cdot \mathbf{x}$ which are independent Gaussian random variables, we get

$$\begin{aligned}\mathbf{E}[\sigma(\mathbf{w}^* \cdot \mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x})\mathbf{x}^{\perp_{\mathbf{w}}} \cdot \mathbf{u}] &= \mathop{\mathbf{E}}_{z_1,z_2\sim\mathcal{N}}[\sigma(\cos(\theta)z_1 + \sin(\theta)z_2)\mathrm{T}_\rho\sigma'(z_1)\sin(\alpha)z_2] \\ &= \mathop{\mathbf{E}}_{z_1}\Big[\mathop{\mathbf{E}}_{z_2}[\sigma(\cos(\theta)z_1 + \sin(\theta)z_2)z_2 \mid z_1]\mathrm{T}_\rho\sigma'(z_1)\sin(\alpha)\Big] \\ &= \mathop{\mathbf{E}}_{z_1,z_2\sim\mathcal{N}}[\sigma'(\cos(\theta)z_1 + \sin(\theta)z_2)\mathrm{T}_\rho\sigma'(z_1)\sin(\alpha)\sin(\theta)],\end{aligned}$$

where in the last inequality we applied Fact 7.1.4. Moreover, recalling the definition of the Ornstein–Uhlenbeck semigroup, we further have

$$\begin{aligned}\mathbf{E}[\sigma(\mathbf{w}^* \cdot \mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x})\mathbf{x}^{\perp_{\mathbf{w}}} \cdot \mathbf{u}] &= \mathop{\mathbf{E}}_{z_1}[\mathop{\mathbf{E}}_{z_2}[\sigma'(\cos\theta z_1 + \sin\theta z_2) \mid z_1]\mathrm{T}_\rho\sigma'(z_1)\sin\alpha\sin\theta] \\ &\le \mathop{\mathbf{E}}_{z_1\sim\mathcal{N}}[\mathrm{T}_{\cos\theta}\sigma'(z_1)\mathrm{T}_\rho\sigma'(z_1)]\sin\theta,\end{aligned}$$

where the last inequality holds since $\mathbf{E}_{z_1\sim\mathcal{N}}[\mathrm{T}_{\cos\theta}\sigma'(z_1)\mathrm{T}_\rho\sigma'(z_1)] = \|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \ge 0$. Plugging in these bounds on the first and second terms of Equation (7.16), we get

$$\|\mathbf{g}(\mathbf{w})\|_2 \le \sqrt{\mathrm{OPT}}\sqrt{\mathbf{E}[(\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x}))^2]} + \mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w} \cdot \mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x})]\sin\theta.$$

As we have argued in the proof of Proposition 7.2.3, if $\rho \le \cos\theta$, then $\mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w} \cdot \mathbf{x})\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x})] \ge \mathbf{E}[(\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x}))^2]$, hence we further get that if in addition it holds

$$\sin\theta \ge 4\sqrt{\mathrm{OPT}}/\sqrt{\mathbf{E}[(\mathrm{T}_\rho\sigma'(\mathbf{w} \cdot \mathbf{x}))^2]},$$

then we obtain

$$\begin{aligned}\|\mathbf{g}(\mathbf{w})\|_2 &\leq (\mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathrm{T}_{\rho}\sigma'(\mathbf{w}\cdot\mathbf{x})] + (1/4)\,\mathbf{E}[(\mathrm{T}_{\rho}\sigma'(\mathbf{w}\cdot\mathbf{x}))^2])\sin\theta \\ &\leq (5/4)\,\mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathrm{T}_{\rho}\sigma'(\mathbf{w}\cdot\mathbf{x})]\sin\theta,\end{aligned}$$

completing the proof. □

We now proceed to determine the sample complexities required to estimate the gradient. Lemma 7.3.4 provides the sample complexity to approximate the norm of the population gradient $\|\mathbf{g}(\mathbf{w})\|_2$ and the inner product between the population gradient and $\mathbf{w}^*$.

**Lemma 7.3.8.** *Suppose $\sigma$ is a $(B, L)$-Regular activation. If $0 < \rho \leq \cos\theta < 1$ and $\sin\theta \geq 4(\sqrt{\mathrm{OPT}})/\|\mathrm{T}_{\rho}\sigma'\|_{L_2}$, then using*

$$n = \Theta\left(\frac{dB^2}{\sin^2\theta\|\mathrm{T}_{\rho}\sigma'\|_{L_2}^2\delta}\right)$$

*samples, with probability at least $1-\delta$, we have*

$$\begin{aligned}\|\widehat{\mathbf{g}}(\mathbf{w})\|_2 &\leq (3/2)\,\mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathrm{T}_{\rho}\sigma'(\mathbf{w}\cdot\mathbf{x})]\sin\theta\,, \\ \widehat{\mathbf{g}}(\mathbf{w})\cdot\mathbf{w}^* &\leq -\frac{1}{2}\,\mathbf{E}[\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathrm{T}_{\rho}\sigma'(\mathbf{w}\cdot\mathbf{x})]\sin^2\theta.\end{aligned}$$

*Proof.* Observe first that, by Chebyshev inequality,

$$\mathbf{Pr}[\|\widehat{\mathbf{g}}(\mathbf{w}) - \mathbf{g}(\mathbf{w})\|_2 \geq t] \leq \frac{\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\|\mathbf{g}(\mathbf{w};\mathbf{x},y) - \mathbf{g}(\mathbf{w})\|_2^2]}{nt^2}. \tag{7.17}$$

Now we proceed to bound the variance $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\|\mathbf{g}(\mathbf{w};\mathbf{x},y) - \mathbf{g}(\mathbf{w})\|_2^2]$. Let $\mathbf{e}_1, \ldots, \mathbf{e}_d$

be the standard basis of $\mathbb{R}^d$; we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|\mathbf{g}(\mathbf{w};\mathbf{x},y)-\mathbf{g}(\mathbf{w})\|_2^2] \le \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\|\mathbf{g}(\mathbf{w};\mathbf{x},y)\|_2^2] &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\sum_{j=1}^d(\mathbf{g}(\mathbf{w};\mathbf{x},y)\cdot\mathbf{e}_j)^2\right] \\ &= \sum_{j=1}^d \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y\mathrm{T}_\rho\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}\cdot(\mathbf{e}_j)^{\perp_\mathbf{w}})^2] \\ &\le dB^2\|\mathrm{T}_\rho\sigma'\|_{L_2}^2\,\mathbf{E}[(\mathbf{x}\cdot(\mathbf{e}_1)^{\perp_\mathbf{w}})^2] \le dB^2\|\mathrm{T}_\rho\sigma'\|_{L_2}^2.\end{aligned}$$

In the second inequality above, we used $|y|\le B$, which is w.l.o.g., as shown in Claim 7.5.6. Therefore, plugging the upper bound on the variance back into Equation (7.17), we get

$$\mathbf{Pr}[\|\widehat{\mathbf{g}}(\mathbf{w})-\mathbf{g}(\mathbf{w})\|_2\ge t]\le\frac{dB^2\|\mathrm{T}_\rho\sigma'\|_{L_2}^2}{nt^2}.$$

Now choosing $t=\frac{1}{6}\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2\sin\theta$ and setting

$$n=\Theta\left(\frac{dB^2\|\mathrm{T}_\rho\sigma'\|_{L_2}^2}{\sin^2\theta\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^4\delta}\right),$$

we obtain that with probability at least $1-\delta$, it holds

$$\|\widehat{\mathbf{g}}(\mathbf{w})-\mathbf{g}(\mathbf{w})\|_2\le\frac{1}{6}\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2\sin\theta\,, \tag{7.18}$$

and hence

$$\|\widehat{\mathbf{g}}(\mathbf{w})\|_2\le\|\mathbf{g}(\mathbf{w})\|_2+(1/6)\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2\sin\theta.$$

Applying the upper bound on $\|\mathbf{g}(\mathbf{w})\|_2$ we have provided in Lemma 7.3.7, we obtain

$$\|\widehat{\mathbf{g}}(\mathbf{w})\|_2\le\sqrt{\mathrm{OPT}}\|\mathrm{T}_\rho\sigma'\|_{L_2}+(7/6)\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2\sin\theta.$$

In particular, if $0 < \rho \leq \cos\theta < 1$ and $\sin\theta \geq 4\sqrt{\mathrm{OPT}}/\|\mathrm{T}_\rho \sigma'\|_{L_2}$, then we have

$$\|\widehat{\mathbf{g}}(\mathbf{w})\|_2 \leq (3/2)\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin\theta.$$

Since $0 < \rho \leq \cos\theta < 1$, it must be $\|\mathrm{T}_\rho \sigma'\|_{L_2} \leq \|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}$, and thus using

$$n = \Theta\left(\frac{dB^2}{\sin^2\theta \|\mathrm{T}_\rho \sigma'\|_{L_2}^2 \delta}\right) \tag{7.19}$$

samples suffices to guarantee that $\|\widehat{\mathbf{g}}(\mathbf{w})\|_2 \leq (3/2)\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin\theta$.

For the inner product between $\widehat{\mathbf{g}}(\mathbf{w})$ and $\mathbf{w}^*$, let us denote $\mathbf{v} = (\mathbf{w}^*)^{\perp_\mathbf{w}}/\|(\mathbf{w}^*)^{\perp_\mathbf{w}}\|_2$, and $\mathbf{w}^* = \sin\theta\mathbf{v} + \cos\theta\mathbf{w}$. Then, since $\widehat{\mathbf{g}}(\mathbf{w})$ is orthogonal to $\mathbf{w}$, we have $\widehat{\mathbf{g}}(\mathbf{w}) \cdot \mathbf{w}^* = \widehat{\mathbf{g}}(\mathbf{w}) \cdot \mathbf{v}\sin\theta$. Therefore, using Equation (7.18), we obtain that when the batch size $n$ satisfies Equation (7.19), with probability at least $1 - \delta$, we have

$$\begin{aligned}
\widehat{\mathbf{g}}(\mathbf{w}) \cdot \mathbf{w}^* &= (\widehat{\mathbf{g}}(\mathbf{w}) - \mathbf{g}(\mathbf{w})) \cdot \mathbf{v}\sin\theta + \mathbf{g}(\mathbf{w}) \cdot \mathbf{v}\sin\theta \\
&\leq \|\widehat{\mathbf{g}}(\mathbf{w}) - \mathbf{g}(\mathbf{w})\|_2 \sin\theta + \mathbf{g}(\mathbf{w}) \cdot \mathbf{w}^* \\
&\leq (1/6)\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin^2\theta + \mathbf{g}(\mathbf{w}) \cdot \mathbf{w}^*.
\end{aligned} \tag{7.20}$$

Now applying Proposition 7.2.3 we get

$$\widehat{\mathbf{g}}(\mathbf{w}) \cdot \mathbf{w}^* \leq -\frac{5}{6}\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin^2\theta + \sqrt{\mathrm{OPT}}\|\mathrm{T}_\rho \sigma'\|_{L_2} \sin\theta.$$

In particular, when $\sin\theta \geq 3\sqrt{\mathrm{OPT}}/\|\mathrm{T}_\rho \sigma'\|_{L_2}$, in Proposition 7.2.3 we showed that

$$\mathbf{g}(\mathbf{w}) \cdot \mathbf{w}^* \leq -(2/3)\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin^2\theta.$$

Thus, when $\sin\theta \geq 3\sqrt{\mathrm{OPT}}/\|\mathrm{T}_\rho \sigma'\|_{L_2}$, using Equation (7.20) we have that with probability at least $1-\delta$,

$$\widehat{\mathbf{g}}(\mathbf{w}) \cdot \mathbf{w}^* \leq -\frac{1}{2}\|\mathrm{T}_{\sqrt{\rho\cos\theta}}\sigma'\|_{L_2}^2 \sin^2\theta,$$

completing the proof. □

## 7.4 SGD – VA Efficiently Learns Monotone GLMs

We have shown in Section 7.3 that Algorithm 14 converges to a parameter vector $\mathbf{w}$ with an $L_2^2$ error bounded above by $O(\mathrm{OPT}) + \|\mathrm{P}_{>1/(\theta^*)^2}\sigma\|_{L_2}^2$, where $\theta^*$ is a Critical Point. One of the technical difficulties is that in general we cannot bound $\|\mathrm{P}_{>1/(\theta^*)^2}\sigma\|_{L_2}^2$ by OPT. One such example is when $\sigma(t) = \mathrm{He}_{(1/(\theta^*)^2+1)}(t)$; in this case $\|\mathrm{P}_{>1/(\theta^*)^2}\sigma\|_{L_2}^2 = \|\sigma\|_{L_2}^2$, which can be $\omega(\mathrm{OPT})$. In this section, we show that if the activation is also monotone, then given that $\theta^*$ is sufficiently small, we can bound $\|\mathrm{P}_{>1/(\theta^*)^2}\sigma\|_{L_2}^2$ by the Ornstein–Uhlenbeck semigroup of $\sigma'$. Specifically, we provide an initialization method that along with Algorithm 14 gives an algorithm that guarantees error $O(\mathrm{OPT})$. Formally, we show the following.

**Theorem 7.4.1** (Learning Monotone $(B, L)$-Regular Activations)**.** *Let $\epsilon > 0$, and let $\sigma$ be a monotone $(B, L)$-Regular activation. Then, Algorithm 14 draws $N = \tilde{\Theta}(dB^2 \log(L/\epsilon)/\epsilon + d/\epsilon^2)$ samples, runs in $\mathrm{poly}(d, N)$ time and returns a vector $\widehat{\mathbf{w}}$ such that with probability at least $2/3$, $\widehat{\mathbf{w}} \in \mathcal{R}_{\sigma,\theta_0}(O(\mathrm{OPT}) + \epsilon)$, and it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}) - y)^2] \leq C\mathrm{OPT} + \epsilon$, where $C$ is an absolute constant independent of $\epsilon, d, B, L$.*

The main result of this section is an initialization routine that allows us to bound the higher coefficients of the spectrum, $\|\mathrm{P}_{>1/(\theta^*)^2}\sigma\|_{L_2}^2$. In particular, we prove the following.

**Proposition 7.4.2** (Initialization). *Let $\sigma : \mathbb{R} \to \mathbb{R}$, $\sigma \in L_2(\mathcal{N})$, be a monotone $(B, L)$-Regular activation. Let $\mathcal{D}$ be a distribution of labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such that $\mathcal{D}_{\mathbf{x}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Fix a unit vector $\mathbf{w}^* \in \mathbb{R}^d$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2] = \mathrm{OPT}$. There exists an algorithm that draws $N = \widetilde{O}(d/\epsilon^2)$ samples, runs in $\mathrm{poly}(N, d)$ time, and with probability at least $2/3$, returns a unit vector $\mathbf{w}^{(0)} \in \mathbb{R}^d$ such that for any unit $\mathbf{w}' \in \mathbb{R}^d$ with $\theta = \theta(\mathbf{w}', \mathbf{w}^*) \leq \theta(\mathbf{w}^{(0)}, \mathbf{w}^*)$, it holds that*

$$\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \lesssim \sin^2\theta \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 .$$

Combining Theorem 7.3.2 with Proposition 7.4.2, we can the prove Theorem 7.4.1.

*Proof of Theorem 7.4.1.* Theorem 7.3.2 implies that Algorithm 14 generates a vector $\widehat{\mathbf{w}} \in \mathcal{R}_{\sigma,\theta_0}(C\mathrm{OPT} + \epsilon)$ where $C$ is an absolute constant. This implies that $\theta^2\|\mathrm{T}_{\cos(\theta)}\sigma'\|_{L_2}^2 \leq C\mathrm{OPT} + \epsilon$. Since $\theta(\widehat{\mathbf{w}}, \mathbf{w}^*) \leq \theta_0$, combining with Proposition 7.4.2, i.e., $\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \lesssim \sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$, we further have $\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \lesssim \mathrm{OPT} + \epsilon$. Finally, using the error bound on $\mathcal{L}(\widehat{\mathbf{w}})$ developed in Proposition 7.2.7, we get $\mathcal{L}(\widehat{\mathbf{w}}) \leq C\mathrm{OPT} + \epsilon$.

As displayed in Theorem 7.3.2, the main algorithm uses $N_1 = \tilde{\Theta}(dB^2/\epsilon + B^2/\epsilon)$ samples (since according to Lemma 7.3.6, when $\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \lesssim \sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$, using $m = \tilde{\Theta}(B^2/\epsilon)$ samples suffices for testing Algorithm 15), and in Proposition 7.4.2 we showed that the initialization procedure requires $\tilde{\Theta}(d/\epsilon^2)$ samples. Thus, in summary, for monotone $(B, L)$-Regular activations, Algorithm 14 uses $N = \tilde{\Theta}(dB^2/\epsilon + d/\epsilon^2)$ samples and runs in $\mathrm{poly}(d, N)$ times. □

For monotone $b$-Lipschitz activations $\sigma$, we know from Lemma 7.5.12 that $\sigma$ is an $\epsilon$-Extended $(b\log^{1/2}(b/\epsilon), b)$-Regular activation, meaning that there exists a truncated activation $\bar{\sigma}$ that such that $\mathbf{E}_{z\sim\mathcal{N}}[(\bar{\sigma}(z) - \sigma(z))^2] \leq \epsilon$ and $\bar{\sigma}$ is $(b\log^{1/2}(b/\epsilon), b)$-Regular.

Hence applying Theorem 7.4.1 to $\bar{\sigma}$, we obtain the following corollary:

**Corollary 7.4.3** (Learning Monotone & Lipschitz Activations)**.** *Let $\epsilon, b > 0$, and let $\sigma$ be a monotone $b$-Lipschitz activation. Then, Algorithm 14 draws $N = \tilde{\Theta}(db^2/\epsilon + d/\epsilon^2)$ samples, runs in $\mathrm{poly}(d, N)$ time, and returns a vector $\widehat{\mathbf{w}}$ such that with probability at least $2/3$, it holds that $\mathcal{L}(\widehat{\mathbf{w}}) \leq C\mathrm{OPT} + \epsilon$, where $C$ is an absolute constant independent of $\epsilon, d, b$.*

Similarly, if $\sigma$ has bounded $2 + \zeta$ moment $\mathbf{E}_{z\sim\mathcal{N}}[\sigma^{2+\zeta}(z)] \leq B_\sigma$, then according to Lemma 7.5.9 we know that $\sigma$ is an $\epsilon$-Extended $((B_\sigma/\epsilon)^{1/\zeta}, (B_\sigma/\epsilon)^{4/\zeta}/\epsilon^2)$-Regular activation. Therefore, replacing $B$ with $(B_\sigma/\epsilon)^{1/\zeta}$ and replace $L$ with $(B_\sigma/\epsilon)^{4/\zeta}/\epsilon^2$ in Theorem 7.4.1, we obtain:

**Corollary 7.4.4** (Learning Monotone Activations With Bounded $2 + \zeta$ Moments)**.** *Let $\epsilon > 0$, and let $\sigma$ be a monotone activation that satisfies $\mathbf{E}_{z\sim\mathcal{N}}[\sigma^{2+\zeta}(z)] \leq B_\sigma$. Then, Algorithm 14 draws $N = \tilde{\Theta}(d(B_\sigma/\epsilon)^{2/\zeta}\log(B_\sigma/\epsilon)/\epsilon + d/\epsilon^2)$ samples, runs in $\mathrm{poly}(d, N)$ time, and returns a vector $\widehat{\mathbf{w}}$ such that with probability at least $2/3$, it holds that $\mathcal{L}(\widehat{\mathbf{w}}) \leq C\mathrm{OPT} + \epsilon$, where $C$ is an absolute constant independent of $\epsilon, d, B_\sigma, L$.*

The main contents of this section are the following: To prove Proposition 7.4.2, we need to combine two main technical pieces: (1) proving that there exists a threshold $\theta_0$ such that for any $\theta \leq \theta_0$, $\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \lesssim \sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$; (2) proving that there exists an efficient algorithm that finds a parameter $\mathbf{w}^{(0)}$ such that $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \leq \theta_0$.

Section (7.4.1) is devoted to the proof of (1), i.e., that there exists $\theta_0$ such that for $\theta \leq \theta_0$, $\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \lesssim \sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$ (Proposition 7.4.6). Unfortunately, it was technically hard to prove this claim directly for all monotone functions due to the versatility of such functions. Hence, the natural idea is that if we can prove (1) for

a sequence of simple and 'nice' functions $\Phi_k$ that can converge to $\sigma$, then by the convergence theorems the desired claim will also hold true for $\sigma$. In particular, let $\Phi_k$ be a sequence of functions; then, one can show that the higher order coefficients can be bounded by

$$\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \leq 2\|\sigma - \Phi_k\|_{L_2}^2 + 4\|\Phi_k - \mathrm{T}_\rho\Phi_k\|_{L_2}^2 + 4\theta^2\|\mathrm{T}_\rho\Phi_k'\|_{L_2}^2.$$

If $\Phi_k$ converges to $\sigma$ pointwise, one can show that the first term above goes to 0 and the second term converges to $\theta^2\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$, which is the bound we are looking for. Thus, it remains to show that $\|\Phi_k - \mathrm{T}_\rho\Phi_k\|_{L_2}^2 \lesssim \theta^2\|\mathrm{T}_\rho\Phi_k'\|_{L_2}^2$.

Section (7.4.2) proves the claim that $\|\Phi - \mathrm{T}_\rho\Phi\|_{L_2}^2 \lesssim \theta^2\|\mathrm{T}_\rho\Phi'\|_{L_2}^2$ (Proposition 7.4.8), for any $\Phi(z)$ that is a *monotonic staircase function*:

**Definition 7.4.5** (Monotonic Staircase Functions)**.** *For simplicity, denote the indicator function $\mathbb{1}\{z \geq t\}$ by $\phi(z;t)$. Let $m$ be a positive integer and let $M > 0$. The monotonic staircase functions (of $M$-bounded support) are defined by*

$$\mathcal{F}_M \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^{m} A_i\phi(z;t_i) + A_0 : A_0 \in \mathbb{R}; A_i > 0, |t_i| \leq M, \forall i \in [m]; m < \infty \right\}.$$

These staircase functions constitute a dense subset of the monotone function class and have a simple and easy-to-analyze form, therefore they serve well for our purpose. However, though the staircase function $\Phi$ already takes a concise and simple expression, many technical difficulties arise when analyzing $\mathrm{T}_\rho\Phi(z) - \Phi(z)$, mainly due to the complicated form of $\mathrm{T}_\rho\Phi(z)$. Our workaround is to introduce a new type of smoothing/augmentation method, which we call centered augmentation, defined by $\mathrm{T}_\rho(\Phi(z/\rho))$. This recentered augmentation takes a much simpler form compared to $\mathrm{T}_\rho\Phi(z)$. In par-

ticular, we show that when the smoothing parameter $\rho$ is not too small, namely, when $1-\rho^2 \le O(1/\log(1/\epsilon))$, then: (i) the $L_2^2$ distance between $\mathrm{T}_\rho\Phi(z/\rho)$ and $\Phi(z)$ can be bounded above by $(1-\rho^2)\|\mathrm{T}_\rho\Phi'(z/\rho)\|_{L_2}^2$ (Lemma 7.4.12); (ii) the $L_2^2$ distance between $\mathrm{T}_\rho\Phi(z/\rho)$ and $\mathrm{T}_\rho\Phi(z)$ can be bounded above by $(1-\rho^2)(\|\mathrm{T}_\rho\Phi'(z/\rho)\|_{L_2}^2 + \|\mathrm{T}_\rho\Phi'(z)\|_{L_2}^2)$ (Lemma 7.4.13); (iii) finally, choosing the smoothing strength $\rho_1$ slightly larger than $\rho$, we have $\|\mathrm{T}_{\rho_1}\Phi'(z/\rho_1)\|_{L_2}^2 \lesssim \|\mathrm{T}_\rho\Phi'(z)\|_{L_2}^2$ (Lemma 7.4.15). Combining these 3 results on the relations between $\mathrm{T}_\rho\Phi(z/\rho)$ and $\mathrm{T}_\rho\Phi(z)$, we prove Proposition 7.4.8 in Section (7.4.2), completing the last piece of the puzzle in the proof of Proposition 7.4.6.

Finally, in Section (7.4.3), we prove (2) by providing an *SQ* initialization algorithm. The main idea is to transform the labels $y$ to $\tilde{y} = \mathcal{T}(y) \stackrel{\text{def}}{=} \mathbb{1}\{y \ge t'\}$ for a carefully chosen threshold $t'$. Then, we show that there exists a halfspace $\phi(\mathbf{w}^* \cdot \mathbf{x}; t) = \mathbb{1}\{\mathbf{w}^* \cdot \mathbf{x} \ge t\}$ such that the transformed labels $\tilde{y}$ can be viewed as the corrupted labels of $\phi(\mathbf{w}^* \cdot \mathbf{x}; t)$. Then, utilizing the algorithm for learning halfspaces [DKTZ22b], we can obtain an initial vector $\mathbf{w}^{(0)}$ such that $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \le \theta_0$.

Finally, combining (1) and (2), we prove Proposition 7.4.2.

### 7.4.1 Bounding Higher Order Hermite Coefficients of Monotone Activations

The main result of this section is the following:

**Proposition 7.4.6** (From Hermite Tails to Ornstein–Uhlenbeck Semigroup)**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a monotone activation and $\sigma \in L_2(\mathcal{N})$. Let $M$ be the upper bound for the support of $\sigma'(z)$,[1] i.e., $\forall z \in \mathbb{R}$ such that $|z| \ge M$, we have $\sigma'(z) = 0$. For any*

[1] In Claim 7.5.7, we show that for any $\sigma \in \mathcal{H}(B, L)$, the support of $\sigma'$ can always be bounded by $M \lesssim \sqrt{\log(B/\epsilon) - \log\log(B/\epsilon)}$.

$\theta \in [0, \pi]$ *such that* $1 - C/M^2 < \cos^2\theta$ *with* $C > 0$ *an absolute constant, it holds that* $\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \lesssim \sin^2\theta \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$.

*Proof.* Instead of proving Proposition 7.4.6 directly for the activation $\sigma$, we chose another function $\Phi$ that works as a surrogate for $\sigma$ and satisfies certain regularity properties. Let $\Phi$ be any function in $L_2(\mathcal{N})$, then by Young's inequality we have that

$$\begin{aligned}\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 &\leq 2\|\mathrm{P}_{>1/\theta^2}(\sigma - \Phi)\|_{L_2}^2 + 2\|\mathrm{P}_{>1/\theta^2}\Phi\|_{L_2}^2 \\ &\leq 2\|\mathrm{P}_{>1/\theta^2}(\sigma - \Phi)\|_{L_2}^2 + 4\|\mathrm{P}_{>1/\theta^2}(\Phi - \mathrm{T}_\rho\Phi)\|_{L_2}^2 + 4\|\mathrm{P}_{>1/\theta^2}\mathrm{T}_\rho\Phi\|_{L_2}^2 \,.\end{aligned}$$

Observe that $\mathrm{P}_{>m}$ is a non-expansive operator since for any $f \in L_2(\mathcal{N})$, $f(z) \doteq \sum_{i\geq 0} a_i \mathrm{He}_i(z)$ it holds

$$\|\mathrm{P}_{>m}f\|_{L_2}^2 = \sum_{i>m} a_i^2 \leq \sum_{i\geq 0} a_i^2 = \|f\|_{L_2}^2.$$

Therefore, $\|\mathrm{P}_{>1/\theta^2}(\sigma - \Phi)\|_{L_2}^2 \leq \|\sigma - \Phi\|_{L_2}^2$. In addition, note that we have the following inequality for any $f, f' \in L_2(\mathcal{N})$:

$$\|\mathrm{P}_{>m}f\|_2^2 = \sum_{i>m} a_i^2 \leq \sum_{i>m}(i/m)a_i^2 \leq \sum_{i=1}^{m}(i/m)a_i^2 + \sum_{i>m}(i/m)a_i^2 = (1/m)\|f'\|_{L_2}^2,$$

therefore $\|\mathrm{P}_{>1/\theta^2}\mathrm{T}_\rho\Phi\|_{L_2}^2 \leq \theta^2\|(\mathrm{T}_\rho\Phi)'\|_{L_2}^2$. Finally by Fact 7.1.7 we have $\|(\mathrm{T}_\rho\Phi)'\|_{L_2}^2 = \|\rho\mathrm{T}_\rho\Phi'\|_{L_2}^2 \leq \|\mathrm{T}_\rho\Phi'\|_{L_2}^2$ since $\rho < 1$, thus, it holds

$$\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \leq 2\|\sigma - \Phi\|_{L_2}^2 + 4\|\Phi - \mathrm{T}_\rho\Phi\|_{L_2}^2 + 4\theta^2\|\mathrm{T}_\rho\Phi'\|_{L_2}^2 \,. \tag{7.21}$$

Let $\Phi_k$ be any sequence of functions such that $\lim_{k\to\infty} \|\Phi_k - \sigma\|_{L_2} = 0$. For this sequence we have that Equation (7.21) becomes

$$\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \le 2\|\sigma - \Phi_k\|_{L_2}^2 + 4\|\Phi_k - \mathrm{T}_\rho\Phi_k\|_{L_2}^2 + 4\theta^2\|\mathrm{T}_\rho\Phi_k'\|_{L_2}^2 \,. \tag{7.22}$$

In particular, let $\Phi_k$ be a sequence of staircase monotonic functions (see Definition 7.4.5) that converges to $\sigma$ uniformly; then, for $\rho^2 \ge 1 - C/M^2$ where $M$ is the upper bound on the support of $\sigma'$ (which is also the upper bound on the support of all $\Phi_k'$'s) and $C$ is an absolute constant, from Proposition 7.4.8, we conclude that $\|\Phi_k - \mathrm{T}_\rho\Phi_k\|_{L_2}^2 \lesssim (1-\rho^2)\|\mathrm{T}_\rho\Phi_k'\|_{L_2}$ and therefore we have that

$$\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \le 2\|\sigma - \Phi_k\|_{L_2}^2 + 4((1-\rho^2) + \theta^2)\|\mathrm{T}_\rho\Phi_k'\|_{L_2}^2 \,. \tag{7.23}$$

Our next goal is to show that the sequence of smoothed derivatives $\mathrm{T}_\rho\Phi_k'$ also converge to $\sigma'$, as stated in the following lemma.

**Lemma 7.4.7** (Convergence of Derivatives)**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$, $\sigma \in L_2(\mathcal{N})$, and let $\Phi_k : \mathbb{R} \to \mathbb{R}$ be a sequence of functions such that $\|\sigma - \Phi_k\|_{L_2} \to 0$ as $k \to \infty$. Then, for any $\rho \in (0,1)$, it holds that*

$$\|\mathrm{T}_\rho\Phi_k' - \mathrm{T}_\rho\sigma'\|_{L_2} \to 0, \quad \text{as } k \to \infty.$$

*Proof.* For any function $f \in L_2(\mathcal{N})$, we have that (by the definition of Ornstein–Uhlenbeck semigroup and Stein's lemma, stated in Fact 7.1.4)

$$\mathrm{T}_\rho f'(z) = \frac{1}{\sqrt{1-\rho^2}} \mathop{\mathbf{E}}_{t\sim\mathcal{N}}[f(\rho z + \sqrt{(1-\rho^2)}t)t] \,.$$

Therefore, we have that

$$
\begin{aligned}
&\|\mathrm{T}_\rho \Phi_k' - \mathrm{T}_\rho \sigma'\|_{L_2}^2 \\
&= \frac{1}{1-\rho^2} \mathop{\mathbf{E}}_{z\sim\mathcal{N}} \left[ \left( \mathop{\mathbf{E}}_{t\sim\mathcal{N}} \left[ \left( \Phi_k(\rho z + \sqrt{(1-\rho^2)}t) - \sigma(\rho z + \sqrt{(1-\rho^2)}t) \right) t \right] \right)^2 \right] \\
&\leq \frac{1}{1-\rho^2} \mathop{\mathbf{E}}_{z\sim\mathcal{N}} \left[ \mathop{\mathbf{E}}_{t\sim\mathcal{N}} \left[ \left( \Phi_k(\rho z + \sqrt{(1-\rho^2)}t) - \sigma(\rho z + \sqrt{(1-\rho^2)}t) \right)^2 \right] \mathop{\mathbf{E}}_{t\sim\mathcal{N}} \left[ t^2 \right] \right] \\
&= \frac{1}{1-\rho^2} \mathop{\mathbf{E}}_{z\sim\mathcal{N}} \left[ \mathop{\mathbf{E}}_{t\sim\mathcal{N}} \left[ \left( \Phi_k(\rho z + \sqrt{(1-\rho^2)}t) - \sigma(\rho z + \sqrt{(1-\rho^2)}t) \right)^2 \right] \right] \\
&= \frac{1}{1-\rho^2} \mathop{\mathbf{E}}_{z\sim\mathcal{N}} \left[ \left( \Phi_k(z) - \sigma(z) \right)^2 \right] ,
\end{aligned}
$$

where the inequality is by the Cauchy-Schwarz inequality and the last inequality is by $\rho z + \sqrt{(1-\rho^2)}t \sim \mathcal{N}(0,1)$ for independent $z \sim \mathcal{N}(0,1)$, $t \sim \mathcal{N}(0,1)$. In remains to take the limit with $k \to \infty$. □

Combining Lemma 7.4.7 with Equation (7.23), and letting $\rho = \cos\theta$ now completes the proof of Proposition 7.4.6. □

We recall that the assumption that the support of $\sigma'(z)$ is bounded by $M < +\infty$ is without loss of generality, as we have proved in Claim 7.5.7.

### 7.4.2 Bounding the Augmentation Error

In this subsection, we prove the main technical result, which provides an upper bound on the smoothing error of piecewise staircase functions using the $L_2(\mathcal{N})$ norm of the smoothed derivative. We recall the class of the piecewise staircase functions $\mathcal{F}_M$ below:

$$
\mathcal{F}_M \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^{m} A_i \phi(z; t_i) + A_0 : A_0 \in \mathbb{R}; A_i > 0, |t_i| \leq M, \forall i \in [m]; m < \infty \right\} .
$$

Our result is the following proposition:

**Proposition 7.4.8.** *Let $\Phi \in \mathcal{F}_M$ be any staircase function that is consists of $m$ indicator functions with thresholds $t_i$, $i \in [m]$, and suppose $|t_i| \leq M$ for all $i \in [m]$, where $1 < M < +\infty$. For any $\rho \in (0,1)$ such that $\rho^2 \geq 1 - C/M^2$ where $C < M^2/4$ is an absolute constant, we have*

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z) - \Phi(z))^2] \lesssim (1-\rho^2) \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi'(z))^2].$$

As we have remarked in the comment after the proof of Lemma 7.5.12, since $M$ is an upper bound on the support of $\sigma'$, we will assume without loss of generality throughout the rest of the work that $M^2$ is larger than constant $4C$.

Some remarks about the staircase functions are in order. Observe first that according to Claim 7.5.7, when $\sigma$ is $(B,L)$-Regular, we can always bound $M$ by $\sqrt{2\log(4B^2/\epsilon) - \log\log(4B^2/\epsilon)}$. Next, for any function $\Phi \in \mathcal{F}_M$, its derivative can be written as:

$$\Phi'(z) = \sum_{i=1}^m A_i \phi'(z; t_i) = \sum_{i=1}^m A_i \delta(z - t_i),$$

where $\delta(z - t_i)$ is the Dirac delta function. Certainly, when $|z| \geq M$ we have $\Phi'(z) = 0$. Also note that for any non-decreasing function $\sigma$ with the support of its derivative $\sigma'(z)$ bounded by $M$, there exists a sequence of staircase functions $\Phi_k \in \mathcal{F}_M$ such that $\Phi_k$ converges to $\sigma$ uniformly. To prove this claim, we note that since for any $|z| \geq M$, $\sigma'(z) = 0$, therefore $\sigma(z) = \sigma(M)$ when $z \geq M$ and $\sigma(z) = \sigma(-M)$ for all $z \leq -M$.

Hence, let

$$\Phi_k(z) = \sum_{i=1}^{m} \frac{1}{k}\phi(z; t_i) + \sigma(-M),$$

$$\text{where } \begin{cases} m = \lceil \sigma(M) - \sigma(-M)/k \rceil + 1, \\ t_i = \min_{t\in[-M,M]}\{\sigma(t) \geq (i-1)(1/k) + \sigma(-M)\},\ i = 1, \ldots, m-1;\ t_m = M. \end{cases}$$

By construction, we have $|\Phi_k(z) - \sigma(z)| \leq 1/k$ for all $z \in \mathbb{R}$, therefore $\Phi_k$ converges to $\sigma$ uniformly.

To prove Proposition 7.4.8, we decompose $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z) - \Phi(z))^2]$ into the following terms and provide upper bounds on each term respectively:

$$\begin{aligned} &\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z) - \Phi(z))^2] \leq \\ &2\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z) - \mathrm{T}_{\rho_1}\Phi(z/\rho_1))^2] + 2\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi(z/\rho_1) - \Phi(z))^2] \lesssim \\ &\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z) - \mathrm{T}_{\rho_1}\Phi(z))^2] + \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi(z) - \mathrm{T}_{\rho_1}\Phi(z/\rho_1))^2] + \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi(z/\rho_1) - \Phi(z))^2], \end{aligned} \tag{7.24}$$

where we repeatedly used the inequality $(a+b)^2 \leq 2a^2 + 2b^2$. As we have discussed at the beginning of Section 7.4, we introduced this 'recentered smoothing' operator $\mathrm{T}_\rho\Phi(z/\rho)$ to overcome the difficulty of analyzing $\mathrm{T}_\rho\Phi(z) - \Phi(z)$, since $\mathrm{T}_\rho\Phi(z/\rho)$ takes a more simple and easy-to-analyze form. Here, $\rho_1 \in (0,1)$ is a carefully chosen smoothing parameter that is slightly larger than $\rho$, so that we can bound $\|\mathrm{T}_{\rho_1}\Phi'(z/\rho_1)\|_{L_2}^2$ above using $\|\mathrm{T}_\rho\Phi'(z)\|_{L_2}^2$ (Lemma 7.4.15).

Coming back to Equation (7.24), we show that: (1) the first term $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z) - \mathrm{T}_{\rho_1}\Phi(z))^2]$ can be bounded above by $(1-\rho)\|\mathrm{T}_{\rho_1}\Phi'(z)\|_{L_2}^2$, using Lemma 7.1.10; (2) the sec-

ond term $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi(z)-\mathrm{T}_{\rho_1}\Phi(z/\rho_1))^2]$ is bounded above by $(1-\rho)(\|\mathrm{T}_{\rho_1}\Phi'(z/\rho_1)\|_{L_2}^2+\|\mathrm{T}_{\rho_1}\Phi'(z)\|_{L_2}^2)$, using Lemma 7.4.13; and (3) the third term $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi(z/\rho_1)-\Phi(z))^2]$ is bounded above by $(1-\rho)\|\mathrm{T}_{\rho_1}\Phi'(z/\rho_1)\|_{L_2}^2$, using Lemma 7.4.12.

Thus, in summary, we have $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho}\Phi(z)-\Phi(z))^2]\lesssim(1-\rho)(\|\mathrm{T}_{\rho_1}\Phi'(z/\rho_1)\|_{L_2}^2+\|\mathrm{T}_{\rho_1}\Phi'(z)\|_{L_2}^2)$. Since $\rho_1$ is chosen so that $\|\mathrm{T}_{\rho_1}\Phi'(z/\rho_1)\|_{L_2}^2\lesssim\|\mathrm{T}_{\rho}\Phi'(z)\|_{L_2}^2$ (see Lemma 7.4.15), and furthermore, since it holds that $\|\mathrm{T}_{\rho_1}\Phi'(z)\|_{L_2}^2\lesssim\|\mathrm{T}_{\rho}\Phi'(z)\|_{L_2}$, combining these results we prove that $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho}\Phi(z)-\Phi(z))^2]\lesssim(1-\rho)\|\mathrm{T}_{\rho}\Phi'(z)\|_{L_2}^2$.

We first derive an explicit expression for $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho}\Phi'(z))^2]$, for any $\Phi\in\mathcal{F}_M$.

**Lemma 7.4.9.** *For any $\Phi\in\mathcal{F}_M$, it holds that*

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho}\Phi'(z))^2]=\sum_{i,j=1}^{m}\frac{A_iA_j}{2\pi\sqrt{1-\rho^4}}\exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho^4)}+\frac{\rho^2t_it_j}{1-\rho^4}\right).$$

*Proof.* By the linearity of the Ornstein–Uhlenbeck semigroup, we have $\mathrm{T}_{\rho}\Phi'(z)=\sum_{i=1}^{m}A_i\mathrm{T}_{\rho}\phi'(z;t_i)$. In fact, each summand in this summation has an explicit expression, which we derive in the following:

$$\begin{aligned}\mathrm{T}_{\rho}\phi'(z;t_i)&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\phi'(\rho z+\sqrt{1-\rho^2}u;t_i)\exp(-u^2/2)\,\mathrm{d}u\\&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\delta(\rho z+\sqrt{1-\rho^2}u-t_i)\exp(-u^2/2)\,\mathrm{d}u\\&=\frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left(-\frac{(\rho z-t_i)^2}{2(1-\rho^2)}\right),\end{aligned}\tag{7.25}$$

where we have used that $\delta$ is the Dirac delta function, and so $\delta(u)$ satisfies $\delta(au)=\delta(u)/a$

for any real positive number $a$. Therefore, we get that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi'(z))^2] \\
&= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\sum_{i,j=1}^{m}\frac{A_iA_j}{2\pi(1-\rho^2)}\exp\left(-\frac{(\rho z-t_i)^2+(\rho z-t_j)^2}{2(1-\rho^2)}\right)\right] \\
&= \sum_{i,j=1}^{m}\frac{A_iA_j}{2\pi(1-\rho^2)}\exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho^2)}\right)\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\rho^2}{1-\rho^2}z^2+\frac{(t_i+t_j)\rho}{1-\rho^2}z-\frac{z^2}{2}\right)\mathrm{d}z \\
&\overset{(i)}{=} \sum_{i,j=1}^{m}\frac{A_iA_j}{2\pi(1-\rho^2)}\exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho^2)}\right)\sqrt{\frac{1-\rho^2}{1+\rho^2}}\exp\left(\frac{(t_i+t_j)^2\rho^2}{2(1-\rho^4)}\right) \\
&= \sum_{i,j=1}^{m}\frac{A_iA_j}{2\pi\sqrt{1-\rho^4}}\exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho^4)}+\frac{\rho^2t_it_j}{1-\rho^4}\right),
\end{aligned}$$

where in $(i)$ we used the fact that $\int\exp(-az^2+bz)\,\mathrm{d}z=\sqrt{\pi/a}\exp(b^2/4a)$. □

A byproduct of the above proof is that:

**Claim 7.4.10.** *For any $\Phi\in\mathcal{F}_M$, it holds*

$$\mathrm{T}_\rho\Phi'(z)=\sum_{k\geq0}\rho^k\alpha_k\mathrm{He}_k(z),\ \textit{where}\ \alpha_k=\sum_{i=1}^{m}\frac{A_i}{\sqrt{2\pi}}\exp(-t_i^2/2)\mathrm{He}_k(t_i).$$

*Furthermore, the function $\zeta(\rho)\overset{\text{def}}{=}\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi'(z))^2]$ is a non-decreasing function of $\rho\in(0,1)$.*

*Proof.* It is easy to see that $\mathrm{T}_\rho\phi'(z;t_i)$ is square-integrable under the Gaussian measure, therefore the Hermite expansion of $\mathrm{T}_\rho\phi'(z;t_i)$ exists. In particular, using Mehler's formula (Fact 7.1.3), we can derive the Hermite expansion of $\mathrm{T}_\rho\phi'(z;t_i)$ immediately:

$$\mathrm{T}_\rho\phi'(z;t_i)=\frac{1}{\sqrt{2\pi}}\sum_{k\geq0}\rho^k\exp(-t_i^2/2)\mathrm{He}_k(t_i)\mathrm{He}_k(z),$$

which then implies that it holds

$$\begin{aligned}\mathrm{T}_\rho\Phi'(z) &= \sum_{i=1}^{m}\frac{A_i}{\sqrt{2\pi}}\sum_{k\geq 0}\rho^k\exp(-t_i^2/2)\mathrm{He}_k(t_i)\mathrm{He}_k(z)\\ &= \sum_{k\geq 0}\rho^k\left(\sum_{i=1}^{m}\frac{A_i}{\sqrt{2\pi}}\exp(-t_i^2/2)\mathrm{He}_k(t_i)\right)\mathrm{He}_k(z). \qquad (7.26)\end{aligned}$$

For the monotonicity of $\zeta(\rho)$, observe that by the Hermite expansion of $\mathrm{T}_\rho\Phi'(z)$, we have

$$\zeta(\rho) = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi'(z))^2] = \sum_{k\geq 0}\rho^{2k}\alpha_k^2,$$

which is an increasing function of $\rho\in(0,1)$. □

Claim 7.4.10 implies that though $\Phi'(z)$ is not in $L_2(\mathcal{N})$ (since the square of the Dirac delta function $\delta^2(z)$ is not integrable), $\mathrm{T}_\rho\Phi'(z)$ is well-defined and is continuous and smooth. Consequently, all the facts presented in Fact 7.1.7 apply to $\mathrm{T}_\rho\Phi'(z)$ as well.

Proceeding to the analysis of $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z)-\Phi(z))^2]$, however, technical difficulties arise when we try to relate $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z)-\Phi(z))^2]$ with $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi'(z))^2]$. The main obstacle is that it is hard to analyze $\mathrm{T}_\rho\phi(z;t)-\phi(z;t)$, since

$$\mathrm{T}_\rho\phi(z;t)-\phi(z;t) = \mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}[u\geq(t-\rho z)/\sqrt{1-\rho^2}] - \mathbb{1}\{z\geq t\},$$

and the probability term does not have a close form. The workaround is to study the centered augmentation (centered smoothing), and then translate the upper bound on the centered augmentation error back to the upper bound on the standard augmentation error.

**Centered Augmentation**

We define the centered augmentation as the following:

$$\mathrm{T}_\rho \sigma(z/\rho) = \mathop{\mathbf{E}}_{u\sim\mathcal{N}}[\sigma(z + (\sqrt{1-\rho^2}/\rho)u)].$$

Note that for the staircase functions $\Phi \in \mathcal{F}_M$, it holds

$$\begin{aligned}\mathrm{T}_\rho \Phi(z/\rho) &= \sum_{i=1}^m A_i \mathop{\mathbf{E}}_{u\sim\mathcal{N}}[\mathbb{1}\{z + (\sqrt{1-\rho^2}/\rho)u \geq t\}] \\ &= \sum_{i=1}^m A_i \mathop{\mathbf{E}}_{u\sim\mathcal{N}}[\mathbb{1}\{\rho z + \sqrt{1-\rho^2}u \geq \rho t_i\}] = \sum_{i=1}^m A_i \mathrm{T}_\rho \phi(z; \rho t_i).\end{aligned}$$

We first provide explicit expressions for $\mathrm{T}_\rho \Phi'(z/\rho)$ and $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi'(z/\rho))^2]$.

**Lemma 7.4.11.** *For any $\Phi(z) = \sum_{i=1}^m A_i \phi(z; t_i) + A_0 \in \mathcal{F}_M$, we have*

$$\begin{aligned}\mathrm{T}_\rho \Phi'(z/\rho) &= \sum_{i=1}^m \frac{\rho A_i}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{\rho^2(z-t_i)^2}{2(1-\rho^2)}\right), \text{ and} \\ \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi'(z/\rho))^2] &= \sum_{i,j=1}^m \frac{\rho^2 A_i A_j}{2\pi\sqrt{1-\rho^4}} \exp\left(-\frac{\rho^2(t_i^2+t_j^2)}{2(1-\rho^4)} + \frac{\rho^4 t_i t_j}{1-\rho^4}\right).\end{aligned}$$

*Proof.* The proof follows similar steps as the proof of Lemma 7.4.9. Observe first that by the definition of $\Phi(z)$, the derivative of $\Phi$ equals

$$\Phi'(z) = \sum_{i=1}^m A_i \phi'(z; t_i) = \sum_{i=1}^m A_i \delta(z - t_i),$$

where $\delta$ is the Dirac delta function. As $\delta(u)$ satisfies $\delta(au) = \delta(u)/a$ for any real positive

number $a$,

$$\Phi'(z/\rho) = \sum_{i=1}^{m} A_i \delta((z - \rho t_i)/\rho) = \sum_{i=1}^{m} \rho A_i \delta(z - \rho t_i) = \rho \sum_{i=1}^{m} A_i \phi'(z; \rho t_i).$$

This implies that

$$\mathrm{T}_\rho \Phi'(z/\rho) = \rho \sum_{i=1}^{m} A_i \mathrm{T}_\rho \phi'(z; \rho t_i),$$

which leads to the first claim in the statement after combining with Equation (7.25). The second claim now follows from Lemma 7.4.9, by replacing $t_i, t_j$ with $\rho t_i$ and $\rho t_j$. □

We now show that the centered augmentation error can be bounded above by $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi'(z/\rho))^2]$.

**Lemma 7.4.12.** *Let $\Phi \in \mathcal{F}_M$. Then, for any $\rho \in (0, 1)$,*

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi(z/\rho) - \Phi(z))^2] \leq 4((1-\rho^2)/\rho^2) \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi'(z/\rho))^2].$$

*Proof.* Observe that after augmentation, the indicator function $\mathbb{1}\{z \geq t\} = \phi(z; t)$ becomes $\mathrm{T}_\rho \phi(z/\rho; t) = \mathrm{T}_\rho \phi(z; \rho t) = \mathbf{Pr}_{u\sim\mathcal{N}}[u \geq \rho(t-z)/\sqrt{1-\rho^2}]$. Therefore, $\mathrm{T}_\rho \phi(z/\rho; t) - \phi(z; t)$ can be expressed as:

$$\mathrm{T}_\rho \phi(z/\rho; t) - \phi(z; t) = \begin{cases} \mathbf{Pr}_{u\sim\mathcal{N}}[u \geq \rho(t-z)/\sqrt{1-\rho^2}] & z < t, \\ -\mathbf{Pr}_{u\sim\mathcal{N}}[u \leq \rho(t-z)/\sqrt{1-\rho^2}] & z \geq t. \end{cases}$$

Hence, $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z/\rho)-\Phi(z))^2]$ equals:

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z/\rho)-\Phi(z))^2]\\
&=\mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\sum_{i,j=1}^{m}A_iA_j(\mathrm{T}_\rho\phi(z/\rho;t_i)-\phi(z;t_i))(\mathrm{T}_\rho\phi(z/\rho;t_j)-\phi(z;t_j))\right]\\
&=\sum_{i,j=1}^{m}\frac{A_iA_j}{\sqrt{2\pi}}\int_{-\infty}^{\min\{t_i,t_j\}}\mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}\left[u\geq\frac{\rho(t_i-z)}{\sqrt{1-\rho^2}}\right]\mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}\left[u\geq\frac{\rho(t_j-z)}{\sqrt{1-\rho^2}}\right]e^{-z^2/2}\,\mathrm{d}z-\\
&\sum_{i,j=1}^{m}\frac{A_iA_j}{\sqrt{2\pi}}\int_{\min\{t_i,t_j\}}^{\max\{t_i,t_j\}}\mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}\left[u\geq\frac{\rho(\max\{t_i,t_j\}-z)}{\sqrt{1-\rho^2}}\right]\mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}\left[u\leq\frac{\rho(\min\{t_i,t_j\}-z)}{\sqrt{1-\rho^2}}\right]e^{-z^2/2}\,\mathrm{d}z\\
&+\sum_{i,j=1}^{m}\frac{A_iA_j}{\sqrt{2\pi}}\int_{\max\{t_i,t_j\}}^{+\infty}\mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}\left[u\leq\frac{\rho(t_i-z)}{\sqrt{1-\rho^2}}\right]\mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}\left[u\leq\frac{\rho(t_j-z)}{\sqrt{1-\rho^2}}\right]e^{-z^2/2}\,\mathrm{d}z.
\end{aligned}
$$

When $z\leq\min\{t_i,t_j\}$, since both $\rho(t_i-z)$ and $\rho(t_j-z)$ are positive, by standard Gaussian concentration,

$$
\mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}[u\geq\frac{\rho(t_i-z)}{\sqrt{1-\rho^2}}]\leq\frac{1}{2}\exp(-\frac{\rho^2(t_i-z)^2}{2(1-\rho^2)}),\ \mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}[u\geq\frac{\rho(t_j-z)}{\sqrt{1-\rho^2}}]\leq\frac{1}{2}\exp(-\frac{\rho^2(t_j-z)^2}{2(1-\rho^2)}).
$$

The same inequalities hold for $\mathbf{Pr}[u\leq\rho(t_i-z)/\sqrt{1-\rho^2}]$ and $\mathbf{Pr}[u\leq\rho(t_j-z)/\sqrt{1-\rho^2}]$

when $z \geq \max\{t_i, t_j\}$. Thus, we can further upper bound $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z/\rho) - \Phi(z))^2]$ by

$$\begin{aligned}
&\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z/\rho) - \Phi(z))^2] \\
&\leq \sum_{i,j=1}^{m} \frac{A_i A_j}{\sqrt{2\pi}} \int_{-\infty}^{\min\{t_i,t_j\}} \mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}\left[u \geq \frac{\rho(t_i - z)}{\sqrt{1-\rho^2}}\right] \mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}\left[u \geq \frac{\rho(t_j - z)}{\sqrt{1-\rho^2}}\right] e^{-z^2/2}\,\mathrm{d}z \\
&\quad + \sum_{i,j=1}^{m} \frac{A_i A_j}{\sqrt{2\pi}} \int_{\max\{t_i,t_j\}}^{+\infty} \mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}\left[u \leq \frac{\rho(t_i - z)}{\sqrt{1-\rho^2}}\right] \mathop{\mathbf{Pr}}_{u\sim\mathcal{N}}\left[u \leq \frac{\rho(t_j - z)}{\sqrt{1-\rho^2}}\right] e^{-z^2/2}\,\mathrm{d}z \\
&\leq \sum_{i,j=1}^{m} \frac{1}{2}\frac{A_i A_j}{\sqrt{2\pi}} \left(\int_{-\infty}^{\min\{t_i,t_j\}} + \int_{\max\{t_i,t_j\}}^{+\infty}\right) \exp\left(-\frac{\rho^2((t_i - z)^2 + (t_j - z)^2)}{2(1-\rho^2)} - \frac{z^2}{2}\right)\mathrm{d}z \\
&\leq \sum_{i,j=1}^{m} \frac{1}{2}\frac{A_i A_j}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{\rho^2((t_i - z)^2 + (t_j - z)^2)}{2(1-\rho^2)} - \frac{z^2}{2}\right)\mathrm{d}z \\
&= \sum_{i,j=1}^{m} \frac{1}{2} A_i A_j \sqrt{\frac{1-\rho^2}{1+\rho^2}} \exp\left(-\frac{\rho^2(t_i^2 + t_j^2)}{2(1-\rho^4)} + \frac{\rho^4 t_i t_j}{1-\rho^4}\right),
\end{aligned}$$

where in the last inequality we used the definition of Gaussian pdf with variance $\frac{1-\rho^2}{1+\rho^2}$ and the fact that its integral over the real line is equal to one. Comparing with the expression for $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi'(z/\rho))^2]$ from Lemma 7.4.11, we immediately get the claimed bound on $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi'(z/\rho))^2]$. □

Our next result shows that when $\rho$ is close to 1, the centered augmentation $\mathrm{T}_\rho\Phi(z/\rho)$ does not differ much from the uncentered augmentation $\mathrm{T}_\rho\Phi(z)$, as stated below.

**Lemma 7.4.13.** *Let $\Phi \in \mathcal{F}_M$. Suppose $1 > \rho^2 \geq 1 - C/M^2$ for an absolute constant $C \in (0, M^2/2]$. Then:*

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z) - \mathrm{T}_\rho\Phi(z/\rho))^2] \leq C'(1-\rho^2)(\|\mathrm{T}_\rho\Phi'(z/\rho)\|_{L_2}^2 + \|\mathrm{T}_\rho\Phi'(z)\|_{L_2}^2),$$

*where $C'$ is an absolute constant.*

*Proof.* We first observe that since $\mathrm{T}_\rho$ is a linear operator on functionals, we have

$\mathrm{T}_\rho\Phi(z) - \mathrm{T}_\rho\Phi(z/\rho) = \mathrm{T}_\rho(\Phi(z) - \Phi(z/\rho))$. Given a staircase function $\Phi \in \mathcal{F}_M$, $\Phi(z) = \sum_{i=1}^m A_i\phi(z;t_i) + A_0$, let $I_+ = \{i : t_i > 0\}$ and $I_- = \{i : t_i < 0\}$. Expressing $\mathrm{T}_\rho(\Phi(z) - \Phi(z/\rho))$ in terms of the sum of indicator functions we get

$$\begin{aligned}
\Big|\mathrm{T}_\rho(\Phi(z) - \Phi(z/\rho))\Big| &= \left|\mathrm{T}_\rho\left(\sum_{i=1}^m A_i(\phi(z;t_i) - \phi(z/\rho;t_i))\right)\right| \\
&= \left|\mathrm{T}_\rho\left(\sum_{i=1}^m A_i(-\mathbb{1}\{\rho t_i \le z \le t_i, t_i \ge 0\} + \mathbb{1}\{t_i \le z \le \rho t_i, t_i \le 0\})\right)\right| \\
&\le \sum_{i=1}^m A_i\left|\mathrm{T}_\rho\Big( - \mathbb{1}\{\rho t_i \le z \le t_i, t_i \ge 0\} + \mathbb{1}\{t_i \le z \le \rho t_i, t_i \le 0\}\Big)\right| \\
&= \sum_{i\in I_+} A_i\mathrm{T}_\rho(\mathbb{1}\{\rho t_i \le z \le t_i\}) + \sum_{i\in I_-} A_i\mathrm{T}_\rho(\mathbb{1}\{t_i \le z \le \rho t_i\})
\end{aligned}$$

Suppose first $t_i \ge 0$. Then by the definition of Ornstein–Uhlenbeck semigroup, we have

$$g_i(z) \stackrel{\text{def}}{=} \mathrm{T}_\rho(\mathbb{1}\{\rho t_i \le z \le t_i\}) = \mathop{\mathbf{E}}_{u\sim\mathcal{N}}[\mathbb{1}\{\rho t_i \le \rho z + \sqrt{1-\rho^2}u \le t_i\}] = \int_{\rho(t_i-z)/\sqrt{1-\rho^2}}^{(t_i-\rho z)/\sqrt{1-\rho^2}} \frac{e^{-u^2/2}}{\sqrt{2\pi}}\,\mathrm{d}u$$

When $z \le t_i$ or $z \ge t_i/\rho$, $t_i - \rho z$ and $\rho(t_i - z)$ are both positive or negative, therefore, when $z \in (-\infty, t_i] \cup [t_i/\rho, +\infty)$, the function $g_i(z)$ can be bounded by

$$\begin{aligned}
g_i(z) &= \int_{\rho(t_i-z)/\sqrt{1-\rho^2}}^{(t_i-\rho z)/\sqrt{1-\rho^2}} \frac{e^{-u^2/2}}{\sqrt{2\pi}}\,\mathrm{d}u \\
&\le \frac{1}{\sqrt{2\pi}}\left(\frac{t_i - \rho z}{\sqrt{1-\rho^2}} - \frac{\rho(t_i - \rho z)}{\sqrt{1-\rho^2}}\right)\exp\left(-\frac{1}{2}\min\left\{\frac{(t_i-\rho z)^2}{1-\rho^2}, \frac{\rho^2(t_i - z)^2}{1-\rho^2}\right\}\right) \\
&\le \frac{(1-\rho)t_i}{\sqrt{2\pi(1-\rho^2)}}\left(\exp\left(-\frac{(t_i-\rho z)^2}{2(1-\rho^2)}\right) + \exp\left(-\frac{\rho^2(t_i-z)^2}{2(1-\rho^2)}\right)\right).
\end{aligned}$$

Comparing the right-hand side of the inequality above with the expressions for $\mathrm{T}_\rho\phi'(z;t_i)$

and $\mathrm{T}_\rho\phi'(z;\rho t_i)$ displayed in Equation (7.25) and Lemma 7.4.11,

$$\mathrm{T}_\rho\phi'(z;t_i) = \frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left(-\frac{(\rho z - t_i)^2}{2(1-\rho^2)}\right)$$
$$\mathrm{T}_\rho\phi'(z;\rho t_i) = \frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left(-\frac{\rho^2(z - t_i)^2}{2(1-\rho^2)}\right)$$

we obtain that

$$g_i(z)\mathbb{1}\{z \le t_i \text{ or } z \ge t_i/\rho\} \le (1-\rho)t_i(\mathrm{T}_\rho\phi'(z;t_i) + \mathrm{T}_\rho\phi'(z;\rho t_i)).$$

On the other hand, when $z \in [t_i, t_i/\rho]$, since $0 \in [\rho(t_i - z)/\sqrt{1-\rho^2}, (t_i - \rho z)/\sqrt{1-\rho^2}]$ we can bound $g_i(z)$ above by

$$g(z) \le \int_{\rho(t_i-z)/\sqrt{1-\rho^2}}^{(t_i-\rho z)/\sqrt{1-\rho^2}} \frac{1}{\sqrt{2\pi}}\,\mathrm{d}u \le \frac{(1-\rho)t_i}{\sqrt{2\pi(1-\rho^2)}}.$$

Thus, in summary, $g_i(z)$ is bounded above by

$$\begin{aligned} g_i(z) &= g_i(z)\mathbb{1}\{z \le t_i \text{ or } z \ge t_i/\rho\} + g_i(z)\mathbb{1}\{t_i \le z \le t_i/\rho\} \\ &\le (1-\rho)t_i(\mathrm{T}_\rho\phi'(z;t_i) + \mathrm{T}_\rho\phi'(z;\rho t_i)) + \frac{(1-\rho)t_i}{\sqrt{2\pi(1-\rho^2)}}\mathbb{1}\{t_i \le z \le t_i/\rho\}. \end{aligned}$$

Similarly, for $i \in I_-$, with the same arguments we obtain that

$$g_i(z) \le (1-\rho)|t_i|(\mathrm{T}_\rho\phi'(z;t_i) + \mathrm{T}_\rho\phi'(z;\rho t_i)) + \frac{(1-\rho)|t_i|}{\sqrt{2\pi(1-\rho^2)}}\mathbb{1}\{t_i/\rho \le z \le t_i\}.$$

Therefore, the $L_2^2$ difference between $\mathrm{T}_\rho\Phi(z)$ and $\mathrm{T}_\rho\Phi(z/\rho)$ can be bounded by (note

that $A_i, g_i(z) > 0$ for all $i \in [m]$)

$$\begin{aligned}
&\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z) - \mathrm{T}_\rho\Phi(z/\rho))^2] = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\sum_{i=1}^m A_i g_i(z)\right)^2\right] \\
&\le \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\sum_{i=1}^m A_i(1-\rho)|t_i|(\mathrm{T}_\rho\phi'(z;t_i) + \mathrm{T}_\rho\phi'(z;\rho t_i))\right.\right. \\
&\qquad \left.\left. + \frac{A_i(1-\rho)|t_i|}{\sqrt{2\pi(1-\rho^2)}}\sum_{i=1}^m(\mathbb{1}\{z\in[t_i,t_i/\rho]\} + \mathbb{1}\{z\in[t_i/\rho,t_i]\})\right)^2\right] \\
&\le \underbrace{2\mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\sum_{i=1}^m A_i(1-\rho)|t_i|(\mathrm{T}_\rho\phi'(z;t_i) + \mathrm{T}_\rho\phi'(z;\rho t_i))\right)^2\right]}_{(Q_1)} \\
&\qquad + \underbrace{2\mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\frac{(1-\rho)}{\sqrt{2\pi(1-\rho^2)}}\sum_{i=1}^m A_i|t_i|(\mathbb{1}\{z\in[t_i,t_i/\rho]\} + \mathbb{1}\{z\in[t_i/\rho,t_i]\})\right)^2\right]}_{(Q_2)}. \quad (7.27)
\end{aligned}$$

Note that in $(Q_2)$ above, we used the convention that if $a > b$ then $[a,b] = \emptyset$ and $\mathbb{1}\{z \in \emptyset\} = 0$. For $(Q_1)$, using Young's inequality again yields:

$$\begin{aligned}
(Q_1) &\le 2\mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\sum_{i=1}^m A_i(1-\rho)|t_i|\mathrm{T}_\rho\phi'(z;t_i)\right)^2\right] + 2\mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\sum_{i=1}^m A_i(1-\rho)|t_i|\mathrm{T}_\rho\phi'(z;\rho t_i)\right)^2\right] \\
&\le 2(1-\rho)^2\left(\max_{i\in[m]}\{|t_i|\}\right)^2\left(\mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\sum_{i=1}^m A_i\mathrm{T}_\rho\phi'(z;t_i)\right)^2\right] + \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\sum_{i=1}^m A_i\mathrm{T}_\rho\phi'(z;\rho t_i)\right)^2\right]\right) \\
&\le 2(1-\rho)^2M^2(\|\mathrm{T}_\rho\Phi'(z)\|_{L_2}^2 + \|\mathrm{T}_\rho\Phi'(z/\rho)\|_{L_2}^2),
\end{aligned}$$

where in the last inequality, we use the fact that since $\Phi \in \mathcal{F}_M$, we have $|t_i| \le M$ for all $i \in [m]$. Now, by our assumption, $\rho^2 \ge 1 - C/M^2$, therefore, $(1-\rho)(1+\rho) \le C/M^2$ and hence $1 - \rho \le C/M^2$, which implies

$$(Q_1) \le 2C(1-\rho)(\|\mathrm{T}_\rho\Phi'(z)\|_{L_2}^2 + \|\mathrm{T}_\rho\Phi'(z/\rho)\|_{L_2}^2).$$

For $(Q_2)$, since $|t_i| \le M$ for all $i \in [m]$ and $1-\rho \le C/M^2$, expanding the square yields

$$(Q_2) \le \frac{M^2(1-\rho)}{2\pi(1+\rho)} \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\sum_{i=1}^{m} A_i(\mathbb{1}\{z \in [t_i, t_i/\rho]\} + \mathbb{1}\{z \in [t_i/\rho, t_i]\})\right)^2\right] \le$$
$$\frac{C}{4\pi} \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\sum_{i,j\in I_+} A_iA_j\mathbb{1}\{z \in [t_i, t_i/\rho] \cap [t_j, t_j/\rho]\} + \sum_{i,j\in I_-} A_iA_j\mathbb{1}\{z \in [t_i/\rho, t_i] \cap [t_j/\rho, t_j]\}\right]$$

By the symmetry of Gaussian distribution, we have

$$\mathbf{Pr}[z \in [t_i/\rho, t_i] \cap [t_j/\rho, t_j], t_i, t_j \le 0] = \mathbf{Pr}[z \in [|t_i|, |t_i|/\rho] \cap [|t_j|, |t_j|/\rho]],$$

therefore, it suffices to discuss only the case where $t_i, t_j \in I_+$. Suppose without loss of generality that $0 < t_i \le t_j$. Observe that $\mathbf{E}_{z\sim\mathcal{N}}[\mathbb{1}\{z \in [t_i, t_i/\rho] \cap [t_j, t_j/\rho]\}] \ne 0$ if and only if $0 < t_i \le t_j < t_i/\rho \le t_j\rho$, therefore, the expectation of the indicator is bounded by:

$$\begin{aligned}\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\mathbb{1}\{z \in [t_i, t_i/\rho] \cap [t_j, t_j/\rho]\}] = \mathbf{Pr}[z \in [t_j, t_i/\rho]] &= \int_{t_j}^{t_i/\rho} \frac{e^{-u^2/2}}{\sqrt{2\pi}}\,\mathrm{d}u \\ &\le \frac{\exp(-t_j^2/2)}{\sqrt{2\pi}}(t_i/\rho - t_j) \le \frac{(1-\rho)t_i}{\rho\sqrt{2\pi}}\exp\left(-\frac{t_j^2}{2}\right).\end{aligned} \tag{7.28}$$

Recall that in Lemma 7.4.11, we proved

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi'(z/\rho))^2] = \sum_{i,j=1}^{m} \frac{\rho^2A_iA_j}{2\pi\sqrt{1-\rho^4}}\exp\left(-\frac{\rho^2(t_i^2+t_j^2)}{2(1-\rho^4)} + \frac{\rho^4t_it_j}{1-\rho^4}\right),$$

hence our strategy is to show that:

$$\exp\left(-\frac{t_j^2}{2}\right) \le \exp\left(-\frac{\rho^2(t_i^2+t_j^2)}{2(1-\rho^4)} + \frac{\rho^4 t_i t_j}{1-\rho^4}\right), \text{ for } 0 < t_i \le t_j < t_i/\rho \le t_j/\rho. \quad (7.29)$$

We show:

**Claim 7.4.14.** *Let $t_i, t_j > 0$ satisfy $t_i \le t_j \le t_i/\rho$. Then, for any $\rho \in (0,1)$, it holds*

$$-\frac{t_j^2}{2} \le -\frac{\rho^2(t_i^2+t_j^2)}{2(1-\rho^4)} + \frac{\rho^4 t_i t_j}{1-\rho^4}.$$

*Proof of Claim 7.4.14.* Simple algebraic calculation yields

$$\begin{aligned}
-\frac{t_j^2}{2} + \frac{\rho^2(t_i^2+t_j^2)}{2(1-\rho^4)} - \frac{\rho^4 t_i t_j}{1-\rho^4} &= \frac{1}{2(1-\rho^4)}(-t_j^2 + \rho^2 t_i^2 + \rho^2 t_j^2 + \rho^4 t_j^2 - 2\rho^4 t_i t_j) \\
&= \frac{1}{2(1-\rho^4)}(-(1-\rho^2)t_j^2 + \rho^2 t_i^2 + \rho^4(t_j - t_i)^2 - \rho^4 t_i^2) \\
&= \frac{1}{2(1-\rho^4)}(-(1-\rho^2)t_j^2 + \rho^2(1-\rho^2)t_i^2 + \rho^4(t_j - t_i)^2).
\end{aligned}$$

Now since $0 < t_j - t_i < t_i/\rho - t_i = (1-\rho)t_i/\rho$, we further have

$$\begin{aligned}
&-\frac{t_j^2}{2} + \frac{\rho^2(t_i^2+t_j^2)}{2(1-\rho^4)} - \frac{\rho^4 t_i t_j}{1-\rho^4} \\
&\le \frac{1}{2(1-\rho^4)}(-(1-\rho^2)t_j^2 + \rho^2(1-\rho^2)t_i^2 + \rho^2(1-\rho)^2 t_i^2) \\
&\le \frac{1}{2(1-\rho^4)}(-(1-\rho^2) + \rho^2(1-\rho^2) + \rho^2(1-\rho)^2)t_j^2 \\
&= \frac{t_j^2}{2(1+\rho)(1+\rho^2)}(-(1+\rho) + \rho^2(1+\rho) + \rho^2(1-\rho)) = \frac{-t_j^2(1-\rho)(1+2\rho)}{2(1+\rho)(1+\rho^2)} < 0.
\end{aligned}$$

Thus, indeed we have $-t_j^2/2 \le -\rho^2(t_i^2 + t_j^2 - 2\rho^2 t_i t_j)/(2(1-\rho^4))$. □

Therefore, for each $t_i, t_j$, $i, j \in [m]$, the expectation in Equation (7.28) is bounded above by

$$\begin{aligned}
&\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\mathbb{1}\{z \in [t_i, t_i/\rho] \cap [t_j, t_j/\rho], t_i t_j > 0\}] \\
&\leq \frac{(1-\rho)\sqrt{2\pi(1-\rho^4)}|t_i|}{\rho^3} \frac{\rho^2}{2\pi\sqrt{1-\rho^4}} \exp\left( -\frac{\rho^2(t_i^2+t_j^2)}{2(1-\rho^4)} + \frac{\rho^4 t_i t_j}{1-\rho^4} \right),
\end{aligned}$$

which, combining with the fact that $\sqrt{1-\rho} \leq \sqrt{C}/M$ and $|t_i| \leq M$, yields

$$\begin{aligned}
(Q_2) &\leq \frac{C}{4\pi}\Bigg( \sum_{i,j\in I_+} A_i A_j \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\mathbb{1}\{z \in [t_i, t_i/\rho] \cap [t_j, t_j/\rho]\}] \\
&\qquad + \sum_{i,j\in I_-} A_i A_j \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\mathbb{1}\{z \in [t_i/\rho, t_i] \cap [t_j\rho, t_j]\}] \Bigg) \\
&\leq \sum_{i,j=1}^{m} \frac{C(1-\rho)\sqrt{2\pi(1-\rho^4)}|t_i|}{4\pi\rho^3} \frac{\rho^2}{2\pi\sqrt{1-\rho^4}} \exp\left( -\frac{\rho^2(t_i^2+t_j^2)}{2(1-\rho^4)} + \frac{\rho^4 t_i t_j}{1-\rho^4} \right) \\
&\leq C'(1-\rho)\|\mathrm{T}_\rho \Phi'(z/\rho)\|_{L_2}^2.
\end{aligned}$$

Plugging the bounds on $(Q_1)$, $(Q_2)$ back to Equation (7.27), we finally obtain:

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi(z) - \mathrm{T}_\rho \Phi(z/\rho))^2] \leq C''(1-\rho)(\|\mathrm{T}_\rho \Phi'(z/\rho)\|_{L_2}^2 + \|\mathrm{T}_\rho \sigma'(z)\|_{L_2}^2).$$

Since $1-\rho \leq 1-\rho^2$, we complete the proof of Lemma 7.4.13. □

Our last step is to show that $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi'(z/\rho))^2]$ is not much larger than $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi'(z))^2]$ when $\rho$ is close to 1.

**Lemma 7.4.15.** *Let $\Phi \in \mathcal{F}_M$ be any staircase function that is constructed from $m$ indicator functions with thresholds $t_i$, $i \in [m]$, and suppose that $|t_i| \leq M$, for all $i \in [m]$, where $1 < M < +\infty$. For any $\rho \in (0,1)$ such that $\rho^2 \geq 1 - C/M^2$ where $C < M^2$ is an*

*absolute constant, let* $\rho_1 = \sqrt{\rho^2 + C(1-\rho^2)/M^2}$. *Then,*

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z/\rho_1))^2] \le 2e^C \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho}\Phi'(z))^2].$$

*Proof.* Observe first that $1-\rho_1^2 = (1-\rho^2)(1-C/M^2) \in (0,1)$, hence $\rho_1 \in (0,1)$ and $\mathrm{T}_{\rho_1}\Phi'(z/\rho_1)$ is well-defined. To proceed, we compare each term of $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z/\rho_1))]$ and $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho}\Phi'(z))^2]$ that are given in Lemma 7.4.11 and Lemma 7.4.9 separately.

Since $\rho_1^2/(1-\rho_1^4)$ appears in the exponential terms of $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z/\rho_1))]$ while the coefficient in the exponential terms of $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho}\Phi'(z))^2]$ is $1/(1-\rho^4)$, we first need to compare these two factors.

**Claim 7.4.16.** *Let* $\rho_1^2 = \rho^2 + C(1-\rho^2)/M^2$. *If* $1 > \rho^2 \ge 1 - C/M^2$, *then* $\rho_1^2/(1-\rho_1^4) \ge 1/(1-\rho^4)$.

*Proof of Claim 7.4.16.* Since $\rho_1, \rho < 1$, we only need to show that $\rho_1^2(1-\rho^4) \ge 1-\rho_1^4$. Plugging in the value of $\rho_1$, we have

$$\rho_1^2(1-\rho^4) = (\rho^2 + C(1-\rho^2)/M^2)(1-\rho^4) = (\rho^2 + C(1-\rho^2)/M^2)(1-\rho^2)(1+\rho^2);$$
$$1-\rho_1^4 = 1-(\rho^2 + C(1-\rho^2)/M^2)^2 = (1+\rho^2 + C(1-\rho^2)/M^2)(1-\rho^2)(1-C/M^2).$$

Therefore, our goal is to prove that

$$(\rho^2 + C(1-\rho^2)/M^2)(1-\rho^2)(1+\rho^2) \ge (1+\rho^2 + C(1-\rho^2)/M^2)(1-\rho^2)(1-C/M^2).$$

Dividing both sides of the inequality above by $1-\rho^2 > 0$ yields that it is sufficient to

show the following inequality:

$$\rho^2 + C(1-\rho^2)/M^2 + (\rho^2 + C(1-\rho^2)/M^2)\rho^2$$
$$\geq (1 - C/M^2) + (\rho^2 + C(1-\rho^2)/M^2)(1 - C/M^2).$$

Since $\rho^2 \geq 1 - C/M^2$, we have $\rho^2 + C(1-\rho^2)/M^2 \geq 1 - C/M^2$ and

$$(\rho^2 + C(1-\rho^2)/M^2)\rho^2 \geq (\rho^2 + C(1-\rho^2)/M^2)(1 - C/M^2).$$

Thus, it holds that $\rho_1^2(1-\rho^2) \geq 1 - \rho_1^4$. □

Observe that for any $t_i, t_j \in \mathbb{R}$ and $\rho \in (0,1)$, we have $t_i^2 + t_j^2 - 2\rho_1^2 t_i t_j \geq (1-\rho^2)(t_i^2 + t_j^2) \geq 0$, and recalling the expression for $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z/\rho_1))^2]$ given in Lemma 7.4.11, we thus obtain

$$\begin{aligned}
\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z/\rho_1))^2] &= \sum_{i,j=1}^{m} \frac{A_i A_j}{2\pi}\sqrt{\frac{\rho_1^4}{1-\rho_1^4}}\exp\left(-\frac{\rho_1^2(t_i^2+t_j^2-2\rho_1^2 t_i t_j)}{2(1-\rho_1^4)}\right) \\
&\overset{(i)}{\leq} \sum_{i,j=1}^{m} \frac{A_i A_j}{2\pi}\sqrt{\frac{\rho_1^4}{1-\rho_1^4}}\exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho^4)} + \frac{\rho_1^2 t_i t_j}{1-\rho^4}\right) \\
&\overset{(ii)}{=} \sum_{i,j=1}^{m} \frac{A_i A_j}{2\pi}\sqrt{\frac{\rho_1^4}{1-\rho_1^4}}\exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho^4)} + \frac{\rho^2 t_i t_j}{1-\rho^4} + \frac{C(1-\rho^2)t_i t_j}{(1-\rho^4)M^2}\right) \\
&= \sum_{i,j=1}^{m} \frac{A_i A_j}{2\pi}\sqrt{\frac{\rho_1^4}{1-\rho_1^4}}\exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho^4)} + \frac{\rho^2 t_i t_j}{1-\rho^4}\right)\exp\left(\frac{C t_i t_j}{(1+\rho^2)M^2}\right),
\end{aligned} \tag{7.30}$$

where in $(i)$ we plugged in Claim 7.4.16 and in $(ii)$ we used the definition that $\rho_1^2 = \rho^2 + C(1-\rho^2)/M^2$.

Since $M$ is an upper bound on $|t_i|, i \in [m]$, we can assume without loss of generality that $M^2 \geq 2C$. We next observe that when $M^2 \geq 2C$, we have the following inequality.

**Claim 7.4.17.** *If $M^2 \geq 2C$ then $\rho_1^4/(1-\rho_1^4) \leq 4/(1-\rho^4)$.*

*Proof of Claim 7.4.17.* For any fixed $\rho \in (0,1)$, let us define

$$h(M) = \frac{(\rho^2 + C(1-\rho^2)/M^2)(1-\rho^4)}{1-(\rho^2+C(1-\rho^2)/M^2)^2}.$$

It is easy to see that $h(M)$ is a decreasing function with respect to $M > 0$, therefore, for any fixed $\rho \in (0,1)$ and any $M^2 \geq 2C$, we have

$$h(M) = \frac{(\rho^2 + C(1-\rho^2)/M^2)(1-\rho^4)}{1-(\rho^2+C(1-\rho^2)/M^2)^2} \leq h(\sqrt{2C}) = \frac{((1+\rho^2)/2)^2(1-\rho^4)}{1-((1+\rho^2)/2)^2} \leq 4.$$

Therefore, for any $\rho \in (0,1)$, it holds $\rho_1^4/(1-\rho_1^4) \leq 4/(1-\rho^2)$. □

Thus, plugging in Claim 7.4.17 into Equation (7.30), and recalling that it is assumed $|t_i|^2 \leq M^2$ for any $i \in [m]$, we further get

$$\begin{aligned}
\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z/\rho_1))^2] &\leq \sum_{i,j=1}^{m} \frac{A_iA_j}{\pi}\frac{1}{\sqrt{1-\rho^4}}\exp(-\frac{t_i^2+t_j^2}{2(1-\rho^4)}+\frac{\rho^2 t_it_j}{1-\rho^4})\exp(\frac{Ct_it_j}{(1+\rho^2)M^2}) \\
&\leq 2e^C \sum_{i,j=1}^{m} \frac{A_iA_j}{2\pi\sqrt{1-\rho^4}}\exp(-\frac{t_i^2+t_j^2}{2(1-\rho^4)}+\frac{\rho^2 t_it_j}{1-\rho^4}) \\
&= 2e^C \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho}\Phi'(z))^2],
\end{aligned}$$

which completes the proof. □

### Proof of Proposition 7.4.8

We can now restate Proposition 7.4.8 and present its proof.

**Proposition 7.4.18.** *Let $\Phi \in \mathcal{F}_M$ be any staircase function that is consists of $m$ indicator functions with thresholds $t_i$, $i \in [m]$, and suppose $|t_i| \leq M$ for all $i \in [m]$,*

*where* $1 < M < +\infty$. *For any* $\rho \in (0,1)$ *such that* $\rho^2 \geq 1 - C/M^2$ *where* $C < M^2/4$ *is an absolute constant, we have*

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi(z) - \Phi(z))^2] \lesssim (1-\rho^2) \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi'(z))^2].$$

*Proof of Proposition 7.4.8.* Observe that $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi(z) - \Phi(z))^2]$ can be bounded as

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi(z) - \Phi(z))^2] \leq 2 \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi(z) - \mathrm{T}_{\rho_1}\Phi(z/\rho_1))^2] + 2 \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1} \Phi(z/\rho_1) - \Phi(z))^2]. \tag{7.31}$$

We first bound the second term in Equation (7.31). Since we have assumed that $\rho^2 \geq 1 - C/M^2$, where $C$ is an absolute constant, using Lemma 7.4.12 and Lemma 7.4.15 and plugging in $\rho_1 = \sqrt{\rho^2 + C(1-\rho^2)/M^2}$, we get

$$\begin{aligned}
\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1} \Phi(z/\rho_1) - \Phi(z))^2] &\leq \frac{4(1-\rho_1^2)}{\rho_1^2} \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1} \Phi'(z/\rho_1))^2] \\
&\leq \frac{8e^C(1-\rho^2 - C(1-\rho^2)/M^2)}{(\rho^2 + C(1-\rho^2)/M^2)} \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho} \Phi'(z))^2] \\
&\lesssim e^C (1-\rho^2) \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho} \Phi'(z))^2]. \qquad (7.32)
\end{aligned}$$

Now we turn to bounding the first term in Equation (7.31). First, we add and subtract $\mathrm{T}_{\rho_1}\Phi'(z)$ in the squared parentheses to obtain:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi(z) - \mathrm{T}_{\rho_1}\Phi(z/\rho_1))^2] \\
&\leq 2 \underbrace{\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho \Phi(z) - \mathrm{T}_{\rho_1}\Phi(z))^2]}_{Q_1} + 2 \underbrace{\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1} \Phi(z) - \mathrm{T}_{\rho_1}\Phi(z/\rho_1))^2]}_{Q_2}. \qquad (7.33)
\end{aligned}$$

For $Q_1$, observe that since $\rho < \rho_1 < 1$, using the property of Ornstein–Uhlenbeck

semigroup presented in Fact 7.1.7, we have $\mathrm{T}_\rho\Phi(z) = \mathrm{T}_{\rho_1(\rho/\rho_1)}\Phi(z) = \mathrm{T}_{\rho/\rho_1}(\mathrm{T}_{\rho_1}\Phi(z))$. Therefore, using Lemma 7.1.10 with $f(z) = \mathrm{T}_{\rho_1}\Phi(z)$ we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho/\rho_1}(\mathrm{T}_{\rho_1}\Phi(z)) - \mathrm{T}_{\rho_1}\Phi(z))^2] &\le 3(1-\rho/\rho_1)\mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\frac{\mathrm{d}}{\mathrm{d}z}\mathrm{T}_{\rho_1}\Phi(z)\right)^2\right] \\
&\overset{(i)}{=} \frac{3(\rho_1-\rho)}{\rho_1}\rho_1^2\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z))^2] \\
&= \frac{3(\rho_1^2-\rho^2)\rho_1}{\rho_1+\rho}\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z))^2] \\
&\overset{(ii)}{=} \frac{3C(1-\rho^2)\rho_1}{M^2(\rho_1+\rho)}\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z))^2], \qquad (7.34)
\end{aligned}$$

note that in $(i)$ we applied Fact 7.1.7, Part 2(g), and in $(ii)$ we brought in the definition of $\rho_1^2 = \rho^2 + C(1-\rho^2)/M^2$. It remains to bound $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z))^2]$ above by $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho}\Phi'(z))^2]$.

**Claim 7.4.19.** *Let $\rho^2 \ge 1 - C/M^2$ and $\rho_1^2 = \rho^2 + C(1-\rho^2)/M^2$. Then, $\|\mathrm{T}_{\rho_1}\Phi'(z)\|_{L_2}^2 \le e^C\|\mathrm{T}_\rho\Phi'(z)\|_{L_2}^2$.*

*Proof.* To prove this claim, we recall the explicit expression of $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z))^2]$ using

the formula displayed in Lemma 7.4.9:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z))^2] = \sum_{i,j=1}^{m} \frac{A_iA_j}{2\pi\sqrt{1-\rho_1^4}} \exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho_1^4)} + \frac{\rho_1^2 t_i t_j}{1-\rho_1^4}\right) \\
&= \sum_{i,j=1}^{m} \frac{A_iA_j}{2\pi\sqrt{1-\rho_1^4}} \exp\left(-\frac{(t_i^2+t_j^2-2\rho_1^2 t_i t_j)}{2(1-\rho_1^4)}\right) \\
&\overset{(i)}{\leq} \sum_{i,j=1}^{m} \frac{A_iA_j}{2\pi\sqrt{(1-\rho_1^2)(1+\rho_1^2)}} \exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho^4)} + \frac{\rho_1^2 t_i t_j}{1-\rho^4}\right) \\
&\overset{(ii)}{\leq} \sum_{i,j=1}^{m} \frac{A_iA_j}{2\pi\sqrt{(1-\rho^2)(1-C/M^2)(1+\rho_1^2)}} \exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho^4)} + \frac{\rho^2 t_i t_j}{1-\rho^4} + \frac{(\rho_1^2-\rho^2)t_i t_j}{1-\rho^4}\right) \\
&\overset{(iii)}{\leq} 2\sum_{i,j=1}^{m} \frac{A_iA_j}{2\pi\sqrt{1-\rho^4}} \exp\left(-\frac{t_i^2+t_j^2}{2(1-\rho^4)} + \frac{\rho_1^2 t_i t_j}{1-\rho^4}\right) \exp\left(\frac{(\rho_1^2-\rho^2)t_i t_j}{1-\rho^4}\right) \\
&\leq 2\,\mathbf{E}[(\mathrm{T}_{\rho}\Phi'(z))^2] \exp\left(\frac{(\rho_1^2-\rho^2)M^2}{1-\rho^4}\right).
\end{aligned}$$

Inequality $(i)$ is due to the facts that $(t_i^2+t_j^2-2\rho_1^2 t_i t_j) \geq 0$ for any $t_i, t_j \in \mathbb{R}$ and that $1/(1-\rho_1^4) \geq 1/(1-\rho^2)$ since $\rho < \rho_1 < 1$; in $(ii)$ we plugged in the definition $\rho_1$; $(iii)$ comes from the assumption that $C/M^2 \leq 1/4$ and the fact that $1+\rho_1^2 \geq 1+\rho^2$ since $\rho_1 > \rho$. Finally, for the term $\exp((\rho_1^2-\rho^2)t_i t_j/(1-\rho^4))$, bringing in the definition of $\rho_1^2$ and the fact that $|t_i| \leq M, i \in [m]$ we get

$$\exp\left(\frac{(\rho_1^2-\rho^2)M^2}{1-\rho^4}\right) = \exp\left(\frac{(1-\rho^2)(C/M^2)M^2}{(1-\rho^2)(1+\rho^2)}\right) \leq e^C.$$

Thus, in summary, we have $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z))^2] \leq e^C\,\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho}\Phi'(z))^2]$. □

Plugging the upper bound on $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho_1}\Phi'(z))^2]$ from Claim 7.4.19 back into Equa-

tion (7.34) yields

$$Q_1 = 2 \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\rho/\rho_1}(\mathrm{T}_{\rho_1}\Phi(z)) - \mathrm{T}_{\rho_1}\Phi(z))^2] \leq \frac{C(1-\rho^2)\rho_1}{M^2(\rho_1+\rho)} 4e^C \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi'(z))^2].$$

Since $C/M^2 \leq 1/4$ we have $1 > \rho_1 > \rho \geq 1/2$, thus we finally get

$$Q_1 \leq 2e^C(1-\rho^2)\|\mathrm{T}_\rho\Phi'\|_{L_2}^2.$$

We now turn to bounding the term $Q_2$ in Equation (7.32). Applying Lemma 7.4.13 with $\rho_1$, we obtain

$$Q_2 \leq C''(1-\rho^2)(\|\mathrm{T}_{\rho_1}\Phi'(z)\|_{L_2}^2 + \|\mathrm{T}_{\rho_1}\Phi'(z/\rho_1)\|_{L_2}^2).$$

Applying Claim 7.4.19 and Lemma 7.4.15 again, we obtain

$$Q_2 \leq 4C''(1-\rho^2)e^C\|\mathrm{T}_\rho\Phi'\|_{L_2}^2.$$

Plugging the upper bounds on $Q_1$ and $Q_2$ back into Equation (7.33) yields $\mathbf{E}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z) - \mathrm{T}_{\rho_1}\Phi(z/\rho_1))^2] \lesssim e^C(1-\rho^2)\|\mathrm{T}_\rho\Phi'(z)\|_{L_2}^2$. Finally, combining with Equation (7.32), we obtain

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_\rho\Phi(z) - \Phi(z))^2] \lesssim e^C(1-\rho^2)\|\mathrm{T}_\rho\Phi'\|_{L_2}^2.$$

Since $e^C$ is an absolute constant, this completes the proof of Proposition 7.4.8. □

### 7.4.3 Initialization Algorithm for Monotone Activations

In this section, we provide an initialization algorithm for $\sigma$ that is an $\epsilon$-Extended monotone $(B, L)$-Regular activation. The algorithm generates a vector $\mathbf{w}^{(0)}$ satisfying $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \leq C/M$, where $C$ is an absolute constant and $M$ is at most $\sqrt{\log(B/\epsilon) - \log\log(B/\epsilon)}$. Our key idea is to convert the regression problem to a problem of robustly learning halfspaces via *data transformation*. In particular, we transform $y$ to $\tilde{y} \in \{0, 1\}$ by truncating the labels $y$ to $\tilde{y} = \mathbb{1}\{y \geq t'\}$, where this $t'$ is a carefully chosen threshold. Then we utilize a previous algorithm from [DKTZ22b] to robustly learn $\mathbf{w}^*$.

**Proposition 7.4.20.** *Let $\sigma$ be a non-decreasing $(B, L)$-Regular function. Let $M$ be defined as in Claim 7.5.7. Then, there exists an algorithm that draws $O(d/\epsilon^2 \log(1/\delta))$ samples, it runs in $\mathrm{poly}(d, N)$ time, and, with probability at least $1 - \delta$, it outputs a vector $\mathbf{w}$ such that $\theta(\mathbf{w}, \mathbf{w}^*) \leq \min\{\pi/6, C/M\}$, where $C > 0$ is a universal constant, independent of any problem parameters.*

The proof of Proposition 7.4.2 follows from Proposition 7.4.20 and Proposition 7.4.6.

*Proof of Proposition 7.4.2.* Proposition 7.4.20 implies that there exists an algorithm using $O(d/\epsilon^2)$ samples and outputs a vector $\mathbf{w}^{(0)}$ such that $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \leq C/M$. Now for any $\theta \leq \theta_0$, it holds $\cos\theta^2 \leq 1 - \theta^2 \geq 1 - C^2/M^2$. Thus, using Proposition 7.4.6 we know that $\|\mathrm{P}_{>1/\theta^2}\sigma\|_{L_2}^2 \lesssim \sin^2\theta \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$. This finishes the proof of Proposition 7.4.2. □

Since we are only looking for a constant factor approximate solution, it is sufficient to truncate the activation $\sigma$ to $\tilde{\sigma}$ so that $\|\sigma - \tilde{\sigma}\|_{L_2}^2 \leq C_1 \mathrm{OPT}$ for some absolute constant $C_1$ in Claim 7.5.7. Hence, given an activation $\sigma \in \mathcal{H}(B, L)$, the parameter $M$ is defined as following: fix an absolute constant $C_1 \geq 1$; there exists a function $\tilde{\sigma} \in \mathcal{H}(B, L)$ satisfying

$\|\tilde{\sigma}-\sigma\|_{L_2}^2 \leq 2C_1\epsilon$ such that $\sigma'(z) = 0$ for all $|z| \geq M$. In fact, in the proof of Claim 7.5.7, we chose $\tilde{\sigma}(z) = \sigma(z)\mathbb{1}\{-M_- \leq z \leq M_+\} + \sigma(M_+)\mathbb{1}\{z \geq M_+\} + \sigma(-M_-)\mathbb{1}\{z \leq -M_-\}$, such that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(M_+))^2\mathbb{1}\{z \geq M_+\}] = C_1\epsilon$ and $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(-M_-))\mathbb{1}\{z \leq -M_-\}] = C_1\epsilon$, then $M = \max\{M_+, M_-\} \leq \sqrt{\log(B/\epsilon) - \log\log(B/\epsilon)}$. In the following, let us assume without loss of generality that $M = M_+ \geq M_-$ since if $M_+ \leq M_-$, we can instead consider $-\sigma(-z)$.

Our goal is to show that, there exists $M^* \geq M$ such that it holds: $\mathbf{Pr}[\tilde{y} \neq \mathbb{1}\{\mathbf{w}^* \cdot \mathbf{x} \geq M^*\}] \leq (4/\sqrt{C_1})\,\mathbf{Pr}[\mathbf{w}^* \cdot \mathbf{x} \geq M^*]$, where $\tilde{y} = \mathcal{T}(y) = \mathbb{1}\{y \geq \sigma(M^*)\}$. Then, we will use the following fact from [DKTZ22b], which states:

**Fact 7.4.21** ([DKTZ22b], Corollary of Lemma C.3 and Theorem C.1)**.** *There is an algorithm that given a halfspace $\phi(\mathbf{w}^* \cdot \mathbf{x}; t)$ and a distribution $(\mathbf{x}, \tilde{y}) \sim \mathcal{D}$ of labeled examples with $\mathrm{OPT}'$-adversarial noise—meaning that $\mathbf{Pr}[\phi(\mathbf{w}^* \cdot \mathbf{x}; t) \neq \tilde{y}] \leq \mathrm{OPT}'$, if $\exp(-t^2/2)/t \geq C_2\mathrm{OPT}'$, where $C_2 > 1$ is a large universal constant—returns $\mathbf{w}$ such that $\theta(\mathbf{w}, \mathbf{w}^*)\exp(-t^2/2) \leq C_3\mathrm{OPT}'$ and $\theta(\mathbf{w}, \mathbf{w}^*) \leq \pi/6$, where $C_3$ is also a universal constant, with probability at least $1-\delta$, using $O(d/\epsilon^2\log(1/\delta))$ samples.*

With the error $\mathrm{OPT}' \leq \mathbf{Pr}[z \geq M^*]$ and $t = M^*$ in Fact 7.4.21, we obtain a vector $\mathbf{w}$ that satisfies $\theta(\mathbf{w}, \mathbf{w}^*) \leq C_3\exp((M^*)^2/2)\,\mathbf{Pr}[z \geq M^*] \leq C_3/M^* \leq C_3/M$, where we used the fact that $\mathbf{Pr}[z \geq M^*] \approx \exp(-(M^*)^2/2)/M^*$. This will complete our initialization argument.

To proceed, we prove the following key lemma:

**Lemma 7.4.22.** *Fix $C > 1$. Let $f$ be a monotone function such that $f \geq 0$ and $\|f - y\|_{L_2}^2 \leq \epsilon$. Assume that for all $q > 0$ it holds that $\mathbf{E}[|\mathbb{1}\{f(z) \geq q\} - \mathbb{1}\{y \geq q\}|] \geq \mathbf{Pr}[f(z) \geq q]/C$. Then, it holds that $\|f\|_{L_2}^2 \leq 5C^2\epsilon$.*

*Proof.* Let $T(q) = \mathbf{Pr}[f(z) \geq q]$ and $\Delta(q) = \mathbf{E}[|\mathbb{1}\{f(z) \geq q\} - \mathbb{1}\{y \geq q\}|]$. From the assumption we have that $T(q) \leq C\Delta(q)$. Therefore, we have that

$$\begin{aligned}
\mathbf{E}[f^2] &= \int_0^\infty 2qT(q)dq \leq \int_0^\infty 2q(C\Delta(q))dq \\
&= 2C\int_0^\infty q(\mathbf{Pr}[f \geq q, y < q] + \mathbf{Pr}[f < q, y \geq q])dq \\
&= 2C\left(\mathbf{E}\left[\frac{f^2 - y^2}{2}\mathbb{1}\{f > y\}\right] + \mathbf{E}\left[\frac{y^2 - f^2}{2}\mathbb{1}\{y > f\}\right]\right) \\
&= 2C\,\mathbf{E}\left[\frac{|f^2 - y^2|}{2}\right] = C\,\mathbf{E}[|f - y||f + y|] \\
&\leq C\sqrt{\mathbf{E}[(f-y)^2]\,\mathbf{E}[(f+y)^2]} \leq C\sqrt{\epsilon\,\mathbf{E}[(f+y)^2]}.
\end{aligned}$$

Note that $\mathbf{E}[(f + y)^2] \leq 4\,\mathbf{E}[f^2] + 4\epsilon$, therefore we have that

$$\mathbf{E}[f^2] \leq C\sqrt{4\epsilon(\mathbf{E}[f^2] + \epsilon)}\ .$$

Let $\tau = \mathbf{E}[f^2]$, the above becomes

$$\tau^2 \leq 4C^2\epsilon\tau + 4C^2\epsilon^2\ .$$

Maximizing over $\tau$, we have that $\tau \leq 5C^2\epsilon$ provided that $C > 1$. Therefore $\mathbf{E}[f^2] \leq 5C^2\epsilon$. □

We can now prove Proposition 7.4.20.

*Proof of Proposition 7.4.20.* Let $\sigma \in \mathcal{H}(B, L)$. Throughout the proof, we make the following assumptions that are without loss of generality:

1. There exists $\bar{M} < \infty$ such that $\sigma(z) = \sigma(\bar{M})$ when $z \geq \bar{M}$ and $\sigma(z) = \sigma(-\bar{M})$ when $z \leq -\bar{M}$. This is without loss of generality according to Claim 7.5.7.

2. $B = \max\{|\sigma(\bar{M})|, |\sigma(-\bar{M})|\} = \sigma(\bar{M})$, since we can always shift $\sigma(z)$ to $\sigma(z) + |\sigma(-\bar{M})|$ without affecting any of the results.

3. By Claim 7.5.6, it holds $|y| \leq B = \sigma(\bar{M})$ without loss of generality.

4. $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq \epsilon$, and $h(z) = \sigma(0)$ is not an approximate solution, i.e., for any absolute constant $C$ we have $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - h(\mathbf{w}^* \cdot \mathbf{x}))^2] \geq C\epsilon$.

5. There exists $M \in [0, \bar{M}]$ such that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(M))^2 \mathbb{1}\{z \geq M\}] \geq C_1\epsilon$, where $C_1 > 1$ is a large absolute constant. In the following of the proof, we will denote $M$ as the smallest value in $[0, \bar{M}]$ such that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(M))^2 \mathbb{1}\{z \geq M\}] \geq C_1\epsilon$. Such $M$ exists because if $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(0))^2 \mathbb{1}\{z \geq 0\}] < C_1\epsilon$, then it must hold that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(0))^2 \mathbb{1}\{z \leq 0\}] \geq C_1\epsilon$, otherwise we have $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(0))^2] \leq 2C_1\epsilon$ which further implies $h(z) = \sigma(0)$ is a constant approximate solution as $\mathbf{E}_{z\sim\mathcal{N}}[(y - h(z))^2] \leq 2\,\mathbf{E}_{z\sim\mathcal{N}}[(y - \sigma(z))^2] + 2\,\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - h(z))^2] \leq (2C_1 + 2)\epsilon$, contradicting the assumption above. Since $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(0))^2 \mathbb{1}\{z \leq 0\}] \geq C_1\epsilon$, we can consider $\tilde{\sigma}(z) = -\sigma(-z)$ as the activation for the initialization algorithm, and we have $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(0))^2 \mathbb{1}\{z \geq 0\}] \geq C_1\epsilon$. Since $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(t))^2 \mathbb{1}\{z \geq t\}]$ is a decreasing function from $t = 0$ to $t = +\infty$ and $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(\bar{M}))^2 \mathbb{1}\{z \geq \bar{M}\}] = 0$, we know such $M$ must exist.

6. In fact, using the same argument in the bullet above, we can assume without loss of generality that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(-M))^2 \mathbb{1}\{z \leq -M\}] \leq \mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(M))^2 \mathbb{1}\{z \geq M\}]$, because if $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(-M))^2 \mathbb{1}\{z \leq -M\}] \geq \mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(M))^2 \mathbb{1}\{z \geq M\}]$ we can consider $\tilde{\sigma} = -\sigma(-z)$ and choose $M' \geq M$ that is the smallest

value such that $\mathbf{E}_{z\sim\mathcal{N}}[(\tilde{\sigma}(z)-\tilde{\sigma}(M))^2\mathbb{1}\{z\geq M\}]\geq C_1\epsilon$, Then $\mathbf{E}_{z\sim\mathcal{N}}[(\tilde{\sigma}(z)-\tilde{\sigma}(-M'))^2\mathbb{1}\{z\leq -M'\}]\leq \mathbf{E}_{z\sim\mathcal{N}}[(\tilde{\sigma}(z)-\tilde{\sigma}(M'))^2\mathbb{1}\{z\geq M'\}]$. Thus, in summary, $M$ satisfies the condition of Claim 7.5.7 and hence it is enough to initialize to $\mathbf{w}^0$ such that $\theta(\mathbf{w}^0,\mathbf{w}^*)\leq C/M$.

Given the assumptions above, we claim that there must exist $q'\in[0,B-\sigma(M)]$ such that

$$\mathbf{Pr}[\mathbb{1}\{y\geq\sigma(M)+q'\}\neq\mathbb{1}\{\sigma(z)\geq\sigma(M)+q'\}]\leq 4/\sqrt{C_1}\,\mathbf{Pr}[\sigma(z)\geq\sigma(M)+q']. \quad (7.35)$$

Suppose for the sake of contradiction that for any $q\in[0,B-\sigma(M)]$ it holds $\mathbf{Pr}[\mathbb{1}\{y\geq\sigma(M)+q\}\neq\mathbb{1}\{\sigma(z)\geq\sigma(M)+q\}]\geq 4/\sqrt{C_1}\,\mathbf{Pr}[\sigma(z)\geq\sigma(M)+q]$. Note that for $q\geq B-\sigma(M)$, since $\sigma(z)\leq B$ and $y\leq B$, we have $\mathbb{1}\{y\geq\sigma(M)+q\}=\mathbb{1}\{\sigma(z)\geq\sigma(M)+q\}=0$. Thus in summary, we have $\mathbf{Pr}[\mathbb{1}\{y\geq\sigma(M)+q\}\neq\mathbb{1}\{\sigma(z)\geq\sigma(M)+q\}]\geq 4/\sqrt{C_1}\,\mathbf{Pr}[\sigma(z)\geq\sigma(M)+q]$ for all $q\geq 0$.

Now let $f(z)=(\sigma(z)-\sigma(M))\mathbb{1}\{z\geq M\}$, $y'=(y-\sigma(M))\mathbb{1}\{y\geq\sigma(M)\}$. Then, for any $q\geq 0$,

$$\begin{aligned}
4/\sqrt{C_1}\,\mathbf{Pr}[\sigma(z)\geq\sigma(M)+q] &= 4/\sqrt{C_1}\,\mathbf{Pr}[f(z)\geq q] \\
&\leq \mathbf{Pr}[\mathbb{1}\{y\geq\sigma(M)+q\}\neq\mathbb{1}\{\sigma(z)\geq\sigma(M)+q\}] \\
&= \mathbf{Pr}[\mathbb{1}\{y'\geq q\}\neq\mathbb{1}\{f(z)\geq q\}].
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)-y')^2] = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[((\sigma(z)-\sigma(M))\mathbb{1}\{z\geq M\} - (y-\sigma(M))\mathbb{1}\{y\geq\sigma(M)\})^2] \\
&\leq 2\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z)-y)^2\mathbb{1}\{z\geq M\}] + 2\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(y-\sigma(M))^2(\mathbb{1}\{y\geq\sigma(M)\}-\mathbb{1}\{z\geq M\})^2] \\
&\leq 2\epsilon + 2\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(y-\sigma(M))^2(\mathbb{1}\{y\geq\sigma(M)\}-\mathbb{1}\{z\geq M\})^2].
\end{aligned}$$

Note that it holds $0 \leq (y-\sigma(M))\mathbb{1}\{y\geq\sigma(M), z<M\} \leq (y-\sigma(z))\mathbb{1}\{y\geq\sigma(M), z<M\}$ and $0 \leq (\sigma(M)-y)\mathbb{1}\{y<\sigma(M), z\geq M\} \leq (\sigma(z)-y)\mathbb{1}\{y<\sigma(M), z\geq M\}$. Therefore,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(y-\sigma(M))^2(\mathbb{1}\{y\geq\sigma(M)\}-\mathbb{1}\{z\geq M\})^2] \\
&= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(y-\sigma(z))^2\mathbb{1}\{y\geq\sigma(M), z<M\}] + \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(y-\sigma(z))^2\mathbb{1}\{y<\sigma(M), z\geq M\}] \leq 2\epsilon.
\end{aligned}$$

Combining with the upper bound on $\mathbf{E}_{z\sim\mathcal{N}}[(f(z)-y')^2]$ yields $\mathbf{E}_{z\sim\mathcal{N}}[(f(z)-y')^2] \leq 6\epsilon$. Hence the conditions of Lemma 7.4.22 are satisfied, and applying Lemma 7.4.22 we obtain $\mathbf{E}_{z\sim\mathcal{N}}[f^2] = \mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z)-\sigma(M))^2\mathbb{1}\{z\geq M\}] \leq 2(C_1/16)(6\epsilon) \leq (3/4)C_1\epsilon$. However, recall that $M$ is chosen such that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z)-\sigma(M))^2\mathbb{1}\{z\geq M\}] \geq C_1\epsilon$, therefore we have reached a contradiction.

Now let $M^* = \operatorname{argmin}\{0\leq M'\leq \bar{M} : \sigma(M') = \sigma(M)+q'\}$, $q'$ satisfying Equation (7.35), we have $M^* \in [M, \bar{M}]$. Note that this $q'$ can be found via a grid search on the interval $[0, B-\sigma(M)]$, as we can discretize the label $y$ and activation $\sigma$ using a $\sqrt{\epsilon}$ grid, therefore, there will only be $\mathrm{poly}(1/\sqrt{\epsilon}, B)$ number of possible choices of $q'$. The procedure is standard and we omit it here. The argument above implies that it hold $\mathrm{OPT}' = \mathbf{Pr}[\mathbb{1}\{y\geq\sigma(M)+q'\} \neq \mathbb{1}\{\sigma(z)\geq\sigma(M)+q'\}] \leq 4/\sqrt{C_1}\,\mathbf{Pr}[z\geq M^*] =$

$(1/C_2)\exp(-(M^*)^2/2)/M^*$ for $C_2$ being a large absolute constant. Hence the conditions of Fact 7.4.21 are satisfied, which then implies that there exists an algorithm that given labels $\tilde{y} = \mathbb{1}\{y \geq \sigma(M^*)\}$ and a target halfspace $\phi(\mathbf{w}^* \cdot \mathbf{x}; M^*)$, returns a vector $\mathbf{w}$ such that $\theta(\mathbf{w}, \mathbf{w}^*) \leq \min\{\pi/6, C_3 \exp((M^*)^2/2)\mathrm{OPT}'\} = \min\{\pi/6, C_3/M^*\} \leq \min\{\pi/6, C_3/M\}$. This completes the proof of Proposition 7.4.20. □

## 7.5 Discussion on Regular Activations

Let us first recall the definitions of the (Extended-)$(B, L)$-Regular activations.

**Definition 7.5.1** ($(B, L)$-Regular Activations)**.** *Given parameters $B, L > 0$, we define the class of $(B, L)$-Regular activations, denoted by $\mathcal{H}(B, L)$, as the class containing all functions $\sigma : \mathbb{R} \to \mathbb{R}$ such that 1) $\|\sigma\|_{L_\infty} \leq B$ and 2) $\|\sigma'\|_{L_2} \leq L$.*

*Given $\epsilon > 0$, we define the class of $\epsilon$-Extended $(B, L)$-Regular activations, denoted by $\mathcal{H}_\epsilon(B, L)$, as the class containing all activations $\sigma_1 : \mathbb{R} \to \mathbb{R}$ for which there exists $\sigma_2 \in \mathcal{H}(B, L)$ such that $\|\sigma_1 - \sigma_2\|_{L_2}^2 \leq \epsilon$.*

We remark that our algorithm can be applied to many non-differentiable activations.

**Remark 7.5.2** (On Differentiability)**.** In the definition of (Extended-)$(B, L)$-Regular activations, the differentiability of $\sigma$ is required. However, this restriction can be relaxed for any activation that is a locally-Lipschitz[2] function, since they are differentiable almost-everywhere [Fed69]. Therefore, since the set of non-differentiable points is measure-zero, we can define the derivative of $\sigma$ at those non-differentiable points freely (for example, using Clarke Differentials [Cla90]).

[2]We say a function $\sigma$ is locally-Lipschitz if for any $z_0 \in \mathbb{R}$, there exists positive reals $b$ and $\delta$ such that for any $z \in [z_0 - \delta, z_0 + \delta]$, it holds $|\sigma(z) - \sigma(z_0)| \leq b|z - z_0|$.

Furthermore, our results can also be applied to functions that are not even locally-Lipschitz. In particular, functions that have finite 'monotone jumps' like $\sigma(z) = \text{sign}(z-t)$ are subject to our results (Lemma 7.5.12).

In fact, the set of smoothed functions are dense to our functions (i.e., there always exists a $\rho \in (0,1)$ such that $\|\sigma - \mathrm{T}_\rho \sigma\|_{L_2}^2 \leq \epsilon$). Therefore, statistically, there is no difference in using either of the functions.

### 7.5.1 Rescaling to the Unit Sphere

Next, we comment on the impact of rescaling the activation $\sigma$.

**Remark 7.5.3** (Rescaling the Parameter)**.** Let $\sigma$ be a monotone (Extended-)$(B, L)$-regular activation. In our approach, it is without loss of generality to assume that $\|\mathbf{w}\|_2 = 1$. This is because, for any nonzero vector $\mathbf{w} \in \mathbb{B}(W)$, we can always rescale the activation $\sigma(\mathbf{w} \cdot \mathbf{x})$ to $\sigma(\|\mathbf{w}\|_2(\mathbf{w}/\|\mathbf{w}\|_2) \cdot \mathbf{x}) \stackrel{\text{def}}{=} \bar{\sigma}((\mathbf{w}/\|\mathbf{w}\|_2) \cdot \mathbf{x}) = \bar{\sigma}(\mathbf{w}' \cdot \mathbf{x})$ where $\|\mathbf{w}'\|_2 = 1$. In other words, we define $\bar{\sigma}(z) = \sigma(\|\mathbf{w}^*\|_2 z)$, where $\mathbf{w}^*$ is one of the target vectors.

After rescaling, the second moment of $\bar{\sigma}'$ increases to $\|\bar{\sigma}'\|_{L_2} \leq \|\mathbf{w}^*\|_2 \|\sigma'(\|\mathbf{w}^*\|_2 z)\|_{L_2}$ (which can be further bounded by using that $\sigma$ is close to a function $\hat{\sigma}$ with $\|\hat{\sigma}'\|_\infty \leq \|\mathbf{w}^*\|_2 B/\epsilon$). Therefore, the parameter $L$ can potentially scale with $W$.

For instance, if $\sigma$ is a $b$-Lipschitz activation, then the derivative of the rescaled function satisfies $|\bar{\sigma}'(z)| = \|\mathbf{w}^*\|_2 |\sigma'(\|\mathbf{w}^*\|_2 z)| \leq \|\mathbf{w}^*\|_2 b$ meaning that $\bar{\sigma}$ effectively becomes a $Wb$-Lipschitz activation. However, we emphasize that our approximation error obtained in Theorem 7.3.2 does not scale with any of these parameters $B, L$. These parameters only influence the sample complexity and runtime of our algorithm in a polynomial manner.

In the following lemma, we show that without loss of generality we can assume that we know the norm of the unknown vector $\mathbf{w}^*$ as we can reduce the problem into testing $1/\mathrm{poly}(\epsilon, 1/L, 1/B)$ different values for the norm.

**Lemma 7.5.4.** *Fix $\epsilon > 0$ and $\delta \in (0,1)$. Let $W > 0$ and let $\sigma$ be an activation such that for all $\lambda \in (0, W)$, $\sigma(\lambda z)$ is a $(B, L)$-Regular activation. Fix a unit vector $\mathbf{w}$ and assume that for some $0 < \lambda \leq W$, it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\lambda \mathbf{w} \cdot \mathbf{x}) - y)^2] \leq \epsilon$. Then, with $N = \mathrm{poly}(1/\epsilon, W, B, L)$ samples and $\mathrm{poly}(d, N)$ runtime, we can find $\lambda' > 0$ so that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\lambda' \mathbf{w} \cdot \mathbf{x}) - y)^2] \leq 4\epsilon$.*

*Proof.* Let $r = \mathrm{poly}(\epsilon, 1/L, 1/B)$ and we fix the following grid $(1+r)^k$, for $k = 1, \ldots, O(\log(W)/r)$. From Lemma 7.5.5, for some $k \leq O(\log(W)/r)$ it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\lambda \mathbf{w} \cdot \mathbf{x}) - \sigma((1+r)^k \mathbf{w} \cdot \mathbf{x}))^2] \leq 2\epsilon$. Hence, by testing all the possible choices and outputting the one with the minimum error suffices. This testing can be done with $\mathrm{poly}(1/\epsilon, W, B, L)$ samples. □

**Lemma 7.5.5.** *Let $\sigma$ be a $(B, L)$-Regular activation. Let $\epsilon > 0$ sufficiently small, then for $r \leq \mathrm{poly}(\epsilon, 1/L, 1/B)$, it holds that $\mathbf{E}_{z\sim\mathcal{N}(0,1)}[(\sigma(z) - \sigma((1+r)z))^2] \leq \epsilon$.*

*Proof.* From Lemma 7.1.10, it holds that $\mathbf{E}_{z\sim\mathcal{N}(0,1)}[(\sigma(z) - \mathrm{T}_\delta\sigma(z))^2] \leq 3(1-\delta)L^2$. Using Markov's inequality we have that

$$\mathbf{Pr}[|\sigma(z) - \mathrm{T}_\delta\sigma(z)| \geq \epsilon] \leq \frac{\mathbf{E}_{z\sim\mathcal{N}(0,1)}[(\sigma(z) - \mathrm{T}_\delta\sigma(z))^2]}{\epsilon^2} \leq \frac{3(1-\delta)L^2}{\epsilon^2} .$$

Furthermore, note that

$$\mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,1)}[|\sigma((1+r)z) - \mathrm{T}_\delta\sigma((1+r)z)| \geq \epsilon] = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,(1+r)^2)}[|\sigma(z) - \mathrm{T}_\delta\sigma(z)| \geq \epsilon] .$$

Furthermore, note that the total variation distance between two zero mean Guassians with variance $\sigma_1$ and $\sigma_2$ is bound from above by $\sqrt{\log(\sigma_1/\sigma_2) - \sigma_2^2/(2\sigma_1^2) - 1/2}$, for our case this is smaller than $2r$. Therefore, we have that

$$\Pr_{z\sim\mathcal{N}(0,1)}[|\sigma((1+r)z) - \mathrm{T}_\delta\sigma((1+r)z)| \geq \epsilon] \leq \Pr_{z\sim\mathcal{N}(0,1)}[|\sigma(z) - \mathrm{T}_\delta\sigma(z)| \geq \epsilon] + 2r \ .$$

Therefore,

$$\Pr_{z\sim\mathcal{N}(0,1)}[|\sigma((1+r)z) - \mathrm{T}_\delta\sigma((1+r)z)| \geq \epsilon] \leq \frac{3(1-\delta)L^2}{\epsilon^2} + 2r \ .$$

Combining, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}[(\sigma((1+r)z) - \mathrm{T}_\delta\sigma((1+r)z))^2] &\leq \epsilon^2 \Pr_{z\sim\mathcal{N}(0,1)}[|\sigma((1+r)z) - \mathrm{T}_\delta\sigma((1+r)z)| \leq \epsilon] \\ &\quad + B^2 \Pr_{z\sim\mathcal{N}(0,1)}[|\sigma((1+r)z) - \mathrm{T}_\delta\sigma((1+r)z)| \geq \epsilon] \\ &\leq \epsilon^2 + B^2(2r + \frac{3(1-\delta)L^2}{\epsilon^2}) \ .\end{aligned}$$

Furthermore, from Fact 7.1.7 it holds that $\mathrm{T}_\delta\sigma(z)$ is $B/(1-\delta^2)^{1/2}$-Lipschitz. Therefore, we have that

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}[(\mathrm{T}_\delta\sigma(z) - \mathrm{T}_\delta\sigma((1+r)z))^2] \leq \frac{B^2r^2}{(1-\delta^2)} \ . \tag{7.36}$$

Choosing $\delta$ so that $1-\delta, 1-\delta^2 \leq O(\epsilon^4/(B^2+L^2))$ and $r = \epsilon(1-\delta^2)/(B+1)$, we get that

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}[(\sigma(z) - \sigma((1+r)z))^2] \leq \epsilon \ .$$

□

### 7.5.2 Truncating the Regular Activations

Next, we observe that for any $\sigma \in \mathcal{H}_\epsilon(B, L)$, one can assume without loss of generality that the labels $y$ are bounded and the support of $\sigma'$ is also bounded. First, we show that we can truncate the labels $y$ without loss of generality.

**Claim 7.5.6.** *Let $\sigma$ be a $(B, L)$-Regular activation. Let $\bar{y} = \operatorname{sign}(y)\min\{|y|, B\}$. Then, $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\bar{y} - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq \mathrm{OPT}$. Furthermore, for any $\widehat{\mathbf{w}}$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\bar{y} - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] \leq O(\mathrm{OPT})$, we have $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] \leq O(\mathrm{OPT})$. Hence it is w.l.o.g. to assume that $|y| \leq B$.*

*Proof.* Let $\Pi(u) = \operatorname{sign}(u)\min\{|u|, B\}$ be the projection operator projecting $u \in \mathbb{R}$ to the interval $[-B, B]$ and let $\bar{y} \stackrel{\text{def}}{=} \Pi(y)$. Since $|\sigma(z)| \leq B$ almost surely, we have $\Pi(\sigma(z)) = \sigma(z)$. Thus by the property of projection operators, we have $|y - \sigma(\mathbf{w}^* \cdot \mathbf{x})| \geq |\Pi(y) - \Pi(\sigma(\mathbf{w}^* \cdot \mathbf{x}))|$. Therefore, we have $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\bar{y} - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\Pi(y) - \Pi(\sigma(\mathbf{w}^* \cdot \mathbf{x})))^2] \leq \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq \mathrm{OPT}$. The arguments above shows that $\mathbf{w}^*$ is also an OPT solution when $y$ is truncated.

Now let $\widehat{\mathbf{w}}$ be a constant approximate solution with respect to the truncated labels: $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\bar{y} - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] \leq C\mathrm{OPT}$. Then, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] &\leq 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \bar{y})^2] + 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\bar{y} - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] \\
&\leq 4 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] + 4 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\bar{y} - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] + 2C\mathrm{OPT} \\
&\leq (8 + 2C)\mathrm{OPT}.
\end{aligned}$$

Therefore, $\widehat{\mathbf{w}}$ is also an absolute constant approximate solution with respect to the true labels $y$, thus, it is without loss of generality to consider the $L_2^2$ loss with the truncated

labels $\bar{y}$. $\square$

Finally, we show that for $\sigma \in \mathcal{H}(B, L)$, $\sigma$ can be truncated so that the support of $\sigma'$ can be bounded by $M < \infty$.

**Claim 7.5.7.** *Let $\sigma$ be a $(B, L)$-Regular activation. Then, there exists a function $\tilde{\sigma} \in \mathcal{H}(B, L)$ that satisfies $\|\tilde{\sigma} - \sigma\|_{L_2}^2 \leq \epsilon$ and such that the support of $\tilde{\sigma}'$ is bounded by*

$$M \leq \sqrt{2\log(4B^2/\epsilon) - \log\log(4B^2/\epsilon)}.$$

*Moreover, if $\widehat{\mathbf{w}}$ satisfies $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \tilde{\sigma}(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] \leq O(\mathrm{OPT}) + \epsilon$, then also $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] \leq O(\mathrm{OPT}) + \epsilon$. Thus, one can replace $\sigma$ with $\tilde{\sigma}$ and assume without loss of generality that the support of $\sigma'$ is bounded by $M$.*

*Proof.* Note that by choosing $M = \sqrt{2\log(4B^2/\epsilon) - \log\log(4B^2/\epsilon)}$, using Fact 7.1.5 we have

$$\mathbf{Pr}[|z| \geq M] \leq \frac{2\exp(-M^2/2)}{M} = \frac{(\epsilon/(4B^2))\sqrt{\log(4B^2/\epsilon)}}{\sqrt{2\log(4B^2/\epsilon) - \log\log(4B^2/\epsilon)}} \leq \frac{\epsilon}{4B^2}.$$

Let us define

$$\tilde{\sigma}(z) = \begin{cases} \sigma(z), \text{ when } |z| \leq M \\ \sigma(M), \text{ when } z \geq M \\ \sigma(-M), \text{ when } z \leq -M. \end{cases}$$

Then, since $\|\sigma\|_\infty \leq B$, we have $\|\tilde{\sigma}\|_\infty \leq B$, and it holds

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z) - \tilde{\sigma}(z))^2] = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z) - \tilde{\sigma}(z))^2 \mathbb{1}\{|z| \geq M\}] \leq 4B^2 \mathbf{Pr}[|z| \geq M] \leq \epsilon.$$

In addition, $\|\tilde{\sigma}'(z)\|_{L_2} = \|\sigma'(z)\mathbb{1}\{|z| \le M\}\|_{L_2} \le L$. In other words, there exists an activation $\tilde{\sigma} \in \mathcal{H}(B, L)$ such that $\|\tilde{\sigma} - \sigma\|_{L_2}^2 \le \epsilon$. Furthermore, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \tilde{\sigma}(\mathbf{w}^* \cdot \mathbf{x}))^2] &\le 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] + 2\,\mathbf{E}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - \tilde{\sigma}(\mathbf{w}^* \cdot \mathbf{x}))^2] \\ &\le C\mathrm{OPT} + \epsilon,\end{aligned}$$

Now let $\widehat{\mathbf{w}}$ satisfy $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \tilde{\sigma}(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] \le C\mathrm{OPT} + \epsilon$. We show that $\mathcal{L}(\widehat{\mathbf{w}}) \le O(\mathrm{OPT}) + \epsilon$. We only need to observe that

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] &\le 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \tilde{\sigma}(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] + 2\,\mathbf{E}[(\tilde{\sigma}(\widehat{\mathbf{w}} \cdot \mathbf{x}) - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] \\ &\le 2C\mathrm{OPT} + 4\epsilon.\end{aligned}$$

Hence we can replace $\sigma$ with $\tilde{\sigma} \in \mathcal{H}(B, L)$ and focus on the $L_2^2$ loss with respect to $\tilde{\sigma}$. Therefore, we can assume without loss of generality that $\sigma(z)$ is a constant when $|z| \ge M$, in other words, for any $|z| \ge M$, we have $\sigma'(z) = 0$, and the support of $\sigma'$ is indeed bounded by $M$. □

**Remark 7.5.8.** Since $M$ is an upper bound on the support of $\sigma'$, we will assume without loss of generality throughout the rest of the work that $M^2$ is larger than any absolute constant $C$.

### 7.5.3 Examples of Regular Activations

We now show that $\mathcal{H}_\epsilon(B, L)$ contains a wide range of activations. First, we show that all monotone functions with bounded $2 + \zeta$-moment are Extended Regular activations:

**Lemma 7.5.9.** *If $\sigma$ satisfies $\mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)^{2+\zeta}] \leq B_\sigma$ for some $\zeta > 0$ and $\sigma$ is monotone, then $\sigma \in \mathcal{H}_\epsilon(c_1 D, c_2 D^4/\epsilon^2)$ where $D = (B_\sigma/4\epsilon)^{1/\zeta}$ and $c_1, c_2$ are absolute constants.*

*Proof.* For some $\zeta > 0$, we have $\mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)^{2+\zeta}] \leq B_\sigma$. From Markov's inequality, we have that

$$\mathbf{Pr}[|\sigma(z)| \geq T] \leq \frac{\mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)^{2+\zeta}]}{T^{2+\zeta}} \leq \frac{B_\sigma}{T^{2+\zeta}}.$$

Note that $\mathbf{E}[\sigma^2(z)] = \int_0^\infty \mathbf{Pr}[\sigma^2(z) \geq t]\mathrm{d}t = \int_0^\infty 2u\,\mathbf{Pr}[\sigma^2(z) \geq u^2]\mathrm{d}u$ (the last part is after change of variables to $u^2 = t$). Therefore, we have that

$$\begin{aligned}
\mathbf{E}[\sigma^2(z)\mathbb{1}\{|\sigma(z)| \geq D\}] &= \int_0^\infty 2u\,\mathbf{Pr}[\sigma^2(z)\mathbb{1}\{|\sigma(z)| \geq D\} \geq u^2]\mathrm{d}u \\
&= \int_0^\infty 2u\,\mathbf{Pr}[|\sigma(z)|\mathbb{1}\{|\sigma(z)| \geq D\} \geq u]\mathrm{d}u \\
&= \int_D^\infty 2u\,\mathbf{Pr}[|\sigma(z)| \geq u]\mathrm{d}u \leq \int_D^\infty 2u\frac{B_\sigma}{u^{2+\zeta}} \leq 4\frac{B_\sigma}{D^\zeta}\,.
\end{aligned}$$

Set $D = (B_\sigma/4\epsilon)^{1/\zeta}$ and let $\bar{\sigma}(z) = \mathrm{sign}(\sigma(z))\min\{|\sigma(z)|, D\}$. We show that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \bar{\sigma}(z))^2] \leq \epsilon$:

$$\begin{aligned}
\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z) - \bar{\sigma}(z))^2] &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z) - \bar{\sigma}(z))^2\mathbb{1}\{|\sigma(z)| \geq D\}] \\
&\leq \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z))^2\mathbb{1}\{|\sigma(z)| \geq D\}] \leq \epsilon\,.
\end{aligned}$$

Therefore, $\sigma$ is $\epsilon$-close to a $\bar{\sigma}$ with $\|\bar{\sigma}\|_\infty \leq (B_\sigma/4\epsilon)^{1/\zeta}$.

It remains to show that the activation $\bar{\sigma}$ is also $\epsilon$-close to an activation with bounded $\|\sigma'\|_{L_2}$. Without loss of generality we assume that $\bar{\sigma}(z) \geq 0$ and $\bar{\sigma}(z) \in [0, 2D]$, because we can just add assume that the function if $\bar{\sigma}(z)' = \bar{\sigma}(z) + D$. Note that it holds that

$\bar{\sigma}(z) = \int_0^{2D} \mathbb{1}\{\bar{\sigma}(z) \geq t\}\mathrm{d}t$. It suffices to show that there exists a parameter $\rho$ so that $\|\bar{\sigma}(z) - \mathrm{T}_\rho\bar{\sigma}(z)\|_{L_2}^2 \leq (1-\rho^2)\mathrm{poly}(D)$. We have that

$$\begin{aligned}
\|\bar{\sigma}(z) - \mathrm{T}_\rho\bar{\sigma}(z)\|_{L_2}^2 &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\int_0^{2D} \mathbb{1}\{\bar{\sigma}(z) \geq t\} - \mathbb{1}\{\mathrm{T}_\rho\bar{\sigma}(z) \geq t\}\,\mathrm{d}t\right)^2\right] \\
&\leq \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\int_0^{2D} |\mathbb{1}\{\bar{\sigma}(z) \geq t\} - \mathbb{1}\{\mathrm{T}_\rho\bar{\sigma}(z) \geq t\}|\,\mathrm{d}t\right)^2\right] \\
&\leq 2D \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\int_0^{2D} \left(\mathbb{1}\{\bar{\sigma}(z) \geq t\} - \mathbb{1}\{\mathrm{T}_\rho\bar{\sigma}(z) \geq t\}\right)^2 \mathrm{d}t\right] \\
&= 2D \int_0^{2D} \|\mathbb{1}\{\bar{\sigma}(z) \geq t\} - \mathrm{T}_\rho\mathbb{1}\{\bar{\sigma}(z) \geq t\}\|_{L_2}^2 \mathrm{d}t\;,
\end{aligned}$$

where we used the Jensen's inequality $((1/(b-a))\int_a^b f(z)\,\mathrm{d}z)^2 \leq (1/(b-a))\int_a^b f^2(z)\,\mathrm{d}z$ for positive functions $f$ and we exchange the integrals using the Fubini's theorem. Note that because the function $\bar{\sigma}(z)$ is monotone, then there exists a function $q(z)$ so that $\mathbb{1}\{\bar{\sigma}(z) \geq t\} = \mathbb{1}\{z \geq q(t)\}$. Therefore, using this transformation, it suffices to bound the difference

$$\begin{aligned}
\|\mathbb{1}\{z \geq q(t)\} - \mathrm{T}_\rho\mathbb{1}\{z \geq q(t)\}\|_{L_2}^2 &\leq \mathop{\mathbf{E}}_{x,z\sim\mathcal{N}(0,1)}[(\mathbb{1}\{x \geq q(t)\} - \mathbb{1}\{x\rho + z(1-\rho^2)^{1/2} \geq q(t)\})^2] \\
&\leq 4(1-\rho^2)^{1/2}\;,
\end{aligned}$$

where in the first inequality we used Jensen, and in the second one we used that $\mathbf{E}[|\mathrm{sign}(\mathbf{w}\cdot\mathbf{x} + t) - \mathrm{sign}(\mathbf{v}\cdot\mathbf{x} + t)| \leq \theta(\mathbf{v},\mathbf{u})$ for any two unit vectors $\mathbf{v},\mathbf{w}$ (see Fact C.11 of [DKTZ22b]). Hence, we have that

$$\int_0^{2D} \|\mathbb{1}\{\bar{\sigma}(z) \geq t\} - \mathrm{T}_\rho\mathbb{1}\{\bar{\sigma}(z) \geq t\}\|_{L_2}^2 \mathrm{d}t \leq 8D(1-\rho^2)^{1/2}\;.$$

Therefore, the function $\|\bar{\sigma}(z) - \mathrm{T}_\rho\bar{\sigma}(z)\|_2^2 \leq \epsilon$ for $\rho = \sqrt{1-(\epsilon/(16D^2))^2}$. That means

that $\|(\mathrm{T}_\rho\bar{\sigma}(z))'\|_{L_2}^2 \le 16^2D^4/\epsilon^2$ (cf. Fact 7.1.7(c)). Thus, we conclude that $\sigma \in \mathcal{H}_\epsilon(2D, 16^2D^4/\epsilon^2)$. □

Now let us define a special type of activation that has an 'exponential-tail' property. We will show in Lemma 7.5.12 that all $b$-Lipschitz functions are such kind of activations.

**Definition 7.5.10** ($(R,r)$-Sub-exponential Activations)**.** *We say that an activation $\sigma(z)$ is $(R,r)$-sub-exponential for some positive constants $R, r$, if for any $p > 0$, we have $(\mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)^p])^{1/p} \le Rp^r$.*

We will make use of the following fact in the proof of Lemma 7.5.12.

**Fact 7.5.11** ([Ver18] Theorem 5.2.2)**.** *Let $z \sim \mathcal{N}(0,1)$ and let $\sigma$ be a $b$-Lipschitz function. Then, $\sigma(z)$ is a sub-Gaussian random variable with $\|\sigma(z)\|_{\psi_2} \le cb$, where $\|\cdot\|_{\psi_2}$ is the Orlicz 2-norm and $c$ is an absolute constant.*

We show that all the following function classes belong to the Extended Regular activation class:

**Lemma 7.5.12.** *All of the following activations are $\epsilon$-Extended $(B,L)$-Regular.*

1. *If $\sigma$ satisfies $\mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)^4] \le B_{\sigma,4}$ and $\|\sigma'\|_{L_2} \le L$, then $\sigma \in \mathcal{H}(\sqrt{B_{\sigma,4}/\epsilon}, L)$.*
2. *If $\sigma$ is $(R,r)$-Sub-exponential and $\|\sigma'\|_{L_2} \le L$, then $\sigma \in \mathcal{H}_\epsilon(cR(r+\log(R/\epsilon))^r, L)$, where $c$ is an absolute constant.*
3. *If $\sigma$ is $b$-Lipschitz, then $\sigma \in \mathcal{H}_\epsilon(cb\log^{1/2}(b/\epsilon), b)$, where $c$ is an absolute constant.*
4. *If $\sigma = \sigma_1 + \Phi$, where $\sigma_1 \in \mathcal{H}_\epsilon(B,L)$, $|\Phi(z)| \le A$, $\Phi \in \mathcal{F}_M$ (recall Definition 7.4.5), i.e.,*

$$\Phi(z) = \sum_{i=1}^{m} A_i\phi(z;t_i) + A_0 : A_0 \in \mathbb{R}; A_i > 0, |t_i| \le M, \forall i \in [m]; m < \infty$$

*then* $\sigma \in \mathcal{H}_\epsilon(B + A, L + \max\{A^2 L/\sqrt{\epsilon}, A^4/\epsilon\})$.

*Proof.* We prove each claim in order.

1. Suppose first that $\mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)^4] \le B_{\sigma,4}$. Let $\bar{\sigma}(z) = \text{sign}(\sigma(z)) \min\{|\sigma(z)|, \sqrt{B_{\sigma,4}/\epsilon}\}$, which is an activation in the $(B, L)$-Regular class. We show that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \bar{\sigma}(z))^2] \le \epsilon$:

$$\begin{aligned}
\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z) - \bar{\sigma}(z))^2] &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z) - \bar{\sigma}(z))^2 \mathbb{1}\{|\sigma(z)| \ge \sqrt{B_{\sigma,4}/\epsilon}\}] \\
&\le \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z))^2 \mathbb{1}\{|\sigma(z)| \ge \sqrt{B_{\sigma,4}/\epsilon}\}] \\
&\le \sqrt{\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\sigma^4(z)] \, \mathbf{Pr}[|\sigma(z)| \ge \sqrt{B_{\sigma,4}/\epsilon}]}.
\end{aligned}$$

By Markov's inequality we have

$$\mathbf{Pr}[|\sigma(z)| \ge \sqrt{B_{\sigma,4}/\epsilon}] = \mathbf{Pr}[\sigma^2(z) \ge B_{\sigma,4}/\epsilon] \le \frac{\mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)^4]\epsilon^2}{B_{\sigma,4}^2} \le \frac{\epsilon^2}{B_{\sigma,4}}. \tag{7.37}$$

Therefore, the $L_2^2$ difference between $\sigma$ and $\bar{\sigma}$ is bounded above by

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z) - \bar{\sigma}(z))^2] \le \sqrt{B_{\sigma,4} \frac{\epsilon^2}{B_{\sigma,4}}} \le \epsilon.$$

Therefore, $\sigma$ is an Extended $(\sqrt{B_{\sigma,4}/\epsilon}, L)$-Regular activation.

2. Next, assume that $\sigma$ is $(R, r)$-Sub-exponential. Similarly, let

$$\bar{\sigma}(z) = \text{sign}(\sigma(z)) \min\{|\sigma(z)|, eR(4r\log(4) + \log(R^4/\epsilon^2))^r\}$$

Denote for simplicity $B_\sigma \stackrel{\text{def}}{=} eR(4r\log(4) + \log(R^4/\epsilon^2))^r$ Then, $\bar{\sigma}$ is a $(B_\sigma, L)$-Regular

activation. Using the same arguments as above, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z)-\bar{\sigma}(z))^2] &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z)-\bar{\sigma}(z))^2\mathbb{1}\{|\sigma(z)|\geq B_\sigma\}] \\ &\leq \sqrt{\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\sigma^4(z)]\,\mathbf{Pr}[|\sigma(z)|\geq B_\sigma]}.\end{aligned}$$

Now since $\sigma$ is a $(R,r)$-Sub-exponential activation, we have

$$\mathbf{Pr}[|\sigma(z)|\geq t]\leq \frac{\mathbf{E}_{z\sim\mathcal{N}}[\sigma^p(z)]}{t^p}\leq \left(\frac{Rp^r}{t}\right)^p.$$

Choosing $p=(t/(Re))^{1/r}$, we get

$$\mathbf{Pr}[|\sigma(z)|\geq t]\leq \exp\left(-(t/(Re))^{1/r}\right).$$

Let $t=B_\sigma=Re(4r\log(4)+\log(R^4/\epsilon^2))^r$, then it holds

$$\mathbf{Pr}[|\sigma(z)|\geq B_\sigma]\leq \exp\left(-\log(4^{4r}\epsilon^2/R^4)\right)=\frac{\epsilon^2}{R^4 4^{4r}}.$$

Furthermore, since $\mathbf{E}_{z\sim\mathcal{N}}[\sigma^4]\leq R^4 4^{4r}$, we thus obtain

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\sigma(z)-\bar{\sigma}(z))^2]\leq\epsilon.$$

3. Next, suppose $\sigma$ is $b$-Lipschitz. Then, since $|\sigma'|\leq b$, we have $\|\sigma'\|_{L_2}\leq b$. Next, we show that $\sigma$ is $(b,1/2)$-Sub-exponential. Note that it is without loss of generality to assume that $\mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)]=0$, since we can always consider shifting the activation $\sigma$ and the labels $y$ to $\sigma(z)-\mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)]$, $y-\mathbf{E}_{z\sim\mathcal{N}}[\sigma(z)]$ and obtaining the same results. Since $z\sim\mathcal{N}(0,1)$, we can use Fact 7.5.11, which yields that $\sigma(z)$ is sub-Gaussian

with sub-Gaussian constant $\|\sigma(z)\|_{\psi_2} = cb$. Because $\sigma(z)$ is sub-Gaussian, we know $\|\sigma(z)\|_{L_p} \le c\|\sigma(z)\|_{\psi_2} p^{1/2} \le cbp^{1/2}$, this implies that $\sigma$ is a $(b, 1/2)$-Sub-exponential activation. Using the previous result (Part 2), we immediately obtain that $\sigma$ is an Extended $(cb\log^{1/2}(b/\epsilon), b)$-Regular activation.

4. Finally, consider $\sigma = \sigma_1 + \Phi$ where $\sigma_1 \in \mathcal{H}_\epsilon(B, L)$, $\Phi \in \mathcal{F}_M$, $|\Phi(z)| \le A$. Let $\tilde{\sigma} = \mathrm{T}_{1-\epsilon_0}\sigma_1 + \mathrm{T}_{1-\epsilon_0}\Phi$, where $\epsilon_0 \le \min(\epsilon/L^2, 1/M^2, (\epsilon/A^2)^2)$, and $M$ is defined as in Claim 7.5.7. Then we have

$$\|\tilde{\sigma} - \sigma\|_{L_2}^2 \le 2\|\mathrm{T}_{1-\epsilon_0}\sigma_1 - \sigma_1\|_{L_2}^2 + 2\|\mathrm{T}_{1-\epsilon_0}\Phi - \Phi\|_{L_2}^2.$$

By Lemma 7.1.10, we have $\|\mathrm{T}_{1-\epsilon_0}\sigma_1 - \sigma_1\|_{L_2}^2 \lesssim \epsilon_0\|\sigma_1'\|_{L_2}^2 \le \epsilon$. In addition, applying Proposition 7.4.8, since $\epsilon_0 \le 1/M^2$, we have $\|\mathrm{T}_{1-\epsilon_0}\Phi - \Phi\|_{L_2}^2 \lesssim \epsilon_0\|\mathrm{T}_{1-\epsilon_0}\Phi'(z)\|_{L_2}^2$. Note that $\max_{z\in\mathbb{R}}\Phi(z) = \sum_{i=1}^m A_i \le A$. Then, by Lemma 7.4.9, we have

$$\|\mathrm{T}_{1-\epsilon_0}\Phi'(z)\|_{L_2}^2 \lesssim (1/\sqrt{\epsilon_0})\sum_{i,j=1}^m A_iA_j = (1/\sqrt{\epsilon_0})A^2 \le \max\{A^2L/\sqrt{\epsilon}, A^4/\epsilon\}.$$

Therefore, we further obtain $\|\mathrm{T}_\epsilon\Phi - \Phi\|_{L_2}^2 \le \sqrt{\epsilon_0}A^2 \le \epsilon$. This implies that $\|\tilde{\sigma} - \sigma\|_{L_2}^2 \lesssim \epsilon$. Furthermore, observe that $\|\tilde{\sigma}\|_{L_\infty} \le \|\sigma_1\|_{L_\infty} + \|\Phi\|_{L_\infty} \le B + A$, and $\|\tilde{\sigma}'\|_{L_2} \le \|\mathrm{T}_{1-\epsilon_0}\sigma'\|_{L_2} + \|\mathrm{T}_{1-\epsilon_0}\Phi'\|_{L_2} \le L + \sqrt{\epsilon}$. Thus, we conclude that $\tilde{\sigma} \in \mathcal{H}_\epsilon(B + A, L + \max\{A^2L/\sqrt{\epsilon}, A^4/\epsilon\})$. □

### 7.5.4 Required Assumptions on Activation

Here we point out that it is information-theoretically impossible to learn all monotone activations, in the realizable setting, if we only assume that $\sigma \in L_2(\mathcal{N})$—even if we further assume that $\|\sigma\|_{L_2} \le 1$. The intuition behind this fact is the following: there

exists a monotone function $\sigma$, which is equal to 0 everywhere except in the tails of a direction $\mathbf{v}$. That means that in order to see a point where the labels are non-zero, we need to see a label from the tails.

We show that for any choice of the threshold in the tails, i.e., $\mathbf{Pr}_{z\sim\mathcal{N}}[z \geq t]$, there exists a monotone function that has $\|\sigma\|_{L_2} = 1$ and for any unit vectors $\mathbf{v}, \mathbf{u}$ with $\|\mathbf{v} - \mathbf{u}\|_2 = \Omega(1)$, we have $\|\sigma(\mathbf{v} \cdot \mathbf{x}) - \sigma(\mathbf{u} \cdot \mathbf{x})\|_{L_2}^2 = \Omega(1)$. Formally, we show that:

**Theorem 7.5.13** (Impossibility of Learning All Monotone Functions)**.** *Consider the class $\mathcal{F}$ consisting of all monotone activations $\sigma \in L_2(\mathcal{N})$ satisfying $\|\sigma\|_{L_2} \leq 1$. There is no finite-sample algorithm that realizably learns $\mathcal{F}$ up to error $1/8$.*

*Proof.* Let $\gamma^{-1}(\delta) = \sup_t\{t : \mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[z \geq t] \leq \delta\}$. Let $\delta < 1/16$ and consider the following function $\sigma(t) = (1/\sqrt{\delta})\mathbb{1}\{t \geq \gamma^{-1}(\delta)\}$. Note that this function belongs to the class $\sigma \in L_2(\mathcal{N})$ with $\|\sigma\|_{L_2} \leq 1$, since

$$\mathbf{E}[\sigma^2(z)] = 1/\delta \, \mathbf{Pr}[t \geq \gamma^{-1}(\delta)] = 1 \, .$$

Consider a set of unit vectors $V$ such that for any $\mathbf{u}, \mathbf{v} \in V$ we have $\|\mathbf{u} - \mathbf{v}\|_2 \geq 1/2$. By standard packing arguments, there exists such a set of size $2^{\Theta(d)}$. Let $\theta \stackrel{\text{def}}{=} \theta(\mathbf{u}, \mathbf{v})$. Then we have $\|\mathbf{u} - \mathbf{v}\|_2 = 2\sin(\theta/2) \geq 1/2$, hence $\cos\theta = 1 - 2\sin^2(\theta/2) \leq 7/8$. Note that for any $\mathbf{u}, \mathbf{v} \in V$, with $\mathbf{u} \neq \mathbf{v}$, it holds that

$$\begin{aligned}
\|\sigma(\mathbf{v} \cdot \mathbf{x}) - \sigma(\mathbf{u} \cdot \mathbf{x})\|_{L_2}^2 &= 2(1 - \mathbf{E}[\sigma(\mathbf{v} \cdot \mathbf{x})\sigma(\mathbf{u} \cdot \mathbf{x})]) \\
&= 2\sum_{k\geq 0}(1 - \cos^k\theta)\hat{\sigma}(i)^2 \\
&\geq 2(1 - \hat{\sigma}(0)^2) - \cos\theta \sum_{k\geq 1}\hat{\sigma}(i)^2 \geq 2(1 - \hat{\sigma}(0)^2) - 2\cos\theta\|\sigma\|_{L_2}^2 \, ,
\end{aligned}$$

where $\hat{\sigma}(i)$ are the Hermite coefficients of $\sigma$. Furthermore, note that $\hat{\sigma}(0) = \mathbf{E}[\sigma(z)] = \sqrt{\delta}$. Hence, we have that

$$\|\sigma(\mathbf{v}\cdot\mathbf{x}) - \sigma(\mathbf{u}\cdot\mathbf{x})\|_{L_2}^2 \geq 2(1-\delta-7/8) \geq 1/4 - 2\delta \geq 1/8 \; .$$

Intuitively, in order to learn up to error $\epsilon < 1/2$, we need to see at least one sample $\mathbf{x}$ such that $\sigma(\mathbf{v}\cdot\mathbf{x}) > 0$, which happens with probability $\delta$. Since $\delta$ can be selected to be an arbitrarily small positive number, by taking $\delta \to 0$, we see that in order to observe one sample, we need $\Omega(1/\delta)$ samples; if we choose $\mathbf{v}$ at random, we succeed with probability at most $\exp(-cd)$.

One way to formalize the argument is to reduce the problem of learning Gaussian halfspaces to the above task. Consider the following transformation: $\forall \mathbf{v} \in V$, let $y'_{\mathbf{v}} = 1$ when $\sigma(\mathbf{v}\cdot\mathbf{x}) = 1/\sqrt{\delta}$ and $y'_{\mathbf{v}} = -1$ otherwise. This gives an instance of learning halfspaces under the Gaussian distribution. We have that

$$\|\sigma(\mathbf{v}\cdot\mathbf{x}) - \sigma(\mathbf{u}\cdot\mathbf{x})\|_{L_2}^2 = \mathbf{Pr}[\mathrm{sign}(\mathbf{v}\cdot\mathbf{x} - \gamma^{-1}(\delta)) \neq \mathrm{sign}(\mathbf{u}\cdot\mathbf{x} - \gamma^{-1}(\delta))]/\delta \; .$$

Therefore, in order to get error of $1/8$ for our monotone GLM learning task, we need to learn halfspaces with accuracy better than $\delta/8$. This task is known to have a sample complexity lower bound of $\Omega(d/\delta)$. Therefore, since an algorithm that learns $\sigma(\mathbf{v}\cdot\mathbf{x})$ with error better than $1/8$ will also learn halfspaces, it follows that achieving error better than $1/8$ requires $\gtrsim d/\delta$ samples. As $\delta \to 0$, the number of samples becomes unbounded. If we output a function at random, the probability of success is lower bounded by the number of elements in $|V|$, which gives the result. $\square$

# Chapter 8

# Robustly Learning Monotone SIMs under Gaussian Distribution

## 8.1 Introduction

In this part of the thesis, we extend the results of Chapter 7, for the SIM case, i.e., the setting where the monotone activation function $\sigma$ is unknown. While the previous chapter provided a robust learner for GLMs with a known activation, those methods fundamentally rely on access to the derivative $\sigma'$. Here, we develop a novel algorithmic framework that does not require such knowledge, yielding the first efficient, constant-factor approximate learner for monotone SIMs under the Gaussian distribution.

A key challenge in the unknown activation setting is that gradient-based methods, which depend on the activation's derivative, are no longer applicable. To circumvent this, we depart from traditional optimization of a loss function. Instead, our approach directly constructs an update direction using a novel **spectral subroutine** that acts as a robust proxy for the gradient.

Our main result (Theorem 8.1.1) establishes that for any monotone Lipschitz activation, our algorithm returns a hypothesis $(\hat{\sigma}, \hat{\mathbf{w}})$ with squared error $\mathcal{L}_2(\hat{\mathbf{w}}; \hat{\sigma}) \leq C \cdot \mathrm{OPT} + \epsilon$, where $C$ is a universal constant.

In our context, the target class $\mathcal{C}$ will refer to the class of SIMs with unknown activation and parameter vector, namely functions of the form $f(\mathbf{x}) = \sigma(\mathbf{w} \cdot \mathbf{x})$. The only assumptions we make are that the norm of the parameter (weight) vector is bounded, i.e., $\|\mathbf{w}\|_2 \leq W$ for some $W > 0$, and that the activation function belongs to a known family of structured monotone functions. For a pair $(\mathbf{w}, \sigma)$ defining the SIM $f(\mathbf{x}) = \sigma(\mathbf{w} \cdot \mathbf{x})$, we write $\mathcal{L}_2(\mathbf{w}; \sigma) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2]$ to denote the squared error of $f$.

### Main Results

**Theorem 8.1.1** (Robustly Learning Monotone & Lipschitz SIMs)**.** *There exists a universal constant $C > 1$ such that the following holds. Let $\mathcal{C}$ be the class of all SIMs on $\mathbb{R}^d$ with a monotone and $L$-Lipschitz activation. There is an algorithm that, given $\epsilon > 0$ and $W > 0$, draws $N = d^2\mathrm{poly}(1/\epsilon, W, L)$ samples, runs in $\mathrm{poly}(N)$ time, and returns a predictor $(\widehat{\sigma}, \widehat{\mathbf{w}})$ such that, with high probability, $\mathcal{L}_2(\widehat{\mathbf{w}}; \widehat{\sigma}) \leq C\mathrm{OPT} + \epsilon$.*

It is worth noting that our algorithm does not require the Lipschitz assumption on the unknown activation. In fact, it applies for the broader class of monotone activations with bounded $(2 + \zeta)$ moment, for any $\zeta > 0$ (Corollary 8.2.7).

**Comparison to Prior Work** The most directly related prior work is [ZWDD25], which gave an efficient constant-factor approximate learner for the *known activation* version of our problem (i.e., for a known monotone activation with bounded $(2 + \zeta)$ moment). Independently, [GV24] developed an efficient constant-factor learner for the special case

of a general (biased) ReLU. Interestingly, even for the special case of known activation (e.g., a general LTF or ReLU), obtaining an efficient constant-factor approximation for more general structured distributions (beyond the Gaussian) remains open.

Regarding the SIM version of the problem, prior work either did not achieve a constant-factor approximation or succeeded for a significantly smaller class of activations. Specifically, [GGKS23] gave an efficient SIM learner for monotone 1-Lipschitz activations, for any distribution with bounded second moment, with error guarantee $O(W\sqrt{\mathrm{OPT}})+\epsilon$ (under the technical assumption that the labels are bounded in $[0,1]$). In addition to the sub-optimal dependence on OPT, the multiplicative factor inside the big-O scales with the radius $W$ of the space. Subsequently, [ZWDD24] developed an efficient constant-factor SIM learner under structured distributions (including the Gaussian), albeit for a much smaller family of activations. Specifically, for parameters $a, b > 0$, the activation family considered in [ZWDD24] contains all functions $\sigma : \mathbb{R} \to \mathbb{R}$ such that $|\sigma'(z)| \leq b$ everywhere and $\sigma'(z) \geq a > 0$ for $z \geq 0$. The final error bound of the algorithm in [ZWDD24] is $O\big(\mathrm{poly}(b/a)\big)\mathrm{OPT} + \epsilon$. This guarantee is vacuous as $a \to 0$, i.e., for the class of monotone $b$-Lipschitz functions.

## 8.2 Agnostic Learning SIMs

We consider distributions $\mathcal{D}$ over $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and predictors of the form $f_{\mathbf{w},\sigma}(\mathbf{x}) = \sigma(\mathbf{w} \cdot \mathbf{x})$, where $\|\mathbf{w}\| \leq W$ and $\sigma$ is monotone. We assume that the target activation $\sigma$ is $\epsilon$ close to a $(B, L)$-Regular that satisfies the following conditions:

**Definition 8.2.1** ($(B, L)$-Regular Monotone Activations)**.** *Given parameters $B, L > 0$, we define the class of $(B, L)$-Regular activations, denoted by $\mathcal{H}(B, L)$, as the class*

*containing all functions* $\sigma : \mathbb{R} \to \mathbb{R}$ *such that 1)* $\|\sigma\|_{L_\infty} \leq B$ *and 2)* $\|\sigma'\|_{L_2} \leq L$*. Given* $\epsilon > 0$*, we define the class of* $\epsilon$*-Extended* $(B, L)$*-Regular activations, denoted by* $\mathcal{H}_\epsilon(B, L)$*, as the class containing all activations* $\sigma_1 : \mathbb{R} \to \mathbb{R}$ *for which there exists* $\sigma_2 \in \mathcal{H}(B, L)$ *such that* $\|\sigma_1 - \sigma_2\|_{L_2}^2 \leq \epsilon$*.*

We can show that all monotone activations in $L_2(\mathcal{N})$ are close to a regular activations.

**Lemma 8.2.2.** *Let* $f$ *be a monotone activation in* $L_2(\mathcal{N})$*. Then, for any* $\epsilon > 0$*, there exists* $C(\epsilon) > 0$ *so that* $f \in \mathcal{H}_{\epsilon/2}(\text{poly}(C(\epsilon)/\epsilon), \text{poly}(C(\epsilon)/\epsilon))$*.*

*Proof.* Using the assumption that $f \in L_2(\mathcal{N})$, we have that $\|f\|_{L_2}^2 \leq c < \infty$ for some $c > 0$. Therefore, we have that

$$\|f\|_{L_2}^2 = \int_0^\infty \mathbf{Pr}[f^2(z) \geq t] \mathrm{d}t \leq c \, .$$

Note that the function $\mathbf{Pr}[f^2(z) \geq t]$ is a nonnegative function of $t$. Therefore the sequence $a_n = \int_0^n \mathbf{Pr}[f^2(z) \geq t] \mathrm{d}t$ is non-decreasing for any $n \in N$ and by assumption the limit of the sequence $a_n$ as $n \to \infty$ exists. Therefore from the definition of the convergence of the limits, for any $\epsilon' \in (0, 1)$ there exists $n_{\epsilon'} \in \mathbb{N}$ so that for any $n \geq n_{\epsilon'}$, we have

$$\int_n^\infty \mathbf{Pr}[f^2(z) \geq t] \mathrm{d}t = \left| \int_0^\infty \mathbf{Pr}[f^2(z) \geq t] \mathrm{d}t - \int_0^n \mathbf{Pr}[f^2(z) \geq t] \mathrm{d}t \right| \leq \epsilon' \, .$$

Furthermore, note that for $f_2 = \text{sign}(f)\min(|f|, |f(n_{\epsilon'})|)$, we have that

$$\begin{aligned}\|f - f_2\|_{L_2}^2 &= \int_0^\infty \mathbf{Pr}[|f(z) - f_2(z)|^2 \geq t]\mathrm{d}t \\ &= \int_0^\infty \mathbf{Pr}[|f(z) - f_2(z)|^2 \geq t, |f| \geq |f(n_\epsilon)|]\mathrm{d}t \\ &\leq \int_n^\infty \mathbf{Pr}[f(z)^2 \geq t]\mathrm{d}t \leq \epsilon' \,.\end{aligned}$$

By applying Part 1 of Lemma 7.5.12 to $f_2$ and choose $\epsilon' = \epsilon/2$, we get the result for $C(\epsilon) = |f_2(n_{\epsilon'})|$. □

**Remark 8.2.3.** Instead of directly enforcing a norm bound $\|\mathbf{w}\| \leq W$, one can assume that $\mathcal{H}_\epsilon(B, L)$ that we compete against is chosen so that it contains all the activations $\sigma(z) \mapsto \sigma(\lambda z)$ for all $\lambda \leq W$. This lets us focus on the core statistical challenge without separately tracking a norm constraint on $\mathbf{w}$.

Some important facts about Definition 8.2.1.

**Truncation of Regular Activations** For a target activation $\sigma \in \mathcal{H}_\epsilon(B, L)$, we can make simplifying assumptions that come at no loss of generality. The first assumption is that there exists a finite constant $\bar{M}$ such that outside the interval $[-\bar{M}, \bar{M}]$, the derivative $\sigma'$ of the target activation $\sigma$ is zero. In this case, we call the interval $[-\bar{M}, \bar{M}]$ the support of $\sigma'$ and say the support of $\sigma'$ is bounded by $\bar{M}$. Another way of viewing this assumption is as saying that $\sigma$ is "capped" (or truncated) at $\sigma(\bar{M})$. It turns out that such a truncation ensures all $O(\text{OPT})$ solutions are unaffected. Similarly, the labels can be truncated in the interval $[-B, B]$, and, as a result, we can assume w.l.o.g. that $|y| \leq B$. A formal statement is provided below.

**Fact 8.2.4** (Lemma C.6, C.7 in [ZWDD25]). *Suppose the target activation $\sigma \in \mathcal{H}_\epsilon(B, L)$. Let $\bar{y} = \mathrm{sign}(y)\min\{|y|, B\}$. Then, $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\bar{y} - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] \le \mathrm{OPT} + \epsilon$ and for any $\widehat{\mathbf{w}}$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\bar{y} - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] = O(\mathrm{OPT}) + \epsilon$, we have $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] = O(\mathrm{OPT}) + \epsilon$.*

*Moreover, there exists $\tilde{\sigma} \in \mathcal{H}(B, L)$ such that $\|\tilde{\sigma} - \sigma\|_{L_2}^2 \le \epsilon$, with support of $\tilde{\sigma}'$ bounded by $\bar{M} \le \sqrt{2\log(4B^2/\epsilon) - \log\log(4B^2/\epsilon)}$. If $\widehat{\mathbf{w}}$ satisfies $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \tilde{\sigma}(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] \le O(\mathrm{OPT}) + \epsilon$, then also $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}))^2] \le O(\mathrm{OPT}) + \epsilon$. Thus, one can replace $\sigma$ with $\tilde{\sigma}$ and $y$ by $\bar{y}$, and assume w.l.o.g. that the support of $\sigma'$ is bounded by $\bar{M}$ and $|y| \le B$.*

**Error Bound** In the next fact, we show that for any function $\sigma \in \mathcal{H}_\epsilon(B, L)$, we can bound the $L_2^2$ loss $\mathcal{L}_2(\mathbf{w}; \sigma)$ at vector $\mathbf{w}$ by $\sin^2\theta \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$, where $\theta \stackrel{\mathrm{def}}{=} \theta(\mathbf{w}, \mathbf{w}^*)$.

**Fact 8.2.5** (Error Bound, Lemma D.8 + Proposition 4.5 [ZWDD25]). *Suppose $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2] = \mathrm{OPT}$ holds for a monotone activation $\sigma \in \mathcal{H}_\epsilon(B, L)$ and a unit vector $\mathbf{w}^* \in \mathbb{R}^d$ and suppose the support of $\sigma'$ is bounded by $M > 0$. Given any unit vector $\mathbf{w} \in \mathbb{R}^d$, let $\theta \stackrel{\mathrm{def}}{=} \theta(\mathbf{w}, \mathbf{w}^*)$. Then, if $\theta \le 1/M$, we have $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2] \le C\mathrm{OPT} + C\sin^2\theta \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$ for a universal constant $C > 1$.*

Learning with respect to the class of activations $\mathcal{H}_\epsilon(B, L)$ allows us to make the following simplifying assumption that comes at no loss of generality. We can assume that the labels $y$ are truncated in the interval $[-B, B]$, and, as a result, we can assume that $|y| \le B$. Our main result is that Algorithm 16 efficiently generates a solution pair that achieves $C\mathrm{OPT} + \epsilon$ error:

**Theorem 8.2.6** (Main Result). *Let $\epsilon > 0$ and let $B, L > 0$. Let $\mathcal{D}$ be a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Let $(\mathbf{w}^*, \sigma) \in \mathbb{R}^d \times \mathcal{H}_\epsilon(B, L)$ be a pair of*

**Algorithm 16** Main Algorithm

**Input:** Parameters $B, L, \epsilon, \delta$; data access $(\mathbf{x}, y) \sim \mathcal{D}$; initialize $S^{\mathrm{sol}} \leftarrow \emptyset$
$S^{\mathrm{ini}} \leftarrow$ Initialization$[B, \epsilon]$ (Alg. 17); $S^{\mathrm{sol}} \leftarrow S^{\mathrm{ini}}$
Draw $N$ i.i.d. samples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N} \sim \mathcal{D}$ and form $\widehat{\mathcal{D}}_N$
**for all** $\mathbf{w}^{(0)} \in S^{\mathrm{ini}}$ **do**
  **for all** $\bar{\theta} \in \Theta = \{k\,\epsilon/L : k = 0, \ldots, \lfloor L/\epsilon \rfloor\}$ **do**
    $S \leftarrow$ SpectralOptimization$[\bar{\theta}, \mathbf{w}^{(0)}, \widehat{\mathcal{D}}_N]$ (Alg. 18)
    $S^{\mathrm{sol}} \leftarrow S^{\mathrm{sol}} \cup S$
  **end for**
**end for**
$(\widehat{\mathbf{w}}, \widehat{u}) \leftarrow$ Test$[S^{\mathrm{sol}}]$ (Alg. 19)
**Return:** $(\widehat{\mathbf{w}}, \widehat{u})$

*vector and monotone activation such that* $\mathcal{L}_2(\mathbf{w}^*; \sigma) = \mathrm{OPT}$. *Then Algorithm 16 draws* $N = d^2\,\mathrm{poly}(B, L, 1/\epsilon)$ *samples, it runs in at most* $\mathrm{poly}(N, d)$ *time, and it returns a vector* $\widehat{\mathbf{w}}$ *and a monotone and Lipschitz activation* $\widehat{u} : \mathbb{R} \to \mathbb{R}$, *such that with probability at least* $2/3$, *it holds that* $\mathcal{L}_2(\widehat{\mathbf{w}}; \widehat{u}) \leq O(\mathrm{OPT}) + \epsilon$.

Using the fact that any monotone function $\sigma$ with bounded $2+\zeta$ moment $\mathbf{E}_{z\sim\mathcal{N}}[|\sigma(z)|^{2+\zeta}] \leq B_\sigma$ is an $\epsilon$-Extended $(B, L)$-Regular with $B, L = \mathrm{poly}((B_\sigma/\epsilon)^{1/\zeta}, 1/\epsilon)$, we have:

**Corollary 8.2.7.** *Let* $\epsilon, \zeta > 0$. *Let* $(\mathbf{x}, y) \sim \mathcal{D}$ *with* $\mathcal{D}_{\mathbf{x}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$. *Let* $\mathbf{w}^* \in \mathbb{R}^d$ *be a unit vector and let* $\sigma$ *be a monotone function with bounded* $(2 + \zeta)$ *moment, i.e.,* $\mathbf{E}_{z\sim\mathcal{N}}[|\sigma(z)|^{2+\zeta}] \leq B_\sigma$, *such that* $\mathcal{L}_2(\mathbf{w}^*; \sigma) = \mathrm{OPT}$. *Then, Algorithm 16 draws* $N = d^2\,\mathrm{poly}((B_\sigma/\epsilon)^{1/\zeta}, 1/\epsilon)$ *samples, runs in at most* $\mathrm{poly}(N, d)$ *time, and returns a vector* $\widehat{\mathbf{w}}$ *and a monotone and Lipschitz activation* $\widehat{u} : \mathbb{R} \to \mathbb{R}$, *such that with probability at least* $2/3$, *it holds that* $\mathcal{L}_2(\widehat{\mathbf{w}}; \widehat{u}) \leq O(\mathrm{OPT}) + \epsilon$.

Note further that monotone $L$-Lipschitz functions belong to $\mathcal{H}_\epsilon(B, L)$ with $B = O(L \log(L/\epsilon))$ [ZWDD25, Lemma C.12].

### 8.2.1 Initialization

Recall that the methodology of our initialization algorithm is to find a threshold $t$ such that after transforming the labels $y$ to $\mathcal{T}(y,t) = \mathbb{1}\{y \geq t\}$, we can reduce the regression problem to a robust halfspace learning problem and find a vector $\mathbf{w}^{(0)}$ satisfying $\theta(\mathbf{w}^{(0)}, \mathbf{w}) \leq 1/M$ via the robust halfspace learning algorithm from [DKTZ22b], where $M > 0$ is the smallest parameter such that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(M))^2 \mathbb{1}\{|z| \geq M\}] \leq C(\mathrm{OPT} + \epsilon)$. The detailed description of the initialization algorithm is the following: The following fact guarantees the existence of a target threshold that ensures the desired

**Algorithm 17** Initialization

1: **Input:** Parameters $B$, $\epsilon$; data access $(\mathbf{x}, y) \sim \mathcal{D}$
2: Initialize $S \leftarrow \emptyset$
3: **for** $i = 1$ **to** $\lceil B/\sqrt{\epsilon} \rceil + 1$ **do**
4: $\quad t_i \leftarrow i\sqrt{\epsilon}$
5: $\quad$ Transform data to $\mathcal{D}_i = \{(\mathbf{x}, \mathcal{T}(y; t_i))\}$ where $\mathcal{T}(y; t_i) = \mathbf{1}\{y \geq t_i\}$
6: $\quad$ Run the robust halfspace learner (Fact 8.2.9) on $\mathcal{D}_i$ to obtain $\mathbf{w}^{(0,i)}$
7: $\quad S \leftarrow S \cup \{\mathbf{w}^{(0,i)}\}$
8: **end for**
9: **Return:** $S$

initialization.

**Fact 8.2.8** (Proposition F.19 and Lemma F.21 in [ZWDD25])**.** *Let $\sigma(\mathbf{w}^* \cdot \mathbf{x})$ be an optimal hypothesis that satisfies $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq \epsilon_0$ for $\epsilon_0 \stackrel{\mathrm{def}}{=} \mathrm{OPT} + \epsilon > 0$, where $\sigma$ is a non-decreasing $\epsilon$-Extended $(B, L)$-regular function. Suppose the constant hypothesis $\sigma(z) = c$ is not a constant factor approximate solution for any $c \in \mathbb{R}$. Let $C_1, C_2 > 0$ be large absolute constants. Then, there exists a minimum $M > 0$ that satisfies $\|(\sigma(z) - \sigma(M))\mathbb{1}\{z \geq M\}\|_{L_2}^2 \leq C_1\epsilon_0$, such that $\mathbf{Pr}[\mathbb{1}\{y \geq \sigma(M)\} \neq \mathbb{1}\{\mathbf{w}^* \cdot \mathbf{x} \geq M\}] \leq \mathbf{Pr}[\mathbf{w}^* \cdot \mathbf{x} \geq M]/C_2$.*

Suppose we are given the threshold $t = \sigma(M)$ with $M$ satisfying Fact 8.2.8. After transforming the labels to $\tilde{y} = \mathcal{T}(y; t) = \mathbb{1}\{y \geq t\}$, we can apply the algorithm from [DKTZ22b, ZWDD25] to find a vector $\mathbf{w}^0$ such that $\theta(\mathbf{w}^0, \mathbf{w}^*) \leq 1/M$.

**Fact 8.2.9** (Proposition F.19, Fact F.20 in [ZWDD25])**.** *Let $h^*(\mathbf{x}) = \mathbb{1}\{\mathbf{w}^* \cdot \mathbf{x} \geq M\}$ be a target Gaussian halfspace, i.e., $\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}$ and $\mathcal{D}_{\mathbf{x}}$ is a standard Gaussian distribution. Let $(\mathbf{x}, \tilde{y}) \sim \mathcal{D}$ be a distribution of labeled samples with $\mathrm{OPT}'$-adversarial noise—meaning that $\mathbf{Pr}[\tilde{y} \neq h^*(\mathbf{x})] \leq \mathrm{OPT}'$, with the adversarial noise satisfying $\mathrm{OPT}' \leq (1/C_2)(\exp(-M^2/2)/M) \approx (1/C_2)\,\mathbf{Pr}[h^*(\mathbf{x}) = 1]$, where $C_2$ is a large absolute constant. Then, there is an algorithm that uses $O(d/\epsilon^2 \log(1/\delta))$ samples, returns with probability at least $1 - \delta$ a vector $\mathbf{w}$ such that $\theta(\mathbf{w}, \mathbf{w}^*) \leq \min(\pi/16, 1/M)$.*

Therefore, what remains is to find such a threshold $t$. Since the target activation $\sigma$ is unknown, it is not possible to find $M$ and calculate $\mathbf{Pr}[z \geq M]$ and estimate $\mathbf{Pr}[\mathbb{1}\{y \geq \sigma(M)\} \neq \mathbb{1}\{z \geq M\}]$. Our strategy is to show that we can discretize the labels $y$ and the target activation $\sigma$ so that they only take a finite number of values, which then implies that the target threshold $t = \sigma(M)$ must be an element of a finite set. Then, we can brute force iterates through all the possible values (we show there are polynomially many), and run the robust halfspace learning algorithm from Fact 8.2.9 on each of the label transformations $\mathcal{T}(y; t_i)$, $t_i = i\sqrt{\epsilon}$. Formally, we have:

**Claim 8.2.10.** *Suppose $\|y - \sigma(\mathbf{w}^* \cdot \mathbf{x})\|_{L_2}^2 \leq \epsilon_0$ for some $\epsilon_0 > 0$. Define the truncation operator $\mathrm{tr}(\cdot)$ by $\mathrm{tr}(z) = -B + i\sqrt{\epsilon}$ if $z \in [i\sqrt{\epsilon}, (i+1)\sqrt{\epsilon})$, $i \in [2B/\sqrt{\epsilon}]$. Then $\|\mathrm{tr}(y) - \mathrm{tr}(\sigma(\mathbf{w}^* \cdot \mathbf{x}))\|_{L_2}^2 \leq 9(\epsilon_0 + \epsilon)$.*

*Proof.* Since $\sigma$ is $(B, L)$-regular (thus $\|\sigma\|_{L_\infty} \leq B$) and since w.l.o.g. $|y| \leq B$ (Fact 8.2.4), we have $\|\mathrm{tr}(y) - y\|_{L_2} \leq \sqrt{\epsilon_0}$ and $\|\sigma(\mathbf{w}^* \cdot \mathbf{x}) - \mathrm{tr}(\sigma(\mathbf{w}^* \cdot \mathbf{x}))\|_{L_2} \leq \sqrt{\epsilon_0}$. Direct calculation

yields:

$$\|\mathrm{tr}(y) - \mathrm{tr}(\sigma(\mathbf{w}^* \cdot \mathbf{x}))\|_{L_2} = \|\mathrm{tr}(y) - y + y - \sigma(\mathbf{w}^* \cdot \mathbf{x}) + \sigma(\mathbf{w}^* \cdot \mathbf{x}) - \mathrm{tr}(\sigma(\mathbf{w}^* \cdot \mathbf{x}))\|_{L_2}$$
$$\leq 2\sqrt{\epsilon} + \sqrt{\epsilon}_0.$$

Thus, we have $\|\mathrm{tr}(y) - \mathrm{tr}(\sigma(\mathbf{w}^* \cdot \mathbf{x}))\|_{L_2}^2 \leq 9(\epsilon + \epsilon_0)$. □

**Lemma 8.2.11** (Initialization)**.** *Let $\sigma(\mathbf{w}^* \cdot \mathbf{x})$ be an optimal hypothesis that satisfies $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq \mathrm{OPT} + \epsilon$, where $\sigma$ is a non-decreasing $\epsilon$-Extended $(B, L)$-Regular function. Suppose that no constant hypothesis, i.e., functions of the form $\sigma(z) = c$ for any $c \in \mathbb{R}$, is a constant factor approximate solution. Let $C > 1$ be a large absolute constant and let $M > 0$ be the smallest parameter such that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(M))^2 \mathbb{1}\{z \geq M\}] \leq C(\mathrm{OPT} + \epsilon)$. Then Algorithm 17, using $O(d/\epsilon^2 \log(B/\epsilon))$ samples, with probability at least 99%, returns a list $S^{\mathrm{ini}}$ of $O(B/\sqrt{\epsilon})$ vectors that contains a vector $\mathbf{w}^{(0)}$ such that $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \leq \min(1/M, \pi/16)$.*

*Proof.* By Claim 8.2.10, we can discretize the labels and the target activation $\sigma$ so that it only takes a finite number of values: $i\sqrt{\epsilon}$, $i \in [\lceil 2B/\sqrt{\epsilon} \rceil + 1]$, while only inducing a small error. By Fact 8.2.4, there exits a threshold $t \in \{i\sqrt{\epsilon}\}_{i=0}^{B/\sqrt{\epsilon}}$ so that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - t)^2 \mathbb{1}\{\sigma(z) \geq t\}] \leq C(\mathrm{OPT} + \epsilon)$ and $\mathbf{Pr}[\mathbb{1}\{y \geq t\} \neq \mathbb{1}\{\sigma(\mathbf{w}^* \cdot \mathbf{x}) \geq t\}] \leq \mathbf{Pr}[\sigma(\mathbf{w}^* \cdot \mathbf{x}) \geq t]/C_1$, where $C, C_1$ are large absolute constants. Then by Fact 8.2.9 we know that using the algorithm from [DKTZ22b] we will obtain a vector $\mathbf{w}$ such that $\theta(\mathbf{w}, \mathbf{w}^*) \leq 1/M$ where $M = \sigma^{-1}(t)$. The algorithm requires $O(d/\epsilon^2 \log(1/\delta'))$ samples to return such a vector with probability $1 - \delta'$. Since we run the algorithm for $B/\sqrt{\epsilon}$ times, let $\delta' = \sqrt{\epsilon}\delta/B$, and after a union bound we get that with $O(d/\epsilon^2 \log(B/(\epsilon\delta)))$ samples, the algorithm succeeds with probability $1 - \delta$ for all the iterations. Set $\delta = 0.01$

finishes the proof. □

### 8.2.2 The Spectral Subroutine

In this section, we present and prove our main structural result (Proposition 8.2.12). We show that—even though the target activation $\sigma$ is unknown—we can identify a vector that has a strong correlation with an 'ideal descent direction' $\mathbf{v}_{\mathbf{w}}^* := (\mathbf{w}^*)^{\perp \mathbf{w}}/\|(\mathbf{w}^*)^{\perp \mathbf{w}}\|_2$. It is not hard to see that $\mathbf{v}_{\mathbf{w}}^*$ can be used to rotate $\mathbf{w}$ towards $\mathbf{w}^*$. The vector that we identify is a top eigenvector of a matrix $\mathbf{M}_{\mathbf{w}}$ that can be efficiently estimated using sample access to labeled data. We can only identify such a target vector up to its sign; however, as we argue later, this is sufficient for our argument to go through.

To build up this result, we need the following technical pieces: (1) the spectrum of the matrix $\mathbf{M}_{\mathbf{w}}$ contains information on $\mathbf{v}_{\mathbf{w}}^*$, i.e., $\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{M}_{\mathbf{w}} \mathbf{v}_{\mathbf{w}}^*$ is large (Lemma 8.2.14); (2) All the other directions $\mathbf{u}$ that are orthogonal to $\mathbf{v}_{\mathbf{w}}^*$ have small quadratic form value compared to $\mathbf{v}_{\mathbf{w}}^*$, i.e., $\mathbf{u} \cdot \mathbf{M}_{\mathbf{w}} \mathbf{v} \ll \mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{M}_{\mathbf{w}} \mathbf{v}_{\mathbf{w}}^*$, and therefore, the direction $\mathbf{v}_{\mathbf{w}}^*$ stands out in the spectrum of the matrix $\mathbf{M}_{\mathbf{w}}$ (Lemma 8.2.16); (3) finally, we argue that the top eigenvector $\mathbf{v}_{\mathbf{w}}$ of $\mathbf{M}_{\mathbf{w}}$ correlates strongly with $\mathbf{v}_{\mathbf{w}}^*$ (Lemma 8.2.17).

**Proposition 8.2.12** (Main Structural Result)**.** *Fix constants $B, L > 0$. Let $\mathcal{D}$ be a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ with $\mathcal{D}_{\mathbf{x}} = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and let $\widehat{\mathcal{D}}_N$ be its empirical distribution on $N$ i.i.d. samples. Suppose there is a vector $\mathbf{w}^* \in \mathbb{R}^d$ and an activation $\sigma \in \mathcal{H}_\epsilon(B, L)$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2] = \mathrm{OPT}$. There exists an absolute constant $C > 0$ such that if $N = \Theta(d^2 B^{12} L^8/\epsilon^{10} \log(dBL/\epsilon))$, then with probability at least 99% the following holds: for any unit vector $\mathbf{w} \in \mathbb{R}^d$ satisfying $\sin\theta(\mathbf{w}, \mathbf{w}^*) \geq 40\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos\theta(\mathbf{w},\mathbf{w}^*)}\sigma'\|_{L_2}$, let $\mathbf{u} \in \mathbb{R}^d$ be the top eigenvector of the empirical matrix $\widehat{\mathbf{M}}_{\mathbf{w}}$ from Algorithm 18. Then $|\mathbf{u} \cdot \mathbf{w}^*| \geq (\sqrt{2}/2)\sin\theta(\mathbf{w}, \mathbf{w}^*)$ and $\mathbf{u} \cdot \mathbf{w} = 0$.*

**Algorithm 18** Spectral Optimization

**Input:** Parameter $\theta_0$; initialization $\mathbf{w}^{(0)}$; empirical distribution $\widehat{\mathcal{D}}_N$
Initialize $S^{\mathrm{sol}} \leftarrow \emptyset$
Set $K \leftarrow \mathrm{poly}(1/\epsilon, L), \quad T \leftarrow \Theta\Big(\log(1/(\epsilon L))\Big)$
For $t = 0, 1, \ldots$: $\phi_t \leftarrow \bar{\theta}\,(1 - \frac{1}{128})^t, \quad \eta_t \leftarrow \frac{\sin \phi_t}{8}$
**for** $k = 1$ **to** $K$ **do**
    **for** $t = 0$ **to** $T$ **do**
        For each cell $\mathcal{E}_j$, compute

$$\hat{\mathbf{g}}^{(j)}_{\mathbf{w}^{(t)}} \leftarrow \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}_N}\Big[y\,\mathbf{x}^{\perp \mathbf{w}^{(t)}} \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x} \in \mathcal{E}_j\}\Big].$$

        Form empirical matrix

$$\widehat{\mathbf{M}}_{\mathbf{w}^{(t)}} \leftarrow \sum_{j=1}^{I} \frac{\hat{\mathbf{g}}^{(j)}_{\mathbf{w}^{(t)}}\, \hat{\mathbf{g}}^{\mathsf{T},(j)}_{\mathbf{w}^{(t)}}}{\mathbf{Pr}_{z\sim\mathcal{N}(0,1)}[\, z \in \mathcal{E}_j\,]}.$$

        Let $\mathbf{u}$ be the top eigenvector of $\widehat{\mathbf{M}}_{\mathbf{w}^{(t)}}$, and choose $\mathbf{v}^{(t)} \leftarrow \pm\mathbf{u}$ uniformly at random.
        Update $\mathbf{w}^{(t+1)} \leftarrow \mathrm{proj}_{\mathbb{B}^d}\Big(\mathbf{w}^{(t)} - \eta_t\, \mathbf{v}^{(t)}\Big)$
        $S^{\mathrm{sol}} \leftarrow S^{\mathrm{sol}} \cup \{\mathbf{w}^{(t+1)}\}$
    **end for**
**end for**
**Return:** $S^{\mathrm{sol}}$

The rest of this subsection develops machinery required to prove the proposition above. We will prove Proposition 8.2.12 in Section (8.2.4).

[ZWDD25] showed that given the target activation $\sigma$, the gradient vector of the smoothed $L_2^2$ loss correlates strongly with $\mathbf{w}^*$:

**Fact 8.2.13** (Proposition 2.2, [ZWDD25])**.** *Fix an activation $\sigma : \mathbb{R} \to \mathbb{R}$. Fix vectors $\mathbf{w}^*, \mathbf{w} \in \mathbb{S}^{d-1}$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2] = \mathrm{OPT}$ and let $\theta = \theta(\mathbf{w}^*, \mathbf{w})$. If $\sin\theta \geq 3\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$, then:*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y \mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbf{x}^{\perp \mathbf{w}}] \cdot \mathbf{w}^* \geq (2/3)\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 \sin^2\theta.$$

However, the structural result above relies heavily on the knowledge of the target activation $\sigma$ and is thus not applicable to our setting. Instead, we argue that using the vector $\mathbf{g}_{\mathbf{w}}(z)$ defined below, the top eigenvector of the matrix $\mathbf{M}_{\mathbf{w}} \stackrel{\text{def}}{=} \mathbf{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\mathbf{g}_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{z})\mathbf{g}_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{z})^\intercal]$ correlates with the "ideal" update direction $\mathbf{v}^*_{\mathbf{w}} \stackrel{\text{def}}{=} (\mathbf{w}^*)^{\perp\mathbf{w}}/\|(\mathbf{w}^*)^{\perp\mathbf{w}}\|_2$:

$$\mathbf{g}_{\mathbf{w}}(z) \stackrel{\text{def}}{=} \sum_{i=1}^{I} \frac{\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i\}]}{\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i]}\mathbb{1}\{z\in\mathcal{E}_i\},$$

$$\text{where}\ \begin{cases} \mathcal{E}_i = [a_i, a_{i+1}) = [-M' + (i-1)\Delta, -M' + i\Delta),\ \Delta = \epsilon^2/(B^2L^2); \\ M' = O(\sqrt{\log(BL/\epsilon)})\ \text{satisfies}\ \mathbf{Pr}_{z\sim\mathcal{N}}[|z|\geq M'] \leq \epsilon^2/(B^2L^2); \\ I = O(M'B^2L^2/\epsilon^2) = \tilde{O}(B^2L^2/\epsilon^2). \end{cases} \tag{Grad}$$

Denote $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}} \mid \mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i] \stackrel{\text{def}}{=} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i\}]/\mathbf{Pr}[\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i]$, so that $\mathbf{g}_{\mathbf{w}}(z)$ can be written as $\mathbf{g}_{\mathbf{w}}(z) = \sum_i \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}} \mid \mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i]\mathbb{1}\{z\in\mathcal{E}_i\}$. We remind that on each interval $\mathcal{E}_i$, $\mathbf{g}_{\mathbf{w}}(z) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}} \mid \mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i]$ is a fixed constant vector that depends on $\mathbf{w}$.

**Lemma 8.2.14.** *Let $\mathbf{g}_{\mathbf{w}}(z)$ be the vector defined in (Grad), let $\mathbf{M}_{\mathbf{w}} \stackrel{\text{def}}{=} \mathbf{E}_{z\sim\mathcal{N}}[\mathbf{g}_{\mathbf{w}}(z)\mathbf{g}_{\mathbf{w}}(z)^\intercal]$, and $\mathbf{v}^*_{\mathbf{w}} \stackrel{\text{def}}{=} (\mathbf{w}^*)^{\perp\mathbf{w}}/\|(\mathbf{w}^*)^{\perp\mathbf{w}}\|_2$. Suppose $\epsilon \leq \mathrm{OPT}$ and $\sin\theta \geq 4\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos\theta}\sigma'(z)\|_{L_2}$. Then, $(\mathbf{v}^*_{\mathbf{w}})^\intercal\mathbf{M}_{\mathbf{w}}\mathbf{v}^*_{\mathbf{w}} \geq (1/16)\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|^2_{L_2}$.*

*Proof.* Observe that $\mathbf{v}^*_{\mathbf{w}}\sin\theta = (\mathbf{w}^*)^{\perp\mathbf{w}}$. Consider the following quantity:

$$K(h(\mathbf{w}\cdot\mathbf{x})) := \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{w}^*\cdot\mathbf{x}^{\perp\mathbf{w}}h(\mathbf{w}\cdot\mathbf{x})].$$

Define the function class $\mathcal{H}'$ by

$$\mathcal{H}' = \left\{h:\mathbb{R}\to\mathbb{R} \mid h(z) = \sum_{i=1}^{I} h_i\mathbb{1}\{z\in\mathcal{E}_i\}, \|h\|_{L_2} = 1\right\},$$

where $\mathcal{E}_i, i = 1, \ldots, I$, are the same intervals as in the definition of $\mathbf{g}_{\mathbf{w}}(z)$, (Grad). Our goal is to find $h \in \mathcal{H}'$ that maximizes $K(h(\mathbf{w} \cdot \mathbf{x}))$. Observe that for $h \in \mathcal{H}'$, $K(h(\mathbf{w} \cdot \mathbf{x}))$ can be written as

$$
\begin{aligned}
K(h(\mathbf{w} \cdot \mathbf{x})) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y\mathbf{w}^* \cdot \mathbf{x}^{\perp \mathbf{w}} \sum_{i=1}^{I} h_i \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i\}\right] \\
&\overset{(i)}{=} \sum_{i=1}^{I} h_i \sin\theta \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{v}^*_{\mathbf{w}} \cdot \mathbf{x})\mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i\}] \\
&\overset{(ii)}{=} \sum_{i=1}^{I} h_i \sin\theta \left(\mathbf{v}^*_{\mathbf{w}} \cdot \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp \mathbf{w}} \mid \mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i]\right) \mathbf{Pr}[\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i] \\
&= \sum_{i=1}^{I} h_i \sin\theta \left(\mathbf{v}^*_{\mathbf{w}} \cdot \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp \mathbf{w}} \mid \mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i]\right) \mathbf{Pr}[z \in \mathcal{E}_i] \qquad (8.1) \\
&\overset{(iii)}{=} \sin\theta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\sum_{i=1}^{I} \left(h_i \mathbb{1}\{z \in \mathcal{E}_i\}\right)\left(\mathbf{v}^*_{\mathbf{w}} \cdot \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp \mathbf{w}} \mid \mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i]\mathbb{1}\{z \in \mathcal{E}_i\}\right)\right] \\
&= \sin\theta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[h(z)(\mathbf{v}^*_{\mathbf{w}} \cdot \mathbf{g}_{\mathbf{w}}(z))\right] = \sin\theta \left\langle h(z), \mathbf{v}^*_{\mathbf{w}} \cdot \mathbf{g}_{\mathbf{w}}(z)\right\rangle_{L_2(\mathcal{N}(0,1))},
\end{aligned}
$$

where $(i)$ is by $\mathbf{w}^* \cdot \mathbf{x}^{\perp \mathbf{w}} = (\mathbf{w}^*)^{\perp \mathbf{w}} \cdot \mathbf{x}^{\perp \mathbf{w}} = \mathbf{v}^*_{\mathbf{w}} \cdot \mathbf{x}^{\perp \mathbf{w}} \sin\theta$, $(ii)$ is by the definition of $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp \mathbf{w}} \mid \mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i]$, $(iii)$ is by $\mathbf{Pr}[z \in \mathcal{E}_i] = \mathbf{E}[\mathbb{1}\{z \in \mathcal{E}_i\}] = \mathbf{E}_{z\sim\mathcal{N}}[(\mathbb{1}\{z \in \mathcal{E}_i\})^2]$ and an appropriate grouping of terms, and the remaining inequalities use the definition of the inner product and that $\mathbf{w} \cdot \mathbf{z} \sim \mathcal{N}(0, 1)$. Because $K(h(\mathbf{w} \cdot \mathbf{x}))$ is an inner product in the space of $L_2(\mathcal{N})$ functions, from the $L_2$ norm (self-)duality, this is maximized for $h(z) = (\mathbf{g}_{\mathbf{w}}(z) \cdot \mathbf{v}^*_{\mathbf{w}})/\|\mathbf{g}_{\mathbf{w}}(z) \cdot \mathbf{v}^*_{\mathbf{w}}\|_{L_2}$.

Now we study $(\mathbf{v}^*_{\mathbf{w}})^\intercal \mathbf{M}_{\mathbf{w}} \mathbf{v}^*_{\mathbf{w}}$. Using the result that $K(h(\mathbf{w} \cdot \mathbf{x}))$ is maximized at $h^*(z) = \mathbf{v}^*_{\mathbf{w}} \cdot \mathbf{g}_{\mathbf{w}}(z)/\|\mathbf{v}^*_{\mathbf{w}} \cdot \mathbf{g}_{\mathbf{w}}(z)\|_{L_2}$ from the discussion above, and noting again that since $\mathbf{x}$ follows the standard Gaussian distribution, it holds $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_j] = \mathbf{Pr}_{z\sim\mathcal{N}}[z \in \mathcal{E}_j]$,

we have that

$$
\begin{aligned}
(\mathbf{v}_{\mathbf{w}}^*)^\intercal \mathbf{M}_{\mathbf{w}} \mathbf{v}_{\mathbf{w}}^* &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[(\mathbf{g}_{\mathbf{w}}(z)\cdot\mathbf{v}_{\mathbf{w}}^*)^2\right] && (8.2)\\
&= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\sum_{i=1}^{I}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{x}) \mid \mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i]\mathbb{1}\{z\in\mathcal{E}_i\}\right)^2\right]\\
&= \sum_{i=1}^{I}\left(\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{x}) \mid \mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i]\right)^2 \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\mathbb{1}\{z\in\mathcal{E}_i\}]\\
&= \sum_{i=1}^{I}\left(\mathbf{v}_{\mathbf{w}}^*\cdot\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}} \mid \mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i]\right)^2 \mathbf{Pr}[z\in\mathcal{E}_i]\\
&= \frac{\|\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{g}_{\mathbf{w}}(z)\|_{L_2}}{\sin\theta}K(h^*(\mathbf{w}\cdot\mathbf{x}))\,, && (8.3)
\end{aligned}
$$

where in the last equality we used Equation (8.1) and we brought in $h_i^* = \mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}} \mid \mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i]/\|\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{g}_{\mathbf{w}}(z)\|_{L_2}$ by the definition of $h^*(z)$. Let us define

$$
\widetilde{\mathrm{T}}_{\cos\theta\sigma'}(z) = \sum_{i=1}^{I}\mathrm{T}_{\cos\theta\sigma'}(a_i)\mathbb{1}\{z\in\mathcal{E}_i\}. \tag{8.4}
$$

Recall that we have defined $a_i$ as the left end-point of the interval $\mathcal{E}_i = [a_i, a_{i+1})$ in (Grad). We show that $\widetilde{\mathrm{T}}_{\cos\theta\sigma'}(\mathbf{w}\cdot\mathbf{x})$ is close to $\mathrm{T}_{\cos\theta\sigma'}(\mathbf{w}\cdot\mathbf{x})$ point-wise.

**Claim 8.2.15.** *Suppose* $\sin\theta\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2} \gtrsim \sqrt{\mathrm{OPT}}$ *and* $\epsilon \le \mathrm{OPT}$. *Consider the piece-wise constant function* $\widetilde{\mathrm{T}}_{\cos\theta\sigma'}(z)$ *defined in Equation (8.4), with the intervals* $\mathcal{E}_i$, $i\in[I]$, *and parameter* $M'$ *following the construction in* (Grad). *Then,*

$$
\begin{aligned}
|\widetilde{\mathrm{T}}_{\cos\theta\sigma'}(z) - \mathrm{T}_{\cos\theta\sigma'}(z)|\mathbb{1}\{z\in[-M',M']\} &\le \epsilon/B\,;\\
|\|\mathrm{T}_{\cos\theta\sigma'}(z)\|_{L_2} - \|\widetilde{\mathrm{T}}_{\cos\theta\sigma'}(z)\|_{L_2}| &\le \epsilon\,.
\end{aligned}
$$

*Proof.* By Fact 7.1.7 part (c), we know that for any $\rho\in(0,1)$, $\|(\mathrm{T}_\rho f(z))'\|_{L_\infty} \le$

$\|f\|_{L_\infty}/\sqrt{1-\rho^2}$. In addition, by Fact 7.1.7 part (a) we have $\mathrm{T}_{\cos\theta}\sigma'(z) = \mathrm{T}_{\sqrt{\cos\theta}}(\mathrm{T}_{\sqrt{\cos\theta}}\sigma'(z))$. Therefore we claim that $\mathrm{T}_{\cos\theta}\sigma'(z)$ is a Lipschitz function:

$$\|(\mathrm{T}_{\cos\theta}\sigma'(z))'\|_{L_\infty} = \|(\mathrm{T}_{\sqrt{\cos\theta}}(\mathrm{T}_{\sqrt{\cos\theta}}\sigma'(z)))'\|_{L_\infty} \le \frac{\|\mathrm{T}_{\sqrt{\cos\theta}}\sigma'(z)\|_{L_\infty}}{\sqrt{1-\cos\theta}} \overset{(i)}{=} \frac{\|(\mathrm{T}_{\sqrt{\cos\theta}}\sigma(z))'\|_{L_\infty}}{\sqrt{2\cos\theta}\sin(\theta/2)}$$
$$\le \frac{\|\sigma(z)\|_{L_\infty}}{\sqrt{2\cos\theta}\sin(\theta/2)\sqrt{1-\cos\theta}} \overset{(ii)}{\le} \frac{B}{2\sin^2(\theta/2)\sqrt{\cos\theta}},$$

where in $(i)$ we applied Fact 7.1.7 part (g) and the fact that $1-\cos\theta = 2\sin^2(\theta/2)$ and in $(ii)$ we used the fact that $\|\sigma\|_{L_\infty} \le B$ since $\sigma$ is $\epsilon$-Extended $(B,L)$-regular. Furthermore, note that by our assumption that $\sin\theta \gtrsim \sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos\theta}\sigma'\|_2$, we have $\sin^2\theta \ge \mathrm{OPT}/\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$. Using the fact that $\|\mathrm{T}_\rho f\|_{L_2}^2$ is an non-decreasing function with respect to $\rho$ (Fact 7.1.7, (f)), we have $\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 \le \|\sigma'\|_{L_2}^2 \le L^2$, again by the assumption that $\sigma$ is $\epsilon$-Extended $(B,L)$-Regular. Hence $\sin^2\theta \ge \mathrm{OPT}/L^2$. Finally, our initialization subroutine guarantees that $\cos\theta \ge c$ for some small absolute constant $c>0$. Therefore, in summary, we obtain that $\|(\mathrm{T}_{\cos\theta}\sigma'(z))'\|_{L_\infty} \lesssim BL^2/\mathrm{OPT}$.

Now for any $i \in [I]$, let $z \in [a_i, a_{i+1})$ be any value in the interval $\mathcal{E}_i$. Since we have proved in the last paragraph that $\mathrm{T}_{\cos\theta}\sigma'(z)$ is $O(BL^2/\mathrm{OPT})$-Lipschitz, we have that there exists a sufficiently large absolute constant $C'$ such that

$$|\mathrm{T}_{\cos\theta}\sigma'(z) - \mathrm{T}_{\cos\theta}\sigma'(a_i)| \le C'BL^2/\mathrm{OPT}|z-a_i| \le (C'BL^2/\mathrm{OPT})(\epsilon^2/(CB^2L^2)) \le \epsilon/B.$$

Note that in the last inequality we used (by the definition of $\mathcal{E}_i = [a_i, a_{i+1})$ in (Grad)) that $a_{i+1} - a_i = \Delta \le \epsilon^2/(CB^2L^2)$, where $C \ge C'$ is a sufficiently large absolute constant. Therefore, we conclude that $|\widetilde{\mathrm{T}}_{\cos\theta}\sigma'(z) - \mathrm{T}_{\cos\theta}\sigma'(z)|\mathbb{1}\{z \in [-M', M']\} \le \epsilon/B$, proving the first part of the claim.

For the second part of the claim, note first that

$$\|\mathrm{T}_{\cos\theta\sigma'}(z) - \widetilde{\mathrm{T}}_{\cos\theta\sigma'}(z)\|_{L_2} \le \|(\mathrm{T}_{\cos\theta\sigma'}(z) - \widetilde{\mathrm{T}}_{\cos\theta\sigma'}(z))\mathbb{1}\{z \in [-M', M']\}\|_{L_2} + \|\mathrm{T}_{\cos\theta\sigma'}(z)\mathbb{1}\{|z| \ge M'\}\|_{L_2}$$

Using the first part of the claim, we obtain $\|(\mathrm{T}_{\cos\theta\sigma'}(z) - \widetilde{\mathrm{T}}_{\cos\theta\sigma'}(z))\mathbb{1}\{z \in [-M', M']\}\|_{L_2} \le \epsilon$. Now applying Fact 7.1.7(c) again we have $\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_\infty} \le B/(\cos\theta \sin\theta)$. We have argued in the previous paragraph that $\sin\theta \ge \sqrt{\mathrm{OPT}}/L \ge \sqrt{\epsilon}/L$ and $\cos\theta \ge c$ for some small absolute constant $c$ under our assumptions. Therefore, $\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_\infty} \lesssim BL/\sqrt{\epsilon}$. Since $M' = O(\sqrt{\log(BL/\epsilon)})$ is chosen such that $\mathbf{Pr}[|z| \ge M'] \le \epsilon^2/(CBL)^2$ for a large absolute constant $C$ and a standard Gaussian random variable $z$, we have

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathrm{T}_{\cos\theta\sigma'}(z))^2\mathbb{1}\{|z| \ge M\}] \le (B^2L^2/\epsilon)\,\mathbf{Pr}[|z| \ge M'] \le \epsilon.$$

Therefore, it holds $|\|\mathrm{T}_{\cos\theta\sigma'}(z)\|_{L_2} - \|\widetilde{\mathrm{T}}_{\cos\theta\sigma'}(z)\|_{L_2}| \le 2\epsilon$. □

Now observe that since $\widetilde{\mathrm{T}}_{\cos\theta\sigma'}(z)/\|\widetilde{\mathrm{T}}_{\cos\theta\sigma'}\|_{L_2} \in \mathcal{H}'$, we have $K(h^*(\mathbf{w}\cdot\mathbf{x})) \ge K(\widetilde{\mathrm{T}}_{\cos\theta\sigma'}(\mathbf{w}\cdot\mathbf{x})/\|\widetilde{\mathrm{T}}_{\cos\theta\sigma'}\|_{L_2})$, which implies (from Equation (8.2))

$$\begin{aligned}\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}^*_{\mathbf{w}} &\ge \frac{\|\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{g}_{\mathbf{w}}(z)\|_{L_2}}{\sin\theta}K(\widetilde{\mathrm{T}}_{\cos\theta\sigma'}(\mathbf{w}\cdot\mathbf{x})/\|\widetilde{\mathrm{T}}_{\cos\theta\sigma'}\|_{L_2}) \\ &= \frac{\|\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{g}_{\mathbf{w}}(z)\|_{L_2}}{\|\widetilde{\mathrm{T}}_{\cos\theta\sigma'}\|_{L_2}\sin\theta}\sum_{i=1}^{I}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{w}^*\cdot\mathbf{x}^{\perp\mathbf{w}})\mathrm{T}_{\cos\theta\sigma'}(a_i)\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i\}] \quad (8.5)\end{aligned}$$

Note that by the definition of $\mathbf{M}_{\mathbf{w}}$, we have $(\mathbf{v}^*_{\mathbf{w}})^\intercal\mathbf{M}_{\mathbf{w}}\mathbf{v}^*_{\mathbf{w}} = \mathbf{E}_{z\sim\mathcal{N}}[(\mathbf{v}^*_{\mathbf{w}}\cdot\mathbf{g}_{\mathbf{w}}(z))^2] = \|\mathbf{g}_{\mathbf{w}}(z)\cdot\mathbf{v}^*_{\mathbf{w}}\|^2_{L_2}$. Furthermore, as we have shown in Claim 8.2.15, it holds $\|\widetilde{\mathrm{T}}_{\cos\theta\sigma'}\|_{L_2} \le \|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2} + \epsilon$ and note that since we have assumed $1 \ge \sin\theta \ge 3\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2}$, we have $3\sqrt{\epsilon} \le 3\sqrt{\mathrm{OPT}} \le \|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2}$. Thus, for small $\epsilon$ it holds $\sqrt{\epsilon} \le \|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2}$ and

we obtain $\|\widetilde{\mathrm{T}}_{\cos\theta}\sigma'\|_{L_2} \le 2\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$. Plugging $\|\mathbf{g}(z)\cdot\mathbf{v}_{\mathbf{w}}^*\|_{L_2} = (\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^*)^{1/2}$ and $\|\widetilde{\mathrm{T}}_{\cos\theta}\sigma'\|_{L_2} \le 2\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$ back into Equation (8.5) yields

$$
\begin{aligned}
&(\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{M}\mathbf{v}_{\mathbf{w}}^*)^{1/2}\\
&\ge \frac{1}{2\sin\theta\|\mathrm{T}_{\cos\theta}\sigma'(z)\|_{L_2}}\sum_{i=1}^{I}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{w}^*\cdot\mathbf{x}^{\perp\mathbf{w}})\mathrm{T}_{\cos\theta}\sigma'(a_i)\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i\}]\\
&= \frac{1}{2\sin\theta\|\mathrm{T}_{\cos\theta}\sigma'(z)\|_{L_2}}\underbrace{\sum_{i=1}^{I}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{w}^*\cdot\mathbf{x}^{\perp\mathbf{w}})(\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x}))\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i\}]}_{(Q_1)}\\
&\quad+\frac{1}{2\|\mathrm{T}_{\cos\theta}\sigma'(z)\|_{L_2}}\underbrace{\sum_{i=1}^{I}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{x})(\mathrm{T}_{\cos\theta}\sigma'(a_i)-\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x}))\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i\}]}_{(Q_2)}.
\end{aligned}
\tag{8.6}
$$

For the term $(Q_1)$, we apply Fact 8.2.13 and obtain

$$
\begin{aligned}
(Q_1) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{w}^*\cdot\mathbf{x}^{\perp\mathbf{w}})\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}|\le M'\}]\\
&\ge (2/3)\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{w}^*\cdot\mathbf{x}^{\perp\mathbf{w}})\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}|\ge M'\}]\\
&\ge (2/3)\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 - B\,\mathbf{E}[|\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{x}|]\,\mathbf{E}[|\sin\theta\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})|\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}|\ge M'\}],
\end{aligned}
$$

where in the second inequality we used $|y|\le B$ and $\mathbf{v}_{\mathbf{w}}^*\sin\theta = (\mathbf{w}^*)^{\perp\mathbf{w}}$. Using Cauchy-Schwarz inequality further yields

$$
\begin{aligned}
\mathbf{E}[|\sin\theta\mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})|\mathbb{1}\{|\mathbf{w}\cdot\mathbf{x}|\ge M'\}] &\le \sin\theta\sqrt{\mathbf{E}[(\mathrm{T}_{\cos\theta}\sigma'(z))^2]\,\mathbf{Pr}[|\mathbf{w}\cdot\mathbf{x}|\ge M']}\\
&\le \sin\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}(\epsilon/(CBL)),
\end{aligned}
$$

where we used the fact that $M'$ satisfies $\mathbf{Pr}[|z|\ge M'] \le \epsilon^2/(CBL)^2$ for some large

absolute constant $C$; see the definition of $M'$ in (Grad). When $\sin\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2} \geq 3\sqrt{\mathrm{OPT}} \geq 3\sqrt{\epsilon}$, since $C$ is a large absolute constant and $L$ is a constant larger than 1, we have $\epsilon/(CBL) \leq (1/24B)\sin\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$. Therefore, it holds

$$\begin{aligned}(Q_1) &\geq (2/3)\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 - B\,\mathbf{E}[|\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{x}|](1/24B)\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 \\ &\geq (7/12)\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2.\end{aligned}$$

For the term $(Q_2)$, using Claim 8.2.15 again we have

$$\begin{aligned}|(Q_2)| &\leq \sum_{i=1}^{I} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|y||\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{x}||\mathrm{T}_{\cos\theta}\sigma'(a_i) - \mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \in \mathcal{E}_i\}] \\ &\leq \sum_{i=1}^{I} B\,\mathbf{E}[|\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{x}|]\,\mathbf{E}[|\mathrm{T}_{\cos\theta}\sigma'(a_i) - \mathrm{T}_{\cos\theta}\sigma'(\mathbf{w}\cdot\mathbf{x})|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \in \mathcal{E}_i\}] \\ &\leq \sum_{i=1}^{M} B(\epsilon/B)\,\mathbf{Pr}[z \in \mathcal{E}_i] \leq \epsilon.\end{aligned}$$

Since $\epsilon \leq \mathrm{OPT} \leq (1/12)\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$, we obtain that $|(Q_2)| \leq (1/12)\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$.

Plugging $(Q_1), (Q_2)$ back into Equation (8.6) yields:

$$(\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^*)^{1/2} \geq \frac{7}{24}\sin\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2} - \frac{1}{24}\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2} \geq \frac{1}{4}\sin\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}.$$

Thus, we have $\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^* \geq (1/16)\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$. □

The next lemma shows that any vector $\mathbf{u}$ that is orthogonal to $\mathbf{v}_{\mathbf{w}}^*$ has a small quadratic form.

**Lemma 8.2.16.** *Let* $\mathbf{u}$ *be any unit vector that is orthogonal to* $\mathbf{v}_{\mathbf{w}}^*$. *Then* $\mathbf{u}^\intercal\mathbf{M}_{\mathbf{w}}\mathbf{u} \leq 2\mathrm{OPT}$.

*Proof.* First, if $\mathbf{u} = \mathbf{w}$, then since $\mathbf{g}_{\mathbf{w}}(z) \cdot \mathbf{w} = 0$, we have $\mathbf{w}^\intercal \mathbf{M}_{\mathbf{w}} \mathbf{w} = 0 \leq \mathrm{OPT}$. Now consider $\mathbf{u} \perp \mathbf{w}$. Since $\mathbf{u} \perp \mathbf{w}, \mathbf{v}^*_{\mathbf{w}}$ and since $\mathbf{w}^* = \cos\theta \mathbf{w} + \sin\theta \mathbf{v}^*_{\mathbf{w}}$, we have that $\mathbf{u} \cdot \mathbf{x}$ is independent of $\mathbf{w}^* \cdot \mathbf{x}$. Direct calculation gives (noting again that $\mathbf{Pr}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}[\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_j] = \mathbf{Pr}_{z \sim \mathcal{N}}[z \in \mathcal{E}_j]$ since $\mathcal{D}_{\mathbf{x}}$ is standard Gaussian):

$$
\begin{aligned}
&\mathbf{u}^\intercal \mathbf{M}_{\mathbf{w}} \mathbf{u} = \mathbf{E}[(\mathbf{u} \cdot \mathbf{g}_{\mathbf{w}}(z))^2] \\
&= \mathop{\mathbf{E}}_{z \sim \mathcal{N}} \left[ \left( \sum_{i=1}^{I} \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [y(\mathbf{u} \cdot \mathbf{x}) \mid \mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i] \mathbb{1}\{z \in \mathcal{E}_i\} \right)^2 \right] \\
&= \sum_{i=1}^{I} \left( \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} \left[ y(\mathbf{u} \cdot \mathbf{x}) \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i\} \right] / \mathbf{Pr}[\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i] \right)^2 \mathbf{Pr}[z \in \mathcal{E}_i] \\
&= \sum_{i=1}^{I} \frac{1}{\mathbf{Pr}[z \in \mathcal{E}_i]} \Bigg( \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))(\mathbf{u} \cdot \mathbf{x}) \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i\}] \\
&\qquad\qquad + \mathbf{E}[\sigma(\mathbf{w}^* \cdot \mathbf{x})(\mathbf{u} \cdot \mathbf{x}) \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i\}] \Bigg)^2 \\
&= \sum_{i=1}^{I} \frac{1}{\mathbf{Pr}[z \in \mathcal{E}_i]} \left( \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))(\mathbf{u} \cdot \mathbf{x}) \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i\}] \right)^2,
\end{aligned}
$$

where in the last inequality we used that $\mathbf{u}$ is orthogonal to $\mathbf{w}, \mathbf{w}^*$. Furthermore, it holds:

$$
\begin{aligned}
&\left( \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))(\mathbf{u} \cdot \mathbf{x}) \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i\}] \right)^2 \\
&\overset{(i)}{\leq} \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i\}] \, \mathbf{E}[(\mathbf{u} \cdot \mathbf{x})^2 \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i\}] \\
&\overset{(ii)}{\leq} 2 \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [(y - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2 \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_i\}] \, \mathbf{Pr}[z \in \mathcal{E}_i],
\end{aligned}
$$

where in $(i)$ we applied Cauchy-Schwarz and in $(ii)$ we used the independence between

$\mathbf{u}\cdot\mathbf{x}$ and $\mathbf{w}\cdot\mathbf{x}$ since $\mathbf{u}\perp\mathbf{w}$. Therefore,

$$\begin{aligned}\mathbf{u}^\intercal\mathbf{M}_\mathbf{w}\mathbf{u} &\le \sum_{i=1}^{I} 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_i\}]\\ &\le 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in[-M',M']\}]\le 2\mathrm{OPT}.\end{aligned}$$

□

Then, we can show that the top eigenvector $\mathbf{v}_\mathbf{w}$ of $\mathbf{M}_\mathbf{w}$ correlates strongly with the target direction $\mathbf{v}_\mathbf{w}^*$.

**Lemma 8.2.17.** *Let* $\mathbf{v}_\mathbf{w}$ *be the top eigenvector of* $\mathbf{M}_\mathbf{w}$. *Suppose* $\sin\theta \ge 40\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}$. *Then* $\mathbf{v}_\mathbf{w}\cdot\mathbf{v}_\mathbf{w}^* \ge \sqrt{3}/2$.

*Proof.* Since $\mathbf{v}_\mathbf{w}$ is the top eigenvector of $\mathbf{M}_\mathbf{w}$, by Lemma 8.2.14, it holds $\mathbf{v}_\mathbf{w}^\intercal\mathbf{M}_\mathbf{w}\mathbf{v}_\mathbf{w} \ge (\mathbf{v}_\mathbf{w}^*)^\intercal\mathbf{M}_\mathbf{w}\mathbf{v}_\mathbf{w}^* \ge (4/9)\sin^2\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2 > 0$. Since $\mathbf{M}\mathbf{w}=0$, it must hold $\mathbf{v}_\mathbf{w}\cdot\mathbf{w}=0$, otherwise we would be able to find another eigenvector that has a larger eigenvalue. Suppose $\mathbf{v}_\mathbf{w} = a\mathbf{v}_\mathbf{w}^* + b\mathbf{u}$ where $\mathbf{u}\perp\mathbf{w},\mathbf{w}^*$ and $a^2+b^2=1$. Then, we obtain

$$a^2(\mathbf{v}_\mathbf{w}^*)^\intercal\mathbf{M}_\mathbf{w}\mathbf{v}_\mathbf{w}^* + b^2\mathbf{u}^\intercal\mathbf{M}_\mathbf{w}\mathbf{u} + 2ab\mathbf{u}^\intercal\mathbf{M}_\mathbf{w}\mathbf{v}_\mathbf{w}^* = \mathbf{v}_\mathbf{w}^\intercal\mathbf{M}_\mathbf{w}\mathbf{v}_\mathbf{w} \ge (\mathbf{v}_\mathbf{w}^*)^\intercal\mathbf{M}_\mathbf{w}\mathbf{v}_\mathbf{w}^*.$$

By Lemma 8.2.16 we have $\mathbf{u}^\intercal\mathbf{M}_\mathbf{w}\mathbf{u}\le\mathrm{OPT}$. Furthermore, note that by Cauchy-Schwarz

$$\begin{aligned}\mathbf{u}^\intercal\mathbf{M}_\mathbf{w}\mathbf{v}_\mathbf{w}^* &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathbf{u}\cdot\mathbf{g}_\mathbf{w}(z))(\mathbf{v}_\mathbf{w}^*\cdot\mathbf{g}_\mathbf{w}(z))] \le \sqrt{\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathbf{u}\cdot\mathbf{g}_\mathbf{w}(z))^2]\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(\mathbf{v}_\mathbf{w}^*\cdot\mathbf{g}_\mathbf{w}(z))^2]}\\ &= \sqrt{\mathbf{u}^\intercal\mathbf{M}_\mathbf{w}\mathbf{u}}\sqrt{(\mathbf{v}_\mathbf{w}^*)^\intercal\mathbf{M}_\mathbf{w}\mathbf{v}_\mathbf{w}^*} \le \sqrt{\mathrm{OPT}}\sqrt{(\mathbf{v}_\mathbf{w}^*)^\intercal\mathbf{M}_\mathbf{w}\mathbf{v}_\mathbf{w}^*}.\end{aligned}$$

Since $a^2 + b^2 = 1$, we get

$$(1-a^2)\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^* = b^2\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^* \leq b^2\mathrm{OPT} + 2ab\sqrt{\mathrm{OPT}}\sqrt{\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^*},$$

this implies that

$$\mathbf{v}_{\mathbf{w}} \cdot \mathbf{v}_{\mathbf{w}}^* = a \geq \frac{b}{2}\frac{\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^* - \mathrm{OPT}}{\sqrt{\mathrm{OPT}}\sqrt{\mathbf{v}_{\mathbf{w}}^* \cdot \mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^*}}.$$

Recall that $(\mathbf{v}_{\mathbf{w}}^*)^\intercal \mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^* \geq (1/16)\sin^2\theta \|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2}^2$. Since we have assumed $\sin\theta\|\mathrm{T}_{\cos\theta}\sigma'\|_{L_2} \geq 40\sqrt{\mathrm{OPT}}$, it holds $(\mathbf{v}_{\mathbf{w}}^*)^\intercal \mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^* \geq 100\mathrm{OPT}$. Thus, we obtain $\mathbf{v}_{\mathbf{w}} \cdot \mathbf{v}_{\mathbf{w}}^* = \cos(\theta(\mathbf{v}_{\mathbf{w}}, \mathbf{v}_{\mathbf{w}}^*)) \geq \sin(\theta(\mathbf{v}_{\mathbf{w}}, \mathbf{v}_{\mathbf{w}}^*))(99/20) \geq 3$, hence $\tan(\theta(\mathbf{v}_{\mathbf{w}}, \mathbf{v}_{\mathbf{w}}^*)) \leq 1/3$ and therefore $\mathbf{v}_{\mathbf{w}} \cdot \mathbf{v}_{\mathbf{w}}^* \geq \sqrt{3}/2$. □

### 8.2.3 Determining the Sample Complexity

Since we only have access to the empirical estimate $\widehat{\mathbf{M}}_{\mathbf{w}}$, we will use the following Wedin's theorem to bound the error between the empirical top eigenvector $\widehat{\mathbf{v}}_{\mathbf{w}}$ and the population top eigenvector $\mathbf{v}_{\mathbf{w}}$:

**Fact 8.2.18** (Wedin's theorem)**.** *Let $\theta(\mathbf{v}_{\mathbf{w}}, \widehat{\mathbf{v}}_{\mathbf{w}})$ be the angle between the top eigenvectors $\mathbf{v}_{\mathbf{w}} \in \mathbb{R}^d$ and $\widehat{\mathbf{v}}_{\mathbf{w}} \in \mathbb{R}^d$ of $\mathbf{M}_{\mathbf{w}}$ and $\widehat{\mathbf{M}}_{\mathbf{w}}$ respectively. Let $\rho_1$ and $\rho_2$ be the first 2 eigenvalues of $\mathbf{M}_{\mathbf{w}}$. Then, it holds that:*

$$\sin(\theta(\mathbf{v}_{\mathbf{w}}, \widehat{\mathbf{v}}_{\mathbf{w}})) \leq \frac{\|\mathbf{M}_{\mathbf{w}} - \widehat{\mathbf{M}}_{\mathbf{w}}\|_2}{\rho_1 - \rho_2 - \|\mathbf{M}_{\mathbf{w}} - \widehat{\mathbf{M}}_{\mathbf{w}}\|_2}.$$

Next, we bound the eigengap between the top eigenvalue and the rest.

**Lemma 8.2.19.** *Suppose* $\sin\theta\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2} \geq 40\sqrt{\mathrm{OPT}}$. *Let* $\mathbf{p}$ *be any eigenvector of* $\mathbf{M}_{\mathbf{w}}$ *orthogonal to* $\mathbf{v}_{\mathbf{w}}$. *Then,* $\mathbf{v}_{\mathbf{w}}^{\intercal}\mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}} - \mathbf{p}^{\intercal}\mathbf{M}_{\mathbf{w}}\mathbf{p} \geq (1/24)\sin^2\theta\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2}^2 \geq 60\mathrm{OPT}$.

*Proof.* In Lemma 8.2.17, we showed that $\mathbf{v}_{\mathbf{w}} = a\mathbf{v}_{\mathbf{w}}^* + b\mathbf{u}$ with $\mathbf{u} \perp \mathbf{v}_{\mathbf{w}}^*$, $a \geq \sqrt{3}/2$, $a^2 + b^2 = 1$. Assume that $\mathbf{p} = a_1\mathbf{v}_{\mathbf{w}}^* + b_1\mathbf{u} + \mathbf{u}'$ where $\mathbf{u}'$ is a vector orthogonal to both $\mathbf{v}_{\mathbf{w}}^*$ and $\mathbf{u}$ and $a_1^2 + b_1^2 \leq 1$. Since $\mathbf{p}\cdot\mathbf{v}_{\mathbf{w}} = 0$, we have $aa_1 + bb_1 = 0$, which implies that $a^2a_1^2 = b^2b_1^2 \leq (1-a^2)(1-a_1^2)$. Rearranging the terms, it yields $a_1^2 \leq 1 - a^2 \leq 1/4$, and therefore we have $a_1 \leq 1/2$. Denote $b_1\mathbf{u} + \mathbf{u}' = \mathbf{v}'$, and we have $\mathbf{p} = a_1\mathbf{v}_{\mathbf{w}}^* + \mathbf{v}'$ and $\|\mathbf{v}'\|_2 \leq 1$. Since $\mathbf{v}' \perp \mathbf{v}_{\mathbf{w}}^*$, by Lemma 8.2.16 we have $(\mathbf{v}')^{\intercal}\mathbf{M}_{\mathbf{w}}\mathbf{v}' \leq 2\mathrm{OPT}\|\mathbf{v}'\|_2^2 \leq 2\mathrm{OPT}$. Then the eigenvalue of $\mathbf{p}$ is bounded above by

$$\begin{aligned}\mathbf{p}^{\intercal}\mathbf{M}_{\mathbf{w}}\mathbf{p} &= a_1^2\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^* + a_1\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}' + \mathbf{v}'\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}' \\ &\leq (1/4)\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^* + (1/2)\sqrt{2\mathrm{OPT}(\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^*)} + 2\mathrm{OPT},\end{aligned}$$

where in the last inequality we applied Cauchy-Schwarz. Since we have assumed $\sin\theta\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2} \geq 40\sqrt{\mathrm{OPT}}$, from Lem. 8.2.14 we get $\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^* \geq (1/16)\sin^2\theta\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2}^2 \geq 100\mathrm{OPT}$, therefore we obtain $\mathbf{p}^{\intercal}\mathbf{M}_{\mathbf{w}}\mathbf{p} \leq (1/3)\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^*$. Thus, the eigengap between $\mathbf{v}_{\mathbf{w}}\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}$ and $\mathbf{p}^{\intercal}\mathbf{M}_{\mathbf{w}}\mathbf{p}$ can be bounded above by

$$\mathbf{v}_{\mathbf{w}}^{\intercal}\mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}} - \mathbf{p}^{\intercal}\mathbf{M}_{\mathbf{w}}\mathbf{p} \geq (2/3)\mathbf{v}_{\mathbf{w}}^*\cdot\mathbf{M}_{\mathbf{w}}\mathbf{v}_{\mathbf{w}}^* \geq (1/24)\sin^2\theta\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2}^2 \geq (200/3)\mathrm{OPT}.$$

□

By Lemma 8.2.19 we immediately get that $\rho_1 - \rho_2 \geq c\sin^2\theta\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2}^2$, therefore, to guarantee that $\sin(\theta(\mathbf{v}_{\mathbf{w}}, \widehat{\mathbf{v}}_{\mathbf{w}})) \ll 1$, if suffices to ensure that $\|\mathbf{M}_{\mathbf{w}} - \widehat{\mathbf{M}}_{\mathbf{w}}\|_2 \leq \epsilon \lesssim \mathrm{OPT} \lesssim \sin^2\theta\|\mathrm{T}_{\cos\theta\sigma'}\|_{L_2}^2$.

**Lemma 8.2.20** (Sample Complexity). *Draw* $N = \Theta(d^2 B^{12} L^8/\epsilon^{10} \log(dBL/(\delta\epsilon)))$ *independent samples from* $\mathcal{D}$. *Let*

$$\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)}(\mathbf{x}^{(i)})^{\perp \mathbf{w}} \mathbb{1}\{\mathbf{w}\cdot\mathbf{x}^{(i)} \in \mathcal{E}_j\},\ j \in [I]\ ,$$
$$\widehat{\mathbf{M}}_{\mathbf{w}} \stackrel{\text{def}}{=} \sum_{j=1}^{I} \frac{\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)}(\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)})^{\intercal}}{\mathbf{Pr}[z \in \mathcal{E}_j]} = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\sum_{j=1}^{I} \frac{\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)}\mathbb{1}\{z \in \mathcal{E}_j\}}{\mathbf{Pr}[z \in \mathcal{E}_j]}\right)\left(\sum_{j=1}^{I} \frac{\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)}\mathbb{1}\{z \in \mathcal{E}_j\}}{\mathbf{Pr}[z \in \mathcal{E}_j]}\right)^{\intercal}\right].$$

*Then, with probability at least* $1-\delta$, *for any* $\|\mathbf{w}\|_2 = 1$, *we have* $\|\widehat{\mathbf{M}}_{\mathbf{w}} - \mathbf{M}_{\mathbf{w}}\|_2 \le \epsilon$.

*Proof.* Since $\mathbf{g}_{\mathbf{w}}(z)$ is a piecewise constant function on the intervals $\mathcal{E}_j = [a_j, a_{j+1})$ where $a_j = -M' + j\Delta$ and $\Delta = \epsilon^2/(BL)^2$, as defined in (Grad), it suffices to approximate the (vector) value of $\mathbf{g}_{\mathbf{w}}(z)$ on each interval. First, for any $j \in [I]$, we observe that $\mathbf{r}_j \stackrel{\text{def}}{=} y\mathbf{x}^{\perp\mathbf{w}}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \in \mathcal{E}_j\}$ is a sub-Gaussian random variable with parameter $B$. To see this, it suffices to show that for any unit vector $\mathbf{u}$ it holds $\|\mathbf{u}\cdot\mathbf{r}_j\|_{L_p} \lesssim B\sqrt{p}$ for any $p \ge 1$. Since $\mathbf{w}\cdot\mathbf{r}_j = 0$, we only need to consider $\mathbf{u}$ that is orthogonal to $\mathbf{w}$. Direct calculation yields:

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\mathbf{u}\cdot\mathbf{r}_j|^p] &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left|y(\mathbf{x}^{\perp\mathbf{w}}\cdot\mathbf{u})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \in \mathcal{E}_j\}\right|^p\right] \\ &\le B^p\,\mathbf{E}[|\mathbf{u}\cdot\mathbf{x}|^p]\,\mathbf{Pr}[z \in \mathcal{E}_j] \le B^p(c\sqrt{p})^p.\end{aligned}$$

Thus, $\|\mathbf{u}\cdot\mathbf{r}_j\|_{L_p} \lesssim B\sqrt{p}$ for any unit vector $\mathbf{u}$ and hence $\mathbf{r}_j$ is $B$-sub-Gaussian.

Now sample $N$ independent samples from $\mathcal{D}$, $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$, creating $N$ independent vectors

$$\mathbf{r}_j^{(i)} \stackrel{\text{def}}{=} y^{(i)}(\mathbf{x}^{(i)})^{\perp\mathbf{w}}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}^{(i)} \in \mathcal{E}_j\}.$$

Then we know that $(1/N)\sum_{i=1}^N(\mathbf{r}_j^{(i)} - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{r}_j])$ is a sub-Gaussian random vector with

parameter $B/\sqrt{N}$. Then, using standard sub-Gaussian vector concentration inequality, we obtain

$$\mathbf{Pr}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_j^{(i)} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{r}_j]\right\|_2 \geq s\right] \leq \exp\left(-\frac{cs^2N}{B^2d}\right).$$

We let $s = \epsilon/(CB)(\epsilon/BL)^4$. Observe that $\mathbf{Pr}[z \in \mathcal{E}_j] \geq (a_{j+1} - a_j)\exp(-a_{j+1}^2/2) = \Delta\exp(-a_{j+1}^2/2)$, where $\Delta = \epsilon^2/(BL)^2$. Since $|a_{j+1}| \leq M'$ for all $j \in [I-1]$, we have $\exp(-a_{j+1}^2/2) \gtrsim \exp(-(M')^2/2) \gtrsim \mathbf{Pr}[|z| \geq M'] = \epsilon^2/(BL)^2$. Thus, we have $\mathbf{Pr}[z \in \mathcal{E}_j] \geq (\epsilon/BL)^4$ and hence $s \leq (\epsilon/(CB))\,\mathbf{Pr}[z \in \mathcal{E}_j]$ for all $j \in [I]$. Therefore choosing $N = \Theta(dB^{12}L^8/\epsilon^{10}\log(1/\delta))$, we have that with probability at least $1-\delta$,

$$\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_j^{(i)} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{r}_j]\right\|_2 \leq \frac{\epsilon\,\mathbf{Pr}[z \in \mathcal{E}_j]}{CB},$$

where $C$ is a large absolute constant. However, since there are $I = \tilde{O}((BL/\epsilon)^2)$ pieces of intervals $\mathcal{E}_j$ (by the definition of $I$ in (Grad)), to guarantee that $(1/N)\sum_{i=1}^{N}\mathbf{r}_j$ is close to $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{r}_j]$ on every interval, setting $\delta \leftarrow \delta\epsilon^2/(BL)^2$ and applying a union bound, we obtain that using $N = \Theta(dB^{12}L^8/\epsilon^{10}\log(dBL/(\delta\epsilon)))$ samples, with probability at least $1-\delta$ it holds $\|(1/N)\sum_{i=1}^{N}\mathbf{r}_j - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{r}_j]\|_2 \lesssim \epsilon/(CB)\,\mathbf{Pr}[z \in \mathcal{E}_j]$ for every $j \in [I]$. Thus, letting

$$\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} y^{(i)}(\mathbf{x}^{(i)})^{\perp_{\mathbf{w}}}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}^{(i)} \in \mathcal{E}_j\} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_j^{(i)},$$

we have that with probability at least $1-\delta$, it holds $\|\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp_{\mathbf{w}}}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \in \mathcal{E}_j\}]\|_2 \leq \epsilon/(CBI)$, for any $j \in [I]$.

Now define

$$\widehat{\mathbf{M}}_{\mathbf{w}} \stackrel{\text{def}}{=} \sum_{j=1}^{I} \frac{\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} (\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)})^{\mathsf{T}}}{\mathbf{Pr}[z \in \mathcal{E}_j]} = \mathop{\mathbf{E}}_{z \sim \mathcal{N}} \left[ \left( \sum_{j=1}^{I} \frac{\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} \mathbb{1}\{z \in \mathcal{E}_j\}}{\mathbf{Pr}[z \in \mathcal{E}_j]} \right) \left( \sum_{j=1}^{I} \frac{\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} \mathbb{1}\{z \in \mathcal{E}_j\}}{\mathbf{Pr}[z \in \mathcal{E}_j]} \right)^{\mathsf{T}} \right].$$

Observe that since $\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} \perp \mathbf{w}$ we have $\mathbf{w} \cdot \widehat{\mathbf{M}}_{\mathbf{w}} \mathbf{w} = 0$. Similarly we have $\mathbf{w} \cdot \mathbf{M}_{\mathbf{w}} \mathbf{w} = 0$. Now consider any unit vector $\mathbf{u}$ that is orthogonal to $\mathbf{w}$. Then, by the definition of $\mathbf{g}_{\mathbf{w}}(z)$ and that $\mathbf{M}_{\mathbf{w}} = \mathbf{E}_{z \sim \mathcal{N}}[\mathbf{g}_{\mathbf{w}}(z) \mathbf{g}_{\mathbf{w}}(z)^{\mathsf{T}}]$, we have

$$\begin{aligned}
&\left| \mathbf{u}^{\mathsf{T}} (\widehat{\mathbf{M}}_{\mathbf{w}} - \mathbf{M}_{\mathbf{w}}) \mathbf{u} \right| = \left| \mathbf{u}^{\mathsf{T}} \widehat{\mathbf{M}}_{\mathbf{w}} \mathbf{u} - \mathbf{u}^{\mathsf{T}} \mathop{\mathbf{E}}_{z \sim \mathcal{N}} [\mathbf{g}_{\mathbf{w}}(z) \mathbf{g}_{\mathbf{w}}(z)^{\mathsf{T}}] \mathbf{u} \right| \\
&= \left| \sum_{j=1}^{I} \frac{(\mathbf{u} \cdot \widehat{\mathbf{g}}_{\mathbf{w}}^{(j)})^2}{\mathbf{Pr}[z \in \mathcal{E}_j]} - \sum_{j=1}^{I} \frac{(\mathbf{u} \cdot \mathbf{E}_{(\mathbf{x},y) \sim \mathcal{D}}[y \mathbf{x}^{\perp \mathbf{w}} \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_j\}])^2}{\mathbf{Pr}[z \in \mathcal{E}_j]} \right| \leq \\
&\sum_{j=1}^{I} \frac{1}{\mathbf{Pr}[z \in \mathcal{E}_j]} \left| \mathbf{u} \cdot (\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} - \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [y \mathbf{x}^{\perp \mathbf{w}} \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_j\}]) \right| \left| \mathbf{u} \cdot (\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} + \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [y \mathbf{x}^{\perp \mathbf{w}} \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_j\}]) \right| \\
&\leq \sum_{j=1}^{I} \frac{1}{\mathbf{Pr}[z \in \mathcal{E}_j]} \left\| \widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} - \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [y \mathbf{x}^{\perp \mathbf{w}} \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_j\}] \right\|_2 \left\| \widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} + \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [y \mathbf{x}^{\perp \mathbf{w}} \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_j\}] \right\|_2
\end{aligned}$$

Note that we have $\|\widehat{\mathbf{g}}_{\mathbf{w}}^{(j)} - \mathbf{E}_{(\mathbf{x},y) \sim \mathcal{D}}[y \mathbf{x}^{\perp \mathbf{w}} \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_j\}]\|_2 \leq \epsilon \, \mathbf{Pr}[z \in \mathcal{E}_j]/(CB)$, for any $j \in [I]$, therefore,

$$\begin{aligned}
&\left| \mathbf{u}^{\mathsf{T}} (\widehat{\mathbf{M}}_{\mathbf{w}} - \mathbf{M}_{\mathbf{w}}) \mathbf{u} \right| \\
&\leq \sum_{j=1}^{I} \frac{1}{\mathbf{Pr}[z \in \mathcal{E}_j]} \frac{\epsilon \, \mathbf{Pr}[z \in \mathcal{E}_j]}{CB} \left( \frac{\epsilon \, \mathbf{Pr}[z \in \mathcal{E}_j]}{CB} + 2 \left\| \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [y \mathbf{x}^{\perp \mathbf{w}} \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_j\}] \right\|_2 \right) \\
&\leq \sum_{j=1}^{I} \frac{\epsilon}{CB} \left( \frac{\epsilon \, \mathbf{Pr}[z \in \mathcal{E}_j]}{CB} + 2 \left\| \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim \mathcal{D}} [y \mathbf{x}^{\perp \mathbf{w}} \mathbb{1}\{\mathbf{w} \cdot \mathbf{x} \in \mathcal{E}_j\}] \right\|_2 \right).
\end{aligned}$$

Observe that $\| \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_j\}]\|_2 \lesssim B\,\mathbf{Pr}[z\in\mathcal{E}_j]$ since

$$\begin{aligned}\| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_j\}]\|_2 &= \sup_{\|\mathbf{u}\|_2=1} | \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{u}\cdot\mathbf{x}^{\perp\mathbf{w}})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_j\}]| \\ &\le \sup_{\|\mathbf{u}\|_2=1} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|y||(\mathbf{u}\cdot\mathbf{x}^{\perp\mathbf{w}})|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_j\}] \lesssim B\,\mathbf{Pr}[z\in\mathcal{E}_j],\end{aligned} \tag{8.7}$$

where we used the assumption that $|y|\le B$ and the facts that $\mathbf{u}\cdot\mathbf{x}^{\perp\mathbf{w}}$ and $\mathbf{w}\cdot\mathbf{x}$ are independent and that $|\mathbf{E}[\mathbf{u}\cdot\mathbf{x}^{\perp\mathbf{w}}]|\le 2$. Thus, in summary, we have

$$\left|\mathbf{u}^\intercal(\widehat{\mathbf{M}}_{\mathbf{w}}-\mathbf{M}_{\mathbf{w}})\mathbf{u}\right| \le \sum_{j=1}^{I}\frac{\epsilon}{CB}(\epsilon/(CB)+B)\,\mathbf{Pr}[z\in\mathcal{E}_j]\le\epsilon.$$

This implies that $\|\widehat{\mathbf{M}}_{\mathbf{w}}-\mathbf{M}_{\mathbf{w}}\|_2\le\epsilon$.

Finally, we show that a $\tilde{O}(\epsilon^3)$-net of $\mathbf{w}$'s covers all the functions $\mathbf{g}_{\mathbf{w}}(z)$ and consequently covers all the matrices $\mathbf{M}_{\mathbf{w}}$.

**Claim 8.2.21.** *Given a unit vector* $\mathbf{w}$*, let* $\mathbf{w}'$ *be a vector such that* $\|\mathbf{w}'\|_2=1$ *and* $\|\mathbf{w}'-\mathbf{w}\|_2\le\epsilon^3/(CB^4L^2\sqrt{\log(BL/\epsilon)})$ *for some large absolute constant* $C$*. Then, for all* $z\in\mathbb{R}$*, it holds that* $\|\mathbf{g}_{\mathbf{w}}(z)-\mathbf{g}_{\mathbf{w}'}(z)\|_2\le\epsilon/(CB)$*.*

*Proof.* For any unit vector $\mathbf{w}'$ such that $\|\mathbf{w}'-\mathbf{w}\|_2\le\epsilon^3/(CB^4L^2\sqrt{\log(BL/\epsilon)})\le\epsilon^3/(CB^4L^2M')$ and any unit vector $\mathbf{u}$, we have

$$\begin{aligned}&\mathbf{u}\cdot(\mathbf{g}_{\mathbf{w}}(z)-\mathbf{g}_{\mathbf{w}'}(z)) \\ &=\sum_{j=1}^{I}\frac{\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y((\mathbf{u}\cdot\mathbf{x}^{\perp\mathbf{w}})\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_j\}-(\mathbf{u}\cdot\mathbf{x}^{\perp\mathbf{w}'}\mathbb{1}\{\mathbf{w}'\cdot\mathbf{x}\in\mathcal{E}_j\})]}{\mathbf{Pr}[z\in\mathcal{E}_j]}\mathbb{1}\{z\in\mathcal{E}_j\}.\end{aligned}$$

Let $\mathbf{w}'=\mathbf{w}+\mathbf{q}$ such that $\|\mathbf{q}\|_2\le\epsilon^3/(CB^4L^2\sqrt{\log(BL/\epsilon)})$. Then, $\|\mathbf{u}^{\perp\mathbf{w}}-\mathbf{u}^{\perp\mathbf{w}'}\|_2=$

$\|(\mathbf{w}\cdot\mathbf{u})\mathbf{w}-(\mathbf{w}'\cdot\mathbf{u})\mathbf{w}'\|_2\le\|\mathbf{q}\|_2\le\epsilon/(CB^4L^2)$. Furthermore, note that

$$\begin{aligned}
&\mathbf{E}[|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in[a_j,a_{j+1})\}-\mathbb{1}\{\mathbf{w}'\cdot\mathbf{x}\in[a_j,a_{j+1})\}|]\\
&=\mathbf{E}[|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\ge a_j\}-\mathbb{1}\{\mathbf{w}'\cdot\mathbf{x}\ge a_j\}-(\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\ge a_{j+1}\}-\mathbb{1}\{\mathbf{w}'\cdot\mathbf{x}\ge a_{j+1}\})|]\\
&\le\mathbf{E}[|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\ge a_j\}-\mathbb{1}\{\mathbf{w}'\cdot\mathbf{x}\ge a_j\}|]+\mathbf{E}[|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\ge a_{j+1}\}-\mathbb{1}\{\mathbf{w}'\cdot\mathbf{x}\ge a_{j+1}\}|].
\end{aligned}$$

It is well-known that $\mathbf{Pr}[\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\ge t\}\ne\mathbb{1}\{\mathbf{w}'\cdot\mathbf{x}\ge t\}]\le\theta(\mathbf{w},\mathbf{w}')\exp(-t^2)/(2\pi)$ (see Fact C.11 from [DKTZ22b]). Therefore, since $\theta(\mathbf{w},\mathbf{w}')\lesssim\|\mathbf{w}-\mathbf{w}'\|_2\le\epsilon^3/(CB^4L^2M')$, for small $\epsilon$ we have

$$\begin{aligned}
\mathbf{E}[|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_j\}-\mathbb{1}\{\mathbf{w}'\cdot\mathbf{x}\in\mathcal{E}_j\}|]&\lesssim\frac{\epsilon^3}{B^4L^2M'}(\exp(-(a_{j+1}-\Delta)^2/2)+\exp(-a_{j+1}^2/2))\\
&\lesssim\frac{\epsilon}{B^2}\frac{\epsilon^2}{B^2L^2}\frac{\exp(-a_{j+1}^2)}{a_{j+1}}\le\frac{\epsilon}{B^2}\mathbf{Pr}[z\in\mathcal{E}_j].
\end{aligned}$$

Thus, suppose $z\in\mathcal{E}_j$, we have

$$\begin{aligned}
|\mathbf{u}\cdot(\mathbf{g}_{\mathbf{w}}(z)-\mathbf{g}_{\mathbf{w}'}(z))|&=\frac{1}{\mathbf{Pr}[z\in\mathcal{E}_j]}|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{u}^{\perp\mathbf{w}}-\mathbf{u}^{\perp\mathbf{w}'})\cdot\mathbf{x}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_j\}]|\\
&\quad+\frac{1}{\mathbf{Pr}[z\in\mathcal{E}_j]}|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{u}^{\perp\mathbf{w}}\cdot\mathbf{x}(\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_j\}-\mathbb{1}\{\mathbf{w}'\cdot\mathbf{x}\in\mathcal{E}_j\})]|\\
&\le\frac{1}{\mathbf{Pr}[z\in\mathcal{E}_j]}\mathbf{E}\left[B\left|\frac{\mathbf{u}^{\perp\mathbf{w}}-\mathbf{u}^{\perp\mathbf{w}'}}{\|\mathbf{u}^{\perp\mathbf{w}}-\mathbf{u}^{\perp\mathbf{w}'}\|_2}\cdot\mathbf{x}\right|\right]\mathbf{Pr}[z\in\mathcal{E}_j]\|\mathbf{u}^{\perp\mathbf{w}}-\mathbf{u}^{\perp\mathbf{w}'}\|_2\\
&\quad+\frac{B}{\mathbf{Pr}[z\in\mathcal{E}_j]}\mathbf{E}[|\mathbf{u}^{\perp\mathbf{w}}\cdot\mathbf{x}|]\,\mathbf{E}[|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_j\}-\mathbb{1}\{\mathbf{w}'\cdot\mathbf{x}\in\mathcal{E}_j\}|]\\
&\lesssim\epsilon/B
\end{aligned}$$

Thus, this implies $\|\mathbf{g}_{\mathbf{w}}(z)-\mathbf{g}_{\mathbf{w}'}(z)\|_2\lesssim\epsilon/B$. □

Note that when $\|\mathbf{g}_{\mathbf{w}}(z)-\mathbf{g}_{\mathbf{w}'}(z)\|_2\lesssim\epsilon/B$, we have (recall that in Equation (8.7)

we showed $\|\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}^{\perp\mathbf{w}}\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\in\mathcal{E}_j\}]/\mathbf{Pr}[z\in\mathcal{E}_j]\|_2 \le B$ hence by the definition of $\mathbf{g}_\mathbf{w}(z)$ we have $\|\mathbf{g}_\mathbf{w}(z)\|_2 \le B$):

$$\begin{aligned}\forall \mathbf{u} \text{ s.t. } \|\mathbf{u}\|_2 = 1: \qquad & |\mathbf{u}^\intercal \mathbf{g}_{\mathbf{w}'}(z)\mathbf{g}_{\mathbf{w}'}(z)^\intercal \mathbf{u} - \mathbf{u}^\intercal \mathbf{g}_\mathbf{w}(z)\mathbf{g}_\mathbf{w}(z)^\intercal \mathbf{u}| \\ &= |\mathbf{u}\cdot(\mathbf{g}_{\mathbf{w}'}(z)-\mathbf{g}_\mathbf{w}(z))||\mathbf{u}\cdot(\mathbf{g}_{\mathbf{w}'}(z)+\mathbf{g}_\mathbf{w}(z))| \\ &\le \|\mathbf{g}_{\mathbf{w}'}(z)-\mathbf{g}_\mathbf{w}(z)\|_2\|\mathbf{g}_{\mathbf{w}'}(z)+\mathbf{g}_\mathbf{w}(z)\|_2 \\ &\le \frac{\epsilon}{CB}(2\|\mathbf{g}_\mathbf{w}(z)\|_2+\epsilon/B) \le \epsilon,\end{aligned}$$

indicating that

$$\|\mathbf{M}_{\mathbf{w}'}-\mathbf{M}_\mathbf{w}\|_2 = \sup_{\|\mathbf{u}\|_2} \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\mathbf{u}^\intercal(\mathbf{g}_{\mathbf{w}'}(z)\mathbf{g}_{\mathbf{w}'}(z)^\intercal - \mathbf{g}_\mathbf{w}(z)\mathbf{g}_\mathbf{w}(z)^\intercal)\mathbf{u}] \le \epsilon.$$

Thus, constructing a $\tilde{O}(\epsilon^3/(B^4L^2))$-cover $S$ on the unit sphere and requiring $\|(1/n)\sum_{i=1}^n \mathbf{M}_{\mathbf{w}'}^{(i)} - \mathbf{M}_{\mathbf{w}'}\|_2 \le \epsilon$ on all $\mathbf{w}'\in S$ suffices. Since a $\tilde{O}(\epsilon^3/(B^4L^2))$-cover $S$ on the unit sphere contain $|S| = (O(\epsilon^3/(B^4L^2)))^d$ vectors, applying a union bound we obtain that using $N = \Theta(d^2B^{12}L^8/\epsilon^{10}\log(dBL/(\delta\epsilon)))$ samples, we guarantee that with probability at least $1-\delta$, for any unit vector $\mathbf{w}$, it holds $\|\widehat{\mathbf{M}}_\mathbf{w} - \mathbf{M}_\mathbf{w}\|_2 \le \epsilon$. □

### 8.2.4 Proof of Proposition 8.2.12

*Proof of Proposition 8.2.12.* Let $\mathbf{g}_\mathbf{w}(z)$ be the vector defined in (Grad) and let $\mathbf{M}_\mathbf{w} = \mathbf{E}_{z\sim\mathcal{N}}[\mathbf{g}_\mathbf{w}(z)\mathbf{g}_\mathbf{w}(z)^\intercal]$. In Lemma 8.2.17 we proved that for any vector $\mathbf{w}$, whenever $\sin(\theta(\mathbf{w},\mathbf{w}^*)) \ge 40\sqrt{\mathrm{OPT}}\|\mathrm{T}_{\cos\theta(\mathbf{w},\mathbf{w}^*)}\sigma'\|_{L_2}$, one of the top eigenvector $\mathbf{v}_\mathbf{w}$ of $\mathbf{M}_\mathbf{w}$ correlates with $\mathbf{v}_\mathbf{w}^*$: $\mathbf{v}_\mathbf{w}\cdot\mathbf{v}_\mathbf{w}^* \ge \sqrt{3}/2$, i.e., $\theta(\mathbf{v}_\mathbf{w}^*,\mathbf{v}_\mathbf{w}) \le \pi/6$. Next, in Lemma 8.2.20, we proved that using $N = \Theta(d^2B^{12}L^8/\epsilon^{10}\log(dBL/(\delta\epsilon)))$ samples, for any vector $\mathbf{w}$, the empirical

matrix $\widehat{\mathbf{M}}_{\mathbf{w}}$ satisfies $\|\widehat{\mathbf{M}}_{\mathbf{w}} - \mathbf{M}_{\mathbf{w}}\|_2 \leq \epsilon$. Therefore, using Fact 8.2.18, we know that for any vector $\mathbf{w}$, the top eigenvector $\hat{\mathbf{v}}_{\mathbf{w}}$ of $\widehat{\mathbf{M}}_{\mathbf{w}}$ satisfies $\sin(\theta(\mathbf{v}_{\mathbf{w}}, \hat{\mathbf{v}}_{\mathbf{w}})) \leq \epsilon/(\rho_1 - \rho_2 - \epsilon)$. Since in Lemma 8.2.19 we showed that when $\sin(\theta(\mathbf{w}, \mathbf{w}^*)) \geq 40\sqrt{\mathrm{OPT}}\|\mathrm{T}_{\cos\theta(\mathbf{w},\mathbf{w}^*)}\sigma'\|_{L_2}$, it holds $\rho_1 - \rho_2 \geq 60\mathrm{OPT} \geq 60\epsilon$, we then have that $\theta(\mathbf{v}_{\mathbf{w}}, \hat{\mathbf{v}}_{\mathbf{w}}) \leq 1/59$, indicating $\theta(\hat{\mathbf{v}}_{\mathbf{w}}, \mathbf{v}^*_{\mathbf{w}}) \leq \theta(\hat{\mathbf{v}}_{\mathbf{w}}, \mathbf{v}_{\mathbf{w}}) + \theta(\mathbf{v}_{\mathbf{w}}, \mathbf{v}^*_{\mathbf{w}}) \leq 1/59 + \pi/6 \leq \pi/4$. Therefore, let $\mathbf{u}$ be such eigenvector $\hat{\mathbf{v}}_{\mathbf{w}}$ of $\widehat{\mathbf{M}}_{\mathbf{w}}$ that correlates positively with $\mathbf{v}^*_{\mathbf{w}}$. Note that by definition of $\widehat{\mathbf{M}}_{\mathbf{w}}$, it must hold $\mathbf{u} \perp \mathbf{w}$. Thus we have $\mathbf{u} \cdot \mathbf{w}^* = \mathbf{u} \cdot (\cos(\theta(\mathbf{w}, \mathbf{w}^*))\mathbf{w} + \sin(\theta(\mathbf{w}, \mathbf{w}^*))\mathbf{v}^*_{\mathbf{w}}) = \sin(\theta(\mathbf{w}, \mathbf{w}^*))\mathbf{u} \cdot \mathbf{v}^*_{\mathbf{w}} \geq (\sqrt{2}/2)\sin(\theta(\mathbf{w}, \mathbf{w}^*))$. Let $\delta = 0.01$ finishes the proof. □

**Theorem 8.2.22** (Main Result)**.** *Let $\epsilon > 0$. Fix constants $B, L > 0$. Let $\mathcal{D}$ be a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Suppose there is a vector $\mathbf{w}^* \in \mathbb{R}^d$ and a monotone activation $\sigma \in \mathcal{H}(B, L)$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2] \leq \mathrm{OPT}$. Then Algorithm 16 runs at most $T = \mathrm{poly}(B, L, 1/\epsilon)$ iterations, draws $\Theta(d^2B^{12}L^8/\epsilon^{10}\log(dBL/\epsilon))$ samples, and returns a vector $\hat{\mathbf{w}}$ and a Lipschitz and monotone activation $u : \mathbb{R} \to \mathbb{R}$, such that with probability at least 99%, it holds that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\hat{\mathbf{w}}\cdot\mathbf{x}) - y)^2] \leq O(\mathrm{OPT}) + \epsilon$.*

*Proof of Theorem 8.2.22.* We first show that if we choose correctly the step size and the number of iterations in Algorithm 18, then for any inititialization $\mathbf{w}^{(0)}$, Algorithm 18 with high probability returns a vector $\hat{\mathbf{w}}$ with $\widehat{\theta} = \theta(\hat{\mathbf{w}}, \mathbf{w}^*)$ so that $\sin\widehat{\theta} \leq O(\sqrt{\mathrm{OPT}})/\|\mathrm{T}_{\cos\widehat{\theta}}\sigma'\|_{L_2}$ and $\widehat{\theta} \leq \bar{\theta} = \theta(\mathbf{w}^{(0)}, \mathbf{w}^*)$.

**Proposition 8.2.23.** *Let $\epsilon > 0$. Fix constants $B, L > 0$ and $\delta \in (0, 1)$. Let $\mathcal{D}$ be a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Suppose there is a vector $\mathbf{w}^* \in \mathbb{R}^d$ and an activation $\sigma \in \mathcal{H}_\epsilon(B, L)$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2] \leq \mathrm{OPT}$. Then Algorithm 18, given $\bar{\theta}$, initialization vector $\mathbf{w}^{(0)}$ so that $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \leq \bar{\theta}$ and $T \geq O(\log(L/\epsilon))$ then with probability at least $1 - \delta$ returns a list of vectors $S^{\mathrm{sol}}$ with*

*size* $|S^{\text{sol}}| = \text{poly}(1/\epsilon, L)\log(1/\delta)$ *such that: there exists* $\widehat{\mathbf{w}} \in S^{\text{sol}}$ *so that:* $\widehat{\theta} \leq \bar{\theta}$ *and* $\sin\widehat{\theta} \leq O(\sqrt{\text{OPT}})/\|\mathrm{T}_{\cos\widehat{\theta}}\sigma'\|_{L_2}$ *where* $\widehat{\theta} = \theta(\widehat{\mathbf{w}}, \mathbf{w}^*)$.

*Proof of Proposition 8.2.23.* In the proof, we denote the angle between $\mathbf{w}^{(t)}$ and $\mathbf{w}^*$ by $\theta_t = \theta(\mathbf{w}^{(t)}, \mathbf{w}^*)$. Furthermore, the algorithm uses the following parameters: $\phi_t = \bar{\theta}(1 - c^2/32)^t$ and $\eta_t = c\sin\phi_t/4$ where $c = 1/4 \leq \sqrt{2}/2$. Note that if $\epsilon \geq C\text{OPT}$, then we can run the algorithm with $\epsilon' = \epsilon/(2C)$ and assume that we have more noise of order $\text{OPT}' = 2\epsilon'$. In this case, the final error bound will be $C\text{OPT}' \leq \epsilon/2 \leq \text{OPT} + \epsilon$. So, without loss of generality, we can assume that $\epsilon \leq \text{OPT}$. According to Proposition 8.2.12 as long as $\sin\theta_t \geq 40\sqrt{\text{OPT}}/\|\mathrm{T}_{\cos\theta_t}\sigma'\|_{L_2}$, with probability at least $1/2$, the vector $\mathbf{v}^{(t)}$ returned at Line (9) of Algorithm 18 satisfies $\mathbf{v}^{(t)} \cdot \mathbf{w}^* \leq -c\sin\theta_t$ and $\mathbf{v}^{(t)} \cdot \mathbf{w}^{(t)} = 0$. We denote as $\mathcal{P}_t$ the event that $\mathbf{v}^{(t)}$ negatively correlates with $\mathbf{w}^*$. We consider the following event

$$\mathcal{R}_t \stackrel{\text{def}}{=} \left\{ \sin\theta_t \geq \frac{C\sqrt{\text{OPT}}}{\|\mathrm{T}_{\cos\theta_t}\sigma'\|_{L_2}} \right\},$$

where $C > 0$ is an absolute constant.

First, we show that conditioning on the events $\mathcal{R}_t, \mathcal{P}_t$, for all $t \in T$, it holds that $\phi_t \geq \theta_t$.

**Claim 8.2.24.** *Suppose the events* $\mathcal{R}_t, \mathcal{P}_t$, $t \in [T_1]$, *all hold for some* $T_1 \geq 1$. *Then for all* $t \in [T_1]$, *it holds that* $\phi_t \geq \theta_t$.

*Proof of Claim 8.2.24.* We use induction for this proof. By assumption, we have that $\phi_0 \geq \theta_0$. Next, we assume that $\phi_t \geq \theta_t$. We need to show that $\phi_{t+1} \geq \theta_{t+1}$. We study the distance between $\mathbf{w}^{(t)}$ and $\mathbf{w}^*$ after one iteration from $t$ to $t+1$. Since $\mathbf{v}^{(t)}$ is orthogonal to $\mathbf{w}^{(t)}$, it must be $\|\mathbf{w}^{(t)} - \eta_t\mathbf{v}^{(t)}\|_2 \geq 1$, therefore, $\mathbf{w}^{(t+1)} = \text{proj}_{\mathbb{B}}(\mathbf{w}^{(t)} - \eta_t\mathbf{v}^{(t)})$. By the

non-expansiveness of the projection operator, we have

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 &= \|\mathrm{proj}_{\mathbb{B}}(\mathbf{w}^{(t)} - \eta_t \mathbf{v}^{(t)}) - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}^{(t)} - \eta_t \mathbf{v}^{(t)} - \mathbf{w}^*\|_2^2 \\ &= \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \eta_t^2 \|\mathbf{v}^{(t)}\|_2^2 - 2\eta_t \mathbf{v}^{(t)} \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) \\ &= \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \eta_t^2 + 2\eta_t \mathbf{v}^{(t)} \cdot \mathbf{w}^*. \end{aligned} \tag{8.8}$$

Note that $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 = 2(\cos\theta_{t+1} - \cos\theta_t)$ and using the identity about the sum of cosines, we have

$$4 \sin\left(\frac{\theta_{t+1} - \theta_t}{2}\right) \sin\left(\frac{\theta_{t+1} + \theta_t}{2}\right) \leq \eta_t^2 + 2\eta_t \mathbf{v}^{(t)} \cdot \mathbf{w}^*.$$

First, consider the case where $2\theta_t \geq \phi_t \geq \theta_t$. From Proposition 8.2.12, we have that $\mathbf{v}^{(t)} \cdot \mathbf{w}^* \leq -c \sin\theta_t$ where $c > 0$ is an absolute constant. Hence, since we chose $\eta_t = c \sin\phi_t/4$, it holds $\eta_t^2 + 2\eta_t \mathbf{v}^{(t)} \cdot \mathbf{w}^* \leq -c^2 \sin\phi_t \sin\theta_t/4$. Therefore, in this case, $\theta_{t+1} \leq \theta_t$ hence $\sin((\theta_{t+1} + \theta_t)/2) \leq \sin\theta_t$, and we have that

$$16 \sin\left(\frac{\theta_t - \theta_{t+1}}{2}\right) \geq c^2 \sin\phi_t \ .$$

Using the inequality $x/4 \leq \sin x \leq x$ for $x \in (0, \pi/2)$, we get that

$$\theta_{t+1} \leq \theta_t(1 - c^2/32) \ ,$$

and using that $\phi_{t+1} = \phi_t(1 - c^2/32)$, we have that $\theta_{t+1} \leq \phi_{t+1}$.

Consider the remaining case where $\phi_t \geq 2\theta_t$. In this case, if $\theta_{t+1} \leq \theta_t$, then $\theta_{t+1} \leq \phi_{t+1}$ so we need to consider the case where $\theta_{t+1} \geq \theta_t$. We need to bound the

maximum increase of $\theta_{t+1}$. Applying the triangle inequality and the non-expansiveness of the projection operator, it holds

$$2\sin(\theta_{t+1}/2) = \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 = \|\mathrm{proj}_{\mathbb{B}}(\mathbf{w}^{(t)} - \eta_t \mathbf{v}^{(t)}(\mathbf{w}^{(t)})) - \mathbf{w}^*\|_2 \le \|\mathbf{w}^{(t)} - \eta_t \mathbf{v}^{(t)} - \mathbf{w}^*\|_2$$
$$\le \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 + \eta_t \|\mathbf{v}^{(t)}\|_2 = 2\sin(\theta_t/2) + c\sin\phi_t/4 \,.$$

From the assumption, we have $\theta_t \le \phi_t/2$, therefore, choosing $c \le 1/4$ and since $\sin(x) \le x$ for $x \in (0, \pi/2)$, we have that

$$\sin(\theta_{t+1}/2) \le \sin(\phi_t/4) + c\sin\phi_t/8 \le 9\phi_t/32 \,.$$

Therefore, since $(5/8)x \le \sin x$ when $x \in (0, \pi/2)$ we have $\sin(\theta_{t+1}/2) \ge (5/16)\theta_{t+1}$ and thus, $\theta_{t+1} \le (9/10)\phi_t \le (1 - c^2/32)\phi_t \le \phi_{t+1}$. This completes the proof. □

Next, we condition on the event that all $\mathcal{P}_t$ are satisfied for $t \in [T]$. According to Claim 8.2.24, as long as $\mathcal{R}_t$ is satisfied, we have $\phi_t \ge \theta_t$. Assume $\mathcal{R}_t$ is satisfied for all $t \in [T]$. If $T \ge C' \log(1/(L\epsilon))$ for a sufficiently large constant $C'$, then $\phi_T \le \sqrt{\epsilon}/L$. This would imply $\theta_T \le \sqrt{\epsilon}/L$. If $\mathcal{R}_T$ was satisfied, $\sin\theta_T \ge C\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos\theta_T}\sigma'\|_{L_2}$. So $\theta_T \ge \sin\theta_T \ge C\sqrt{\mathrm{OPT}}/L$ (since $\|\mathrm{T}_{\cos\theta_T}\sigma'\|_{L_2} \le \|\sigma'\|_{L_2} \le L$, by Fact 7.1.7(f)). Then $\sqrt{\epsilon}/L \ge C\sqrt{\mathrm{OPT}}/L$, which means $\sqrt{\epsilon} \ge C\sqrt{\mathrm{OPT}}$. If $\epsilon \le \mathrm{OPT}$, this is a contradiction for $C > 1$. This means that our assumption that $\mathcal{R}_t$ are satisfied for all $t \in [T]$ must be false. Therefore, there must exist some $T_1 \in [T]$ (we take $T_1$ to be the smallest one) such that $\mathcal{R}_{T_1}$ is not satisfied. For all $t < T_1$, $\mathcal{R}_t$ was satisfied (otherwise $T_1$ would be smaller), and thus $\phi_t \ge \theta_t$ for $t < T_1$. At $t = T_1$, $\mathcal{R}_{T_1}$ is false, meaning $\sin\theta_{T_1} < C\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos\theta_{T_1}}\sigma'\|_{L_2}$, and we also have $\theta_{T_1} \le \phi_{T_1} \le \bar{\theta}$. This gives us the desired vector $\widehat{\mathbf{w}} = \mathbf{w}^{(T_1)}$ such that $\widehat{\theta} \le C\sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos\widehat{\theta}}\sigma'\|_{L_2}$.

To conclude, we need to bound the probability that all $\mathcal{P}_t$ (correct direction choices) are satisfied. The events $\mathcal{P}_t$ are independent, and each occurs with probability at least $1/2$. The probability of $T_1$ such events occurring is at least $(1/2)^{T_1}$. Since $T_1 \leq T = O(\log(L/\epsilon))$, this probability is bounded below by $\delta' = (1/2)^T = \mathrm{poly}(\epsilon, 1/L)$. If we rerun the algorithm $K = O((1/\delta')\log(1/\delta))$ times ( Line 5 of Algorithm 18), by standard Chernoff bounds, with probability at least $1-\delta$, there will be at least one run where all $\mathcal{P}_t$ are satisfied for $t \in [T_1]$. □

In Lemma 8.2.11, we showed that the initialization algorithm (Algorithm 17) uses $\tilde{O}(d/\epsilon^2)$ samples and returns a list of vectors $S^{\mathrm{ini}}$, $|S^{\mathrm{ini}}| \leq B/\sqrt{\epsilon}$ that contains a vector $\mathbf{w}^{(0)}$ such that $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \leq 1/M$ and $M$ is the threshold we can truncate $\sigma$ such that $\mathbf{E}_{z\sim\mathcal{N}}[(\sigma(z) - \sigma(M))^2 \mathbb{1}\{|z| \geq M\}] \leq C(\mathrm{OPT} + \epsilon)$, i.e., we can truncate $\sigma$ at $M$ and the overall error is increased by $C(\mathrm{OPT} + \epsilon)$. By Fact 8.2.5, this initialized vector $\mathbf{w}^{(0)}$ ensures that for any unit vector $\mathbf{w}^{(t)}$ such that $\theta_t \stackrel{\mathrm{def}}{=} \theta(\mathbf{w}^{(t)}, \mathbf{w}^*) \leq \theta(\mathbf{w}^{(0)}, \mathbf{w}^*)$, it holds $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}) - y)^2] \leq C\mathrm{OPT} + \sin^2\theta_t \|\mathrm{T}_{\cos\theta_t}\sigma'\|_{L_2}^2$. Therefore, any unit vector $\mathbf{w}^{(t)}$ such that $\theta_t \leq \theta(\mathbf{w}^{(0)}, \mathbf{w}^*)$ and $\sin^2\theta_t \|\mathrm{T}_{\cos\theta_t}\sigma'\|_{L_2}^2 \leq C\mathrm{OPT}$ is a constant factor approximate solution (i.e., it holds $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}) - y)^2] \leq C\mathrm{OPT} + \epsilon$). Since we are iterating through this list $S^{\mathrm{ini}}$, it is guaranteed we will encounter the correct $\mathbf{w}^{(0)}$. Hence it remains to show that we can choose the correct stepsize in Algorithm 18. If we choose $\bar{\theta}$ from the list $\Theta = \{\epsilon/L, \ldots k\epsilon/L\}$ for $k = L/\epsilon$, it is guaranteed that for each $\mathbf{w}^{(0)}$ in the initialization list, there exists $\bar{\theta} \in \Theta$ so that $\theta(\mathbf{w}^{(0)}, \mathbf{w}^*) \in (\bar{\theta}, 2\bar{\theta})$ (see (5) of Algorithm 16). That means according to Proposition 8.2.23, setting $\delta = 0.01$, with probability at least 99% we have that there exist $\widehat{\mathbf{w}}$ in the list $S^{\mathrm{sol}}$, returned by Algorithm 18, that satisfies that $\widehat{\theta} = \theta(\widehat{\mathbf{w}}, \mathbf{w}^*) \leq \bar{\theta}$ and $\sin\widehat{\theta} \leq O(\sqrt{\mathrm{OPT}})/\|\mathrm{T}_{\cos\widehat{\theta}}\sigma'\|_{L_2}$, indicating that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}) - y)^2] \leq C\mathrm{OPT} + \epsilon$

Finally, using Lemma 8.3.3, we know that drawing at most $n = \Theta(\log(BL/\epsilon)B^3L/\epsilon^{3/2})$ new samples, with probability at least 99%, the testing algorithm (Algorithm 19) returns a solution pair $(\widehat{u}_{\widehat{\mathbf{w}}}, \widehat{\mathbf{w}})$ where $\widehat{u}_{\widehat{\mathbf{w}}} \in \mathcal{H}(B, L)$ and $\widehat{\mathbf{w}} \in S^{\mathrm{sol}}$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\widehat{u}_{\widehat{\mathbf{w}}}(\widehat{\mathbf{w}} \cdot \mathbf{x}) - y)^2] \leq C\mathrm{OPT} + \epsilon$. This completes the proof. □

## 8.3 Testing Subroutine

**Algorithm 19** Testing

**Input:** Parameters $B$, $L$, $\epsilon$; data access $(\mathbf{x}, y) \sim \mathcal{D}$; candidate set $S^{\mathrm{sol}}$
Draw $n$ i.i.d. samples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n \sim \mathcal{D}$ and form empirical distribution $\widehat{\mathcal{D}}_n$
**for** $\mathbf{w} \in S^{\mathrm{sol}}$ **do**
    $\widehat{u}_{\mathbf{w}} \leftarrow \arg\min_{u\in\mathcal{H}(B,L)} \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}_n}[(u(\mathbf{w}\cdot\mathbf{x}) - y)^2]$
    $S \leftarrow S \cup \{(\widehat{u}_{\mathbf{w}}, \mathbf{w})\}$
**end for**
**Return:** $(\widehat{u}_{\widehat{\mathbf{w}}}, \widehat{\mathbf{w}}) = \arg\min_{(u_{\mathbf{w}},\mathbf{w})\in S} \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}_n}[(u_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x}) - y)^2]$

Algorithm 16 generates a list of possible parameters $S^{\mathrm{sol}} = \{\mathbf{w}^{(i)}\}_{i=1}^m$, where $m = \mathrm{poly}(1/\epsilon, B, L)$, and we know that there exists a vector $\widehat{\mathbf{w}} \in S^{\mathrm{sol}}$ such that $\sin\theta \leq \sqrt{\mathrm{OPT}}/\|\mathrm{T}_{\cos(\theta)\sigma'}\|_{L_2}$, where $\theta = \theta(\widehat{\mathbf{w}}, \mathbf{w}^*)$. To complete the task of learning SIMs, we need to: (1) find an activation $u \in \mathcal{H}(B, L)$ that is close to the target activation $\sigma$; (2) find the target vector $\widehat{\mathbf{w}} \in S^{\mathrm{sol}}$.

First, we note that given any vector $\mathbf{w}$ and $n$ samples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$, there exists an efficient algorithm that computes a best fitting monotone and $\beta$-Lipschitz function on the sample set $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$, via solving the following constrained optimization

problem:

$$\begin{aligned} &\min_{v_i, i\in[n]} \sum_{i=1}^{n} (v_i - y^{(i)})^2 \\ &\text{s.t. } 0 \le v_{i+1} - v_i \le \beta(\mathbf{w}\cdot\mathbf{x}^{(i+1)} - \mathbf{w}\cdot\mathbf{x}^{(i)}). \end{aligned} \tag{Iso}$$

We remark that $\mathbf{x}^{(i)}$'s are sorted so that $\mathbf{w}\cdot\mathbf{x}^{(i)}$'s are in increasing order.

We use the following fact:

**Fact 8.3.1** (Proposition 1 [HTY24])**.** *Given a sample set $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$, there exists an algorithm that exactly solves* (Iso) *in $O(n\log^2(n))$ time.*

Observe that given the solution $\{v_i\}_{i=1}^n$ of (Iso), we can construct a function $\widehat{u}_{\mathbf{w}}(z)$ by linearly interpolating the points $\{(z_i, v_i)\}_{i=1}^n$ where $z_i = \mathbf{w}\cdot\mathbf{x}_i$. Then, the function $\widehat{u}_{\mathbf{w}}(z)$ is guaranteed to be a monotone and $\beta$-Lipschitz function. In the following claim, we show that the function class $\sigma \in \mathcal{H}(B, L)$ is covered by Lipschitz-continuous functions.

**Claim 8.3.2.** *It is without loss of generality to assume that $\tilde{\sigma} \in \mathcal{H}_\epsilon(B, L)$ is $BL/\sqrt{\epsilon}$-Lipschitz.*

*Proof.* Let $\sigma \in \mathcal{H}(B, L)$ such that $\|\sigma - \tilde{\sigma}\|_{L_2}^2 \le \epsilon$. By Lemma 7.1.10, we have that for any $\rho \in (0, 1)$ it holds $\|\mathrm{T}_\rho\sigma - \sigma\|_{L_2}^2 \le (1-\rho^2)\|\sigma'\|_{L_2}^2 \le (1-\rho^2)L^2$. Therefore, choosing $\rho^2 = 1 - \epsilon/L^2$ we have that for any function $\sigma \in \mathcal{H}(B, L)$, there exists a function $\mathrm{T}_\rho\sigma \in \mathcal{H}(B, L)$ such that $\|\mathrm{T}_\rho\sigma - \sigma\|_{L_2}^2 \le \epsilon$ and hence $\|\mathrm{T}_\rho\sigma - \tilde{\sigma}\|_{L_2}^2 \le 2\epsilon$. Furthermore, the function $\mathrm{T}_\rho\sigma(z)$ is $BL/\sqrt{\epsilon}$-Lipschitz since according to Fact 7.1.7 part (c) we have $\|(\mathrm{T}_\rho\sigma)'\|_{L_\infty} \le \|\sigma\|_{L_\infty}/\sqrt{1-\rho^2} \le BL/\sqrt{\epsilon}$. Therefore, the function class $\mathcal{H}_\epsilon(B, L)$ is covered by $BL/\sqrt{\epsilon}$-Lipschitz functions and hence it is without loss of generality to assume that $\sigma \in \mathcal{H}_\epsilon(B, L)$ is $BL/\sqrt{\epsilon}$-Lipschitz. □

We are now ready to prove the sample complexity and the correctness of the testing algorithm.

**Lemma 8.3.3** (Learning the Predictor and Testing)**.** *Algorithm 19 given* $n = \mathrm{poly}(B, L, 1/\epsilon)$ *samples and a set* $S^{\mathrm{sol}}$ *of* $\mathrm{poly}(B, L, 1/\epsilon)$ *vectors, with probability at least* 99% *returns a solution pair* $(\widehat{u}_{\widehat{\mathbf{w}}}, \widehat{\mathbf{w}})$*, with* $\widehat{u}_{\widehat{\mathbf{w}}}$ *being Lipschitz and monotone, and* $\widehat{\mathbf{w}} \in S^{\mathrm{sol}}$*, such that* $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\widehat{u}_{\widehat{\mathbf{w}}}(\widehat{\mathbf{w}}\cdot\mathbf{x}) - y)^2] \le C \min_{\mathbf{w}\in S^{\mathrm{sol}}} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2] + \epsilon$ *for some universal constant* $C$*.*

*Proof.* Let $\beta = BL/\sqrt{\epsilon}$ (if we know that the target activation $\sigma$ is $b$-Lipschitz, let $\beta = b$). Let $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$ be a set of $n$ samples and let $\widehat{u}_{\mathbf{w}}(z)$ be a solution of (Iso) (via linear interpolation), i.e.,

$$\widehat{u}_{\mathbf{w}}(z) \in \operatorname*{argmin}_{u:\beta-\mathrm{Lipshictz}, u'\ge 0} (1/n)\sum_{i=1}^n (u(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)})^2.$$

Note that in Claim 8.3.2 we showed that all the functions in $\mathcal{H}_\epsilon(B, L)$ are $\beta$-Lipschitz functions, hence we have

$$(1/n)\sum_{i=1}^n (\widehat{u}_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)})^2 \le \operatorname*{argmin}_{u\in\mathcal{H}(B,L), u'\ge 0} (1/n)\sum_{i=1}^n (u(\mathbf{w}\cdot\mathbf{x}^{(i)}) - y^{(i)})^2.$$

Let us denote $\varphi_{u,\mathbf{w}}(\mathbf{x}) \stackrel{\mathrm{def}}{=} u(\mathbf{w}\cdot\mathbf{x})$ and let $\mathcal{U} \stackrel{\mathrm{def}}{=} \{\varphi_{u,\mathbf{w}} : u \text{ is } \beta\text{-Lipschitz}, u' \ge 0, \mathbf{w} \in S^{\mathrm{sol}}\}$ be the family of all such $\varphi_{u,\mathbf{w}}$. Let $\mathcal{L}(\varphi_{u,\mathbf{w}}) \stackrel{\mathrm{def}}{=} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\varphi_{u,\mathbf{w}}(\mathbf{x}) - y)^2]$. Denote the empirical distribution on $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$ by $\widehat{\mathcal{D}}_n$, we define $\widehat{\mathcal{L}}(\varphi_{u,\mathbf{w}}) = \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}_n}[(\varphi_{u,\mathbf{w}}(\mathbf{x}) - y)^2]$. Furthermore, let

$$\widehat{\varphi}^* \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{\varphi\in\mathcal{U}} \widehat{\mathcal{L}}(\varphi), \ \mathcal{L}^* \stackrel{\mathrm{def}}{=} \min_{\varphi\in\mathcal{U}} \mathcal{L}(\varphi).$$

Then, since we know there exist an activation $\sigma \in \mathcal{H}_\epsilon(B, L)$ and a vector $\widehat{\mathbf{w}} \in S^{\text{sol}}$ such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\widehat{\mathbf{w}} \cdot \mathbf{x}) - y)^2] \leq C\text{OPT} + \epsilon$, it holds that $\mathcal{L}^* \leq C\text{OPT} + \epsilon$. Furthermore, by definition we have

$$\widehat{\varphi}^* = \operatorname*{argmin}_{\varphi\in\mathcal{U}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{D}_n}[(\varphi(\mathbf{x}) - y)^2] = \operatorname*{argmin}_{u:\beta-\text{Lipschitz},u'\geq 0,\mathbf{w}\in S^{\text{sol}}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{D}_n}[(y - u(\mathbf{w} \cdot \mathbf{x}))^2],$$

indicating that $\widehat{\varphi}^*$ is a solution of the problem (Iso) with respect to some vector $\widehat{\mathbf{w}} \in S^{\text{sol}}$, i.e., $\widehat{\varphi}^*(\mathbf{x}) = \varphi_{\widehat{u}_{\widehat{\mathbf{w}}},\widehat{\mathbf{w}}}(\mathbf{x}) = \widehat{u}_{\widehat{\mathbf{w}}}(\widehat{\mathbf{w}} \cdot \mathbf{x})$ for some $\beta$-Lipschitz function $\widehat{u}_{\widehat{\mathbf{w}}}$ and $\widehat{\mathbf{w}} \in S^{\text{sol}}$.

We use the following fact to show that $\widehat{\mathcal{L}}(\varphi_{u,\mathbf{w}})$ are close to $\mathcal{L}(\varphi_{u,\mathbf{w}})$ for all $\varphi_{u,\mathbf{w}} \in \mathcal{U}$:

**Fact 8.3.4** (Theorem 1, [SST10])**.** *Suppose there exists a constant $b > 0$ such that for any $\varphi \in \mathcal{U}$, and any $(\mathbf{x}, y) \sim \mathcal{D}$ it holds $(\varphi(\mathbf{x}) - y)^2 \leq b$. Let $\widehat{\mathcal{R}}_n(\mathcal{U})$ be the empirical Rademacher complexity of function class $\mathcal{U}$ with respect to some sample set $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$ and let $\mathcal{R}_n(\mathcal{U}) = \sup_{(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)})} \widehat{\mathcal{R}}_n(\mathcal{U})$. We have that with probability at least $1 - \delta$ over the random sample set of size $n$, for any $\varphi \in \mathcal{U}$, it holds for some universal constant $C' > 0$ that*

$$\mathcal{L}(\varphi) \leq \widehat{\mathcal{L}}(\varphi) + C'\left(\sqrt{\widehat{\mathcal{L}}(\varphi)}\left(\log^{1.5}(n)\mathcal{R}_n(\mathcal{U}) + \sqrt{\frac{b\log(1/\delta)}{n}}\right) + \log^3(n)\mathcal{R}_n^2(\mathcal{U}) + \frac{b\log(1/\delta)}{n}\right),$$

*and*

$$\mathcal{L}(\widehat{\varphi}^*) \leq \mathcal{L}^* + C'\left(\sqrt{\mathcal{L}^*}\left(\log^{1.5}(n)\mathcal{R}_n(\mathcal{U}) + \sqrt{\frac{b\log(1/\delta)}{n}}\right) + \log^3(n)\mathcal{R}_n^2(\mathcal{U}) + \frac{b\log(1/\delta)}{n}\right).$$

Note first that since $u \in \mathcal{U}$ by definition we have $|u| \leq B$ and since $|y| \leq B$ as well, it holds $|\varphi(\mathbf{x})| \leq B$ and $(\varphi(\mathbf{x}) - y)^2 \lesssim B^2$ for any $\varphi \in \mathcal{U}$.

Fact 8.3.4 implies that if $n$ is large enough such that $\mathcal{R}_n(\mathcal{U}) \leq \sqrt{\epsilon}/\log^{3/2}(n)$ and

$B^2 \log(1/\delta)/n \le \epsilon$, then we are guaranteed that: (1) for any activation $u$ and any vector $\mathbf{w} \in S^{\mathrm{sol}}$, it holds $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}\cdot\mathbf{x})-y)^2] \le \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}_n}[(u(\mathbf{w}\cdot\mathbf{x})-y)^2]$; (2) for the solution pair $(\widehat{u}_{\widehat{\mathbf{w}}}, \widehat{\mathbf{w}})$ that achieves the minimal empirical loss among all the vectors $\mathbf{w}$ from $S^{\mathrm{sol}}$ and all $\beta$-Lipschitz activations, i.e., $\widehat{\varphi}^* = \varphi_{\widehat{u}_{\widehat{\mathbf{w}}},\widehat{\mathbf{w}}} \in \operatorname{argmin}_{\varphi\in\mathcal{U}} \mathbf{E}_{(\mathbf{x},y)\sim\widehat{\mathcal{D}}_n}[(\varphi(\mathbf{x})-y)^2]$, we have $\mathcal{L}(\varphi_{\widehat{u}_{\widehat{\mathbf{w}}},\widehat{\mathbf{w}}}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_{\widehat{\mathbf{w}}}(\widehat{\mathbf{w}}\cdot\mathbf{x})-y)^2] \le (C'+1)\mathcal{L}^* \le CC'\mathrm{OPT} + \epsilon$. Therefore, by solving (Iso) and finding the besting fitting activation $u_{\mathbf{w}}$ for each $\mathbf{w} \in S^{\mathrm{sol}}$ and outputting the solution pair $(u_{\widehat{\mathbf{w}}}, \widehat{\mathbf{w}})$ with the smallest empirical error, we are ensured that $u_{\widehat{\mathbf{w}}}(\widehat{\mathbf{w}}\cdot\mathbf{x})$ is a constant factor approximate solution such that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(u_{\widehat{\mathbf{w}}}(\widehat{\mathbf{w}}\cdot\mathbf{x})-y)^2] \le C\mathrm{OPT} + \epsilon$.

Thus, it remains to bound the Rademacher complexity of the function class $\mathcal{U}$ and choose $n$ such that $\mathcal{R}_n(\mathcal{U}) \le \sqrt{\epsilon}/\log^{3/2}(n)$ and $B^2\log(1/\delta)/n \le \epsilon$. To this aim, we use the following fact:

**Fact 8.3.5** (Lemma A.3, [SST10])**.** *For any function class $\mathcal{U}$, let $N_2(\epsilon,\mathcal{U},n)$ be the $\epsilon$-cover of $\mathcal{U}$ with respect to $\ell_2$ norm on sample set $(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)})$. Let $\widehat{\mathcal{D}}_n$ be the empirical distribution on $(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)})$. Then,*

$$\widehat{\mathcal{R}}_n(\mathcal{U}) \le \inf_{\alpha>0}\left\{4\alpha + 10\int_{\alpha}^{\sup_{\varphi\in\mathcal{U}}\sqrt{\mathbf{E}_{\mathbf{x}\sim\widehat{\mathcal{D}}_n}[\varphi^2(\mathbf{x})]}} \sqrt{\frac{\log(|N_2(\epsilon,\mathcal{U},n)|)}{n}}\,\mathrm{d}\epsilon\right\}.$$

Let $\mathcal{F}$ be the family of monotone and $\beta$-Lipschitz functions that maps $[-\bar{M},\bar{M}]$ to $[-B,B]$ (recall that for all $\sigma(z) \in \mathcal{H}_\epsilon(B,L)$ we can truncate the domain of $\sigma(z)$ to $[-\bar{M},\bar{M}]$ where $\bar{M} \lesssim \sqrt{\log(B/\epsilon)}$, as shown in Fact 8.2.4, therefore, it is sufficient to consider the function class of monotone $\beta$-Lipschitz functions $u$ that maps from $[-\bar{M},\bar{M}]$ to $[-B,B]$ that contains the target activation $\sigma$). Then, standard results showed that $|N_2(\epsilon,\mathcal{F},n)| \le |N_\infty(\epsilon,\mathcal{F},n)| = (B/\epsilon)2^{\bar{M}\beta/\epsilon}$ (one can show this via constructing a grid

of width $\epsilon/\beta$ on the domain $[-\bar{M}, \bar{M}]$ and another grid of width $\epsilon$ on the codomain, see e.g., Lemma 6, [KKSK11]). Hence, since $\mathcal{U} = \mathcal{F} \circ S^{\mathrm{sol}}$, we have $|N_2(\epsilon, \mathcal{U}, n)| \leq |N_\infty(\epsilon, \mathcal{U}, n)| \lesssim (B/\epsilon) 2^{\bar{M}\beta/\epsilon} |S^{\mathrm{sol}}|$. Then, choosing $\alpha = 1/n$ in Fact 8.3.5, and noting that $\sqrt{\mathbf{E}_{\mathbf{x}\sim\widehat{\mathcal{D}}_n}[\varphi^2(\mathbf{x})]} \leq \|\varphi\|_{L_\infty} \leq B$, we obtain

$$\begin{aligned}
\widehat{\mathcal{R}}_n(\mathcal{U}) &\lesssim \frac{1}{n} + \sqrt{\frac{1}{n}} \int_{1/n}^{B} \sqrt{\log(|S^{\mathrm{sol}}|) + (\bar{M}\beta/\epsilon)\log(2) + \log(B/\epsilon)}\, \mathrm{d}\epsilon \\
&\lesssim \frac{1}{n} + B\frac{\sqrt{\log(|S^{\mathrm{sol}}|)}}{n} + \sqrt{\frac{1}{n}} \int_{1/n}^{B} \sqrt{(\bar{M}\beta/\epsilon)}\, \mathrm{d}\epsilon + \sqrt{\frac{1}{n}} \int_{1/n}^{B} \sqrt{\log(B/\epsilon)}\, \mathrm{d}\epsilon \\
&\lesssim \frac{1}{n} + B\frac{\sqrt{\log(|S^{\mathrm{sol}}|)}}{n} + \sqrt{\frac{\bar{M}\beta B}{n}} \lesssim \sqrt{\frac{\bar{M}\beta B^2 \log(|S^{\mathrm{sol}}|)}{n}}
\end{aligned}$$

Thus, $\mathcal{R}_n(\mathcal{U}) \lesssim \sqrt{\bar{M}\beta B^2 \log(|S^{\mathrm{sol}}|)/n}$. Recall that we have $|S^{\mathrm{sol}}| = \mathrm{poly}(1/\epsilon, B, L)$, $\bar{M} \leq \sqrt{\log(B/\epsilon)}$, and $\beta = BL/\sqrt{\epsilon}$ (Claim 8.3.2), therefore to guarantee that $\mathcal{R}_n(\mathcal{U}) \leq \sqrt{\epsilon}/\log^{3/2}(n)$ and $B^2 \log(1/\delta)/n \leq \epsilon$, it suffices to choose $n = \Theta(\log(BL/\epsilon)\log(1/\delta)B^3 L/\epsilon^{3/2})$. Letting $\delta = 0.01$ completes the proof. □

## 8.4 Conclusion

This work resolves a recognized open problem in the algorithmic theory of learning SIMs, by developing the first polynomial-time, constant-factor robust SIM learner for monotone activations under Gaussian inputs. At the technical level, our alignment-based spectral framework bypasses the limitations of gradient-based methods and leads to a constant-factor approximation ratio—independent of dimension, radius of optimization, or noise level. An interesting direction for future work is to generalize our algorithmic guarantees beyond Gaussian marginals, e.g., to all isotropic log-concave distributions. This question is open even for agnostic learning of a general (i.e., with arbitrary bias)

halfspace or ReLU, where all known efficient constant-factor learners critically leverage Gaussianity.

# Part III

# Robustly Learning Multi-Index Models

# Chapter 9

# Learning Multi-Index Models with Queries

## 9.1 Introduction

This chapter introduces a powerful and general framework that leverages query access to fundamentally alter the computational landscape of agnostically learning multi-index models (MIMs). We show that for a broad range of models, including deep ReLU networks and intersections of halfspaces, the exponential dependence on dimension inherent in sample-based agnostic learning can be overcome. Our method, which we term **Influence PCA**, provides the first fixed-parameter tractable algorithms for these problems.

Our framework is comprised of two main technical innovations. First, we show how to use oracle (value) queries to efficiently simulate gradient queries to a smoothed version of the gradient label function. This crucial step provides first-order information without changing the underlying Gaussian data distribution. Second, we use these gradient

estimates to construct an *influence matrix*. By performing PCA on this matrix, we can provably identify a low-dimensional subspace that captures the relevant structure of the target function, effectively reducing the problem's dimension from $d$ to one that is polynomial in the model's intrinsic complexity and $1/\epsilon$.

### 9.1.1 Our Results

#### Agnostically Learning Real-valued Multi-index Models

We define again our class of multi-index models, which we refer to as *bounded variation multi-index models*.

**Definition 9.1.1** (Bounded Variation Multi-index Models)**.** *Fix $L, M > 0$ and $k \in \mathbb{Z}_+$. We define the class $\mathfrak{R}(M, L, k)$ of continuous, (almost everywhere) differentiable real-valued functions such that for every $f \in \mathfrak{R}(M, L, k)$:*

1. *It holds $(\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f^4(\mathbf{x})])^{1/2} \leq M$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2] \leq L$.*
2. *There exists a subspace $U$ of $\mathbb{R}^d$ of dimension at most $k$ such that $f$ depends only on $U$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$ it holds that $f(\mathbf{x}) = f(\mathrm{proj}_U \mathbf{x})$, where $\mathrm{proj}_U \mathbf{x}$ is the projection of $\mathbf{x}$ on $U$.*

Our main result is an efficient algorithm that exploits the power of queries to significantly reduce the runtime of agnostically learning the semi-parametric class of Definition 9.1.1.

**Theorem 9.1.2** (Agnostic Query Learner for Real-valued Multi-index Models)**.** *Fix the function class $\mathfrak{R}(M, L, k)$ given in Definition 9.1.1. There exists an algorithm that makes $N_q = \mathrm{poly}(dML/\epsilon)$ queries, draws $N_s = \mathrm{poly}(dML/\epsilon) + k^{\mathrm{poly}(L,M,1/\epsilon)}$ random*

*labeled examples, runs in time* $\mathrm{poly}(N_s, N_q, d)$*, and outputs a polynomial* $h : \mathbb{R}^d \mapsto \mathbb{R}$ *such that with high probability* $h$ *has* $L_2^2$*-excess error* $\mathcal{E}_2(h, \mathfrak{R}(L, M, k); y) \leq \epsilon$*.*

**Concrete Applications** Theorem 9.1.2 applies to a fairly general non-parametric class of functions. Our applications are summarized to the table below:

Table 9.1: Learning Real-Valued Functions using Queries: Running time comparisons of the best known PAC algorithms with our PAC+Queries technique (Influence PCA). For a summary of our results for the above classes, we refer to Section (1.3.8).

| Function Class | PAC (without queries) $L_2$ Regression | PAC+Queries **Influence PCA** |
|---|---|---|
| Single ReLU | $d^{\mathrm{poly}(1/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(1/\epsilon)}$ |
| Sum of $k$ ReLUs | $d^{\mathrm{poly}(1/\epsilon)}$ | $\mathrm{poly}(d)\, O(k)^{\mathrm{poly}(1/\epsilon)}$ |
| Linear Combinations of $k$ ReLUs | $d^{\mathrm{poly}(k/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(k/\epsilon)}$ |
| Deep Networks with $\ell$-Layers, $S$-width | $d^{\mathrm{poly}(\ell S/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(\ell S/\epsilon)}$ |
| Bounded Variation | $d^{\mathrm{poly}(k,L,M,1/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(k,L,M,1/\epsilon)}$ |

### Agnostically Learning Boolean Multi-index Models

We start by describing the family of Boolean functions for which our results are applicable. Roughly speaking, our algorithmic approach can be used to agnostically learn any Boolean concept class $\mathcal{C}$ satisfying the following conditions: (i) $\mathcal{C}$ has bounded Gaussian surface area, (ii) it depends on an unknown low-dimensional subspace, and (iii) it is closed under translations. Under these assumptions, we similarly obtain a "fixed parameter tractable" agnostic learner qualitatively improving over the agnostic PAC setting with random examples only.

Our functon class is the following:

**Definition 9.1.3** (Bounded Surface Area, Low-Dimensional Boolean Concepts)**.** *Fix*

$\Gamma > 0$ *and* $k \in \mathbb{Z}_+$. *We define the class* $\mathfrak{B}(\Gamma, k)$ *of Boolean concepts with the following properties:*

1. *For every* $f \in \mathfrak{B}(\Gamma, k)$, *it holds* $\Gamma(f_{\mathbf{r}}) \leq \Gamma$ *for all* $\mathbf{r} \in \mathbb{R}^d$, *where* $f_{\mathbf{r}}(\mathbf{x}) = f(\mathbf{x} + \mathbf{r})$.

2. *For every* $f \in \mathfrak{B}(\Gamma, k)$, *there exists a subspace* $U$ *of* $\mathbb{R}^d$ *of dimension at most* $k$ *such that* $f$ *depends only on* $U$, *i.e., for every* $\mathbf{x} \in \mathbb{R}^d$ *it holds* $f(\mathbf{x}) = f(\mathrm{proj}_U \mathbf{x})$.

Our main positive result in this context is a query algorithm that agnostically learns the class $\mathfrak{B}(\Gamma, k)$ with running time $\mathrm{poly}(d)k^{\mathrm{poly}(\Gamma/\epsilon)}$. In more detail, we establish the following theorem:

**Theorem 9.1.4** (Agnostic Learner for Boolean Multi-index Models)**.** *Fix the concept class* $\mathfrak{B}(\Gamma, k)$ *given in Definition 9.1.3. There exists an algorithm that makes* $N_q = \mathrm{poly}(d/\epsilon)$ *queries, draws* $N_s = \mathrm{poly}(d/\epsilon) + O(k)^{\mathrm{poly}(\Gamma/\epsilon)}$ *random labeled examples, runs in sample-polynomial time, and outputs a hypothesis* $h : \mathbb{R}^d \to \{\pm 1\}$ *with excess 0-1 error* $\mathcal{E}_{0/1}(h, \mathfrak{B}(\Gamma, k); y) \leq \epsilon$.

**Concrete Applications** Theorem 9.1.4 applies to a fairly general non-parametric class of functions. In Table 9.2, we provide specific applications to well-studied classes of Boolean functions.

### 9.1.2 Detailed Technical Overview

**From Zero- to First-Order: Gradient Queries via Oracle Queries** Intuitively, having access to queries, for some example $\mathbf{x}$, we can ask for the values of $y(\mathbf{x})$ in a "small" neighborhood around $\mathbf{x}$ and therefore estimate the gradient $\nabla_{\mathbf{x}} y(\mathbf{x})$. The first issue that we have to overcome is that the observed label $y(\mathbf{x})$ is not guaranteed to

Table 9.2: Learning Boolean Concepts using Queries: Running time comparisons of the best known agnostic learners (using random samples) with our Influence PCA technique (using queries). For a summary of our results for the above classes, we refer to Section (1.3.8).

| Concept Class | PAC (without queries) $L_1$ Regression [KOS08] | PAC+Queries **Influence PCA (Ours)** |
|---|---|---|
| Single Halfspace | $d^{\mathrm{poly}(1/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(1/\epsilon)}$ |
| Intersections of $k$ Halfspaces | $d^{\mathrm{poly}(\log(k)/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(\log(k)/\epsilon)}$ |
| Functions of $k$ Halfspaces | $d^{\mathrm{poly}(k/\epsilon)}$ | $\mathrm{poly}(d)\, 2^{\mathrm{poly}(k/\epsilon)}$ |
| Degree-$\ell$, $k$-Dim. PTFs | $d^{\mathrm{poly}(\ell/\epsilon)}$ | $\mathrm{poly}(d)\, O(k)^{\mathrm{poly}(\ell/\epsilon)}$ |
| Low-Dim. Geometric Concepts | $d^{\mathrm{poly}(\Gamma/\epsilon)}$ | $\mathrm{poly}(d)\, O(k)^{\mathrm{poly}(\Gamma/\epsilon)}$ |

be a differentiable function (even if the underlying target function is). To circumvent this issue, we employ a strategy similar to the Gaussian convolution technique used in zero-order (gradient-free) optimization [NS17]. In particular, to estimate the gradient of a function $y(\cdot)$ at $\mathbf{x}$ only having access to a value oracle, the method samples $\mathbf{z}$ from a mean-zero Gaussian with small covariance, i.e., $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \rho\mathbf{I})$ for some small $\rho$, and then asks for the value of the function at $\mathbf{x} + \rho\mathbf{z}$. Even if the function $y(\cdot)$ itself is non-smooth, then, by Stein's identity, we have $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[\mathbf{z}\; y(\mathbf{x} + \rho\mathbf{z})] \propto \nabla\widetilde{y}(\mathbf{x})$, where $\widetilde{y}(\mathbf{x})$ is a smoothed version of $y(\mathbf{x})$, specifically $\widetilde{y}(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[y(\mathbf{x} + \rho\mathbf{z})]$. By drawing $N = \mathrm{poly}(d/\epsilon)$ Gaussian samples $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$, we can empirically estimate the gradient of $\widetilde{y}(\cdot)$ at every desired point $\mathbf{x} \in \mathbb{R}^d$. Therefore, by performing $N$ queries on the points $\mathbf{z}^{(i)}$, we obtain an approximation of the gradient $\nabla\widetilde{y}(\mathbf{x})$ for any $\mathbf{x}$. Even though the above technique yields gradient estimates, it comes with a cost: *to obtain the "smooth" label $\widetilde{y}(\mathbf{x})$, we add noise to the (already corrupted) label $y(\mathbf{x})$. Our plan is to argue that learning using the resulting smoothed labels $\widetilde{y}(\mathbf{x})$ yields a good classifier for the original instance — as long as the "smoothing" parameter $\rho$ is sufficiently small.*

**Ornstein–Uhlenbeck Smoothing** One could hope that if we add a small amount of noise to $y(\mathbf{x})$, the smooth label $\widetilde{y}(\mathbf{x})$ will be close to $y(\mathbf{x})$ (at least in the $L_2$-sense). Unfortunately, this is not true (even in one dimension), as $y(\mathbf{x})$ may be an arbitrarily complex function and after smoothing $\widetilde{y}(\mathbf{x})$ may be far from $y(\mathbf{x})$; see Figure 9.1. To be able to learn from the smoothed instance, we need two properties: (i) the resulting marginal distribution on the examples must be close to the initial $\mathbf{x}$-marginal, and (ii) the smoothing operation must not increase the excess error of the functions in the hypothesis class by a lot. In other words, a hypothesis that performs well with respect to the smoothed label $\widetilde{y}(\mathbf{x})$ should also perform well with respect to the original label $y(\mathbf{x})$. Applying the Gaussian convolution smoothing $\mathbf{x} + \rho\mathbf{z}$ yields a normal distribution that has covariance $(1+\rho)\mathbf{I}$. In order to make this distribution be close to a standard normal (say, in total variation distance), one would need to apply a tiny amount of noise, i.e., $\rho$ should be at most $\mathrm{poly}(1/d)$. To avoid changing the $\mathbf{x}$-marginal of the instance, instead of simply convolving with a Gaussian kernel, we apply the Ornstein–Uhlenbeck noise operator $T_\rho$ that rescales $\mathbf{x}$ and corresponds to the transformation $\widetilde{\mathbf{x}} = \sqrt{1-\rho^2}\mathbf{x} + \rho\mathbf{z}$. We observe that $\widetilde{\mathbf{x}}$ follows a standard normal distribution. The resulting "smoothed" label $\widetilde{y}$ is now defined as $T_\rho y(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[y(\widetilde{\mathbf{x}})]$. Even though the marginal of $\widetilde{\mathbf{x}}$ matches exactly with the initial marginal, we have introduced noise to the instance and we still need to show that this does not significantly affect the performance of the hypotheses in the function class of interest.

We show that, regardless of how complex the label $y(\mathbf{x})$ is, if the function class of interest is "well-behaved" — in the sense that it only contains concepts with bounded variation/Gaussian surface area — the Ornstein–Uhlenbeck noise process will not significantly affect the excess error of a hypothesis $h$.

**Proposition 9.1.5** (Informal – Ornstein–Uhlenbeck Smoothing Preserves the Risk-Minimizer)**.** *Let $y : \mathbb{R}^d \mapsto \mathbb{R}$ and $C$ be a class of functions over $\mathbb{R}^d$ such that for every $f \in C$ it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2] \leq L$. Let $\widetilde{f} \in C$ be an $L_2$ risk minimizer with respect to the smoothed label $T_\rho y$ (see Definition 9.2.1), i.e., $\widetilde{f} \in \operatorname{argmin}_{h\in C} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(h(\mathbf{x}) - T_\rho y(\mathbf{x}))^2]$. Then we have that*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}}[(\widetilde{f}(\mathbf{x}) - y(\mathbf{x}))^2] \leq \inf_{f\in C} \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - y(\mathbf{x}))^2] + O(\rho^2 L)\,.$$

At a high-level, the effect of the noise operator $T_\rho$ on the risk minimizer is milder when the function does not change very rapidly. To prove Proposition 9.1.5, we show that the correlation of any hypothesis $f$ with bounded variation is approximately preserved when we replace $y(\mathbf{x})$ with $T_\rho y(\mathbf{x})$. The correlation of $f$ with respect to $T_\rho y(\mathbf{x})$ is $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})T_\rho y(\mathbf{x})]$. However, since $T_\rho$ is a symmetric linear operator, we can equivalently apply the smoothing $T_\rho$ to $f$ and consider $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[T_\rho f(\mathbf{x})y(\mathbf{x})]$. Since $f(\mathbf{x})$ has bounded variation, we can now show via a result on noise sensitivity for real-valued functions, that $T_\rho f(\mathbf{x})$ is indeed close to $f(\mathbf{x})$ in $L_2^2$. Therefore, the correlation $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[T_\rho f(\mathbf{x})y(\mathbf{x})]$ is close to $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})]$. The fact that $T_\rho f$ and $f$ are close is intuitively clear: the smaller the variation of $f$, $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2]$, the smaller the effect of slightly perturbing a point $\mathbf{x}$ will have on the $L_2^2$, as the $L_2^2$ distance between $f(\mathbf{x})$ and $f(\sqrt{1-\rho}\mathbf{x} + \rho z)$ is roughly proportional to $\rho^2\|\nabla f(\mathbf{x})\|_2^2$. For more details, we refer to Section 9.2 and Proposition 9.2.6.

For learning Boolean concepts, we identify their Gaussian Surface Area to be the crucial complexity measure that determines the effect the smoothing operator $T_\rho$ has on the agnostic learning instance. Similarly to our result for real-valued functions, we reduce preserving the excess error to preserving the correlation of concepts, i.e., ensuring

that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})T_\rho y(\mathbf{x})] - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})]$ is small for all concepts of interest $f$ — see Proposition 9.2.10 — and then use a result of Ledoux [Led94] and Pisier [Pis86] to show that correlations are indeed approximately preserved when the concepts have bounded Gaussian Surface Area; see Proposition 9.2.10.

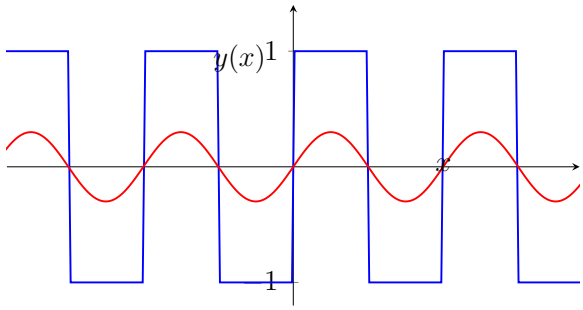



Figure 9.1: Smoothing the label $y(\mathbf{x})$. The label $y(\mathbf{x})$ corresponds to the "square wave" (shown in blue). The smoothed version $\widetilde{y}(\mathbf{x})$ is the red curve. We observe that $y(\mathbf{x})$ and $\widetilde{y}(\mathbf{x})$ are far (in the $L_2$ sense).

### 9.1.3 Learning Bounded Variation Functions via Influence PCA

**Real-Valued MIMs** Up to this point, we have established that (i) we can leverage query access in order to efficiently simulate gradient queries for the Ornstein–Uhlenbeck smoothed label $T_\rho y$, and (ii) learning from the smoothed label $T_\rho y$ is approximately equivalent to learning from the original label $y(\mathbf{x})$. We will now describe an efficient learner that uses the gradient queries to $T_\rho y$.

Our learner is based on estimating the influence matrix of $T_\rho y$ using gradient queries. This corresponds to $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\nabla T_\rho y(\mathbf{x})(\nabla T_\rho y(\mathbf{x}))^\intercal]$. Our main structural result is a general dimension-reduction tool establishing the following: given (an approximation of) the influence matrix of the smooth function $T_\rho y$, we can perform PCA and learn a

low-dimensional subspace $V$ so that a bounded variation function that depends only on $V$ can achieve $\epsilon$ excess error with respect to $T_\rho y$ in $L_2^2$. This dimension-reduction step crucially relies on the target concept being low-dimensional (see Definition 9.1.1).

In fact, our dimension-reduction proof for real-valued concepts shows directly that a low-degree polynomial that depends only on the low-dimensional space $V$ exists.

**Proposition 9.1.6** (Informal– Dimension Reduction via Influence PCA: Real-Valued Functions)**.** *Let* $\widetilde{y}(\mathbf{x}) = T_\rho y(\mathbf{x})$ *and let* $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\nabla\widetilde{y}(\mathbf{x})(\nabla\widetilde{y}(\mathbf{x}))^\intercal]$*. Moreover, let* $V$ *be the subspace spanned by all the eigenvectors of* $\mathbf{M}$ *whose corresponding eigenvalues are at least* $\epsilon^2/(kM)$*. The following holds:*

- *The dimension of* $V$ *is at most* $\mathrm{poly}(M, k, 1/\rho, 1/\epsilon)$*.*
- *There exists a polynomial* $q : V \mapsto \mathbb{R}$ *of degree* $m = O(L/\epsilon^2)$ *such that*

$$\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(q(\mathrm{proj}_V(\mathbf{x})) - \widetilde{y}(\mathbf{x}))^2] \leq \inf_{f\in\mathfrak{R}(M,L,k)} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \widetilde{y}(\mathbf{x}))^2] + \epsilon\,.$$

To prove Proposition 9.1.6, we explicitly construct a low-dimensional polynomial as follows: we first marginalize out the low-influence directions of $\widetilde{y}(\cdot)$, and then we keep its low-degree Hermite approximation.

**Marginalizing Low-Influence Directions** We first construct a low-dimensional (not necessarily polynomial) version of the noisy label $\widetilde{y}$ that preserves the correlation with the target function $f(\cdot)$. By the assumption of Proposition 9.1.8, all directions in the orthogonal complement $V^\perp$ are low-influence, i.e., for $\mathbf{h} \in V^\perp$ it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{h}\cdot\nabla\widetilde{y}(\mathbf{x}))^2] \leq O(\epsilon^2/k)$. In words, the function $\widetilde{y}$ is "approximately constant" along some low-influence direction $\mathbf{h}$. Let us first assume that $\widetilde{y}$ is exactly constant on all

directions of $V^\perp$. Then, in order to preserve the correlation of $\widetilde{y}$ with $f$, *we only need to match the expected value of* $\widetilde{y}$ *over* $V^\perp$. This motivates the following "Gaussian Marginalization Operator" $(\Pi_V g)(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[g(\text{proj}_V \mathbf{x} + \text{proj}_{V^\perp}\mathbf{z})]$ (see Definition 9.3.5 and Lemma 9.3.6). So a natural low-dimensional "approximation" of $\widetilde{y}$ is $\Pi_V \widetilde{y}$. Indeed, if $\widetilde{y}$ was constant on $V^\perp$, using the fact that $\text{proj}_V \mathbf{x}$ and $\text{proj}_{V^\perp}\mathbf{x}$ are independent standard Gaussians, we would obtain that

$$\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}[\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}} \widetilde{y}(\text{proj}_V(\mathbf{x}) + \text{proj}_{V^\perp}(\mathbf{z}))f(\mathbf{x})]] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\widetilde{y}(\mathbf{x})f(\mathbf{x})] = 0 \,.$$

Our goal is to show that the Gaussian marginalization $\Pi_V \widetilde{y}$ achieves similar correlation with $\widetilde{y}$ as $f$, when $\widetilde{y}$ is not constant in $V^\perp$ but "approximately constant", i.e., it has low-influence in directions of $V^\perp$. In Lemma 9.3.12 we show that when $V^\perp$ contains only low-influence directions, the same is approximately true (up to some additive $\epsilon$ error): $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}}[(\widetilde{y}(\mathbf{x}) - \Pi_V \widetilde{y}(\mathbf{x}))f(\mathbf{x})] \leq O(\epsilon)$. To do this, we first observe that, since $f$ depends only on the subspace $U$, it holds that $\Pi_U f = f$ and $\Pi_V f$ depends only on the directions inside the relevant subspace $W = U + V$. We can thus restrict our attention on $W$, i.e., bound the difference $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}_W}[(\widetilde{y}(\mathbf{z}) - \Pi_V \widetilde{y}(\mathbf{z}))f(\mathbf{z})]$, where $\mathcal{N}_W$ is a standard normal on the subspace $W$. We will show that this correlation difference can be bounded by the variance of $\widetilde{y}$ in the irrelevant directions. Indeed, by the Cauchy-Schwarz inequality, we have

$$\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_W}[(\widetilde{y}(\mathbf{z}) - \Pi_V \widetilde{y}(\mathbf{z}))f(\mathbf{z})] \leq \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_W}[f^2(\mathbf{x})]\right)^{1/2} \left(\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_W}[(\widetilde{y}(\mathbf{z}) - \Pi_V \widetilde{y}(\mathbf{z}))^2]\right)^{1/2} .$$

We next relate the $L_2^2$ error introduced by the marginalization operation $\Pi_V$ on $\widetilde{y}$ with the influence matrix $\mathbf{M}$. We use the Gaussian Poincare inequality, which states that

for some $g(t) : \mathbb{R} \mapsto \mathbb{R}$ it holds $\mathbf{Var}_{t\sim\mathcal{N}}[g(t)] \leq \mathbf{E}_{t\sim\mathcal{N}}[(g'(t))^2]$. We obtain that for any subspace $R = \mathbf{r}^\perp$ (the orthogonal complement to the direction $\mathbf{r}$) the variance $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}_W}[(\widetilde{y}(\mathbf{z}) - \Pi_R\widetilde{y}(\mathbf{z}))^2]$ is bounded above by $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_W}[(\nabla\widetilde{y}(\mathbf{x})\cdot\mathbf{r})^2] = \mathbf{r}^\intercal\mathbf{M}\mathbf{r}$. By repeatedly applying the Gaussian Poincare inequality on a basis of the (at most) $k$-dimensional subspace $V^\perp \cap W$, we show that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_W}[(\widetilde{y}(\mathbf{z}) - \Pi_V\widetilde{y}(\mathbf{z}))^2] \leq Mk \max_{\mathbf{r}\in V^\perp, \|\mathbf{r}\|_2=1} \mathbf{r}^\intercal\mathbf{M}\mathbf{r} \leq k\; O(\epsilon^2/(kM) = O(\epsilon^2)\,.$$

In the above bound, we observe that accepting eigenvectors with corresponding eigenvalues at least $\epsilon^2/(Mk)$ ensures that $\Pi_V\widetilde{y}$ achieves at most $O(\epsilon)$ worse correlation with $f$ than $\widetilde{y}$.

**The Low-Degree Polynomial Approximation** We have established that $\Pi_V\widetilde{y}$ is similar to $\widetilde{y}$ in the sense that it has similar (up to $\epsilon^2$) correlation with the target function $f(\cdot)$. To obtain a polynomial with a similar behavior, we use the low-degree Hermite expansion of $\Pi_V\widetilde{y}$, which we denote by $P_m\Pi_V\widetilde{y}$, where $P_m g$ maps the function $g$ to its degree Hermite expansion. We show that in order for $P_m\Pi_V\widetilde{y}$ to achieve low $L_2^2$ excess error, it suffices to pick the degree $m$ so that $P_m f(\mathbf{x})$ is close to $f(\mathbf{x})$ (in $L_2^2$). We show that the following bound for the excess error defined as $\mathcal{E}_2(q, f; \widetilde{y}) = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\widetilde{y}(\mathbf{x}) - q(\mathbf{x}))^2] - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\widetilde{y}(\mathbf{x}) - f(\mathbf{x}))^2]$. We refer to Lemma 9.3.11 for the formal statement and proof.

**Lemma 9.1.7** (Informal – Excess $L_2^2$ Error Decomposition)**.** *It holds*

$$\mathcal{E}_2(P_m\Pi_V\widetilde{y}, f; \psi) \leq O(1)\Big(\underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - P_m f(\mathbf{x}))^2]}_{\textit{Polynomial Approximation Error}} + \underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\widetilde{y}(\mathbf{x}) - \Pi_V\widetilde{y}(\mathbf{x}))f(\mathbf{x})]}_{\textit{Correlation Error}}\Big)\,.$$

Since $f(\mathbf{x})$ has bounded variation (see Definition 9.1.1), we can show using a result from [KTZ19] (see Lemma 9.3.4) that with degree $m = O(L/\epsilon^2)$, it holds that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - P_m f(\mathbf{x}))^2] = \epsilon$. Moreover, in the previous paragraph, we have already established that the correlation error is also $O(\epsilon)$.

**Polynomial Regression in $V$** So far, we have identified the subspace $V$ and we know that there exists a polynomial that depends on $V$ and achieves low $L_2^2$ error with the smoothed label $\widetilde{y} = T_\rho y$. Since we have established that the smoothing operation $T_\rho$ does not affect the excess error of a bounded-surface area concept by a lot (see Proposition 9.1.5), we know that the same concept will achieve low excess-error with respect to the original label $y$. Having established this, for our final step we may directly perform polynomial regression in the low-dimensional subspace $V$ to learn a polynomial with low-excess error. Since the dimension of $V$ is roughly $\mathrm{poly}(Mk/\epsilon)$ and the degree of the polynomial is $\mathrm{poly}(L/\epsilon)$, the total sample and computational complexity of this task is roughly $k^{\mathrm{poly}(L/\epsilon)}$.

**Boolean MIMs** At a high level, the proof and algorithm for Boolean MIMs is similar to that for real-valued MIMs. We show the following dimension reduction lemma that essentially reduces the initial problem to learning a bounded surface area concept in a $\mathrm{poly}(k/\epsilon)$-dimensional subspace $V$.

**Proposition 9.1.8** (Informal – Dimension-Reduction via Influence PCA: Boolean Concepts)**.** *Let $V$ be the subspace spanned by all the eigenvectors of $\mathbf{M}$, where $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\nabla T_\rho y(\mathbf{x})(\nabla T_\rho y(\mathbf{x}))^\intercal]$, whose corresponding eigenvalues are at least $\Omega(\epsilon^2/k)$. The following holds:*

- *The dimension of $V$ is at most* $\mathrm{poly}(k/(\epsilon\rho))$.

- *There exists* $g : \mathbb{R}^d \to \{\pm 1\}$ *with* $\Gamma(g) \leq \Gamma$ *and* $g(\mathbf{x}) = g(\mathrm{proj}_V \mathbf{x})$ *for all* $\mathbf{x} \in \mathbb{R}^d$ *such that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|g(\mathbf{x}) - T_\rho y(\mathbf{x})|] \leq \inf_{f \in \mathfrak{B}(\Gamma,k)} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x}) - T_\rho y(\mathbf{x})|] + \epsilon\,.$$

So far, we have identified the subspace $V$ and we know that there exists a bounded surface area Boolean concept that depends on $V$ and achieves low $L_1$ error with the smoothed label $T_\rho y$. Since we have established that the smoothing operation $T_\rho$ does not affect the excess error of a bounded-surface area concept by a lot (see Proposition 9.1.5 and Lemma 9.2.11), we know that the same concept will achieve low excess-error with respect to the original label $y$. Having established this, for our final step we may use the $L_1$-agnostic learner of [KOS08] on the $k$-dimensional subspace $V$ to learn a PTF of degree $\mathrm{poly}(\Gamma/\epsilon)$ with $(\dim(V))^{\mathrm{poly}(\Gamma/\epsilon)} = k^{\mathrm{poly}(\Gamma/\epsilon)}$ samples and time.

### 9.1.4 Hardness of Proper Agnostic Query Learning for ReLUs and Halfspaces

Here we sketch our hardness reduction, establishing that the exponential dependence in $1/\epsilon$ is inherent for proper agnostic learners, even with query access to the function (see Theorem 9.5.3 and Theorem 9.5.4). In particular, we show that assuming there are no polynomial-time algorithms for the Small-Set Expansion (SSE) problem [RS10], then there are no polynomial time *proper* agnostic learning algorithms for ReLUs and homogeneous halfspaces with respect to the Gaussian distribution.

The basic idea of our argument is to reduce to the problem of (approximately) optimizing a homogeneous degree-4 polynomial over the unit sphere (for the case of halfspaces we reduce to optimizing a degree-5 polynomial). As there are already

known reductions from SSE to the problem of finding approximate maxima of degree-4 polynomials (and for halfspaces we can do a simple reduction from degree-4 to degree-5) this will suffice.

For this, we note that if $f(\mathbf{x})$ is a polynomial and $g(\mathbf{x}) = \mathrm{ReLU}(\mathbf{v} \cdot \mathbf{x})$ for $\mathbf{v}$ a unit vector, then $\mathbf{E}[f(\mathbf{x})g(\mathbf{x})]$ is a low-degree polynomial in $\mathbf{v}$. In fact, by specifying $f$, we can make this into any homogeneous degree-5 polynomial we desire. This gives us SSE hardness of approximating $\mathbf{E}[f(\mathbf{x})g(\mathbf{x})]$.

If $f$ were a Boolean function we would be done. However, as this is not the case, we need two additional steps. Firstly, we scale down $f$ and truncate it so that its values stay within $[-1, 1]$ (note that this introduces only a small error if the average size of $f$ is small). Second, we replace $f$ by a random Boolean function $\tilde{f}$ so that $\mathbf{E}[\tilde{f}(\mathbf{x})] = f(\mathbf{x})$. Doing this, it is not hard to see that with high probability over the randomness of defining $\tilde{f}$ that $\mathbf{E}[\tilde{f}(\mathbf{x})g(\mathbf{x})]$ is arbitrarily close to $\mathbf{E}[f(\mathbf{x})g(\mathbf{x})]$ for all functions $g$.

Now even if the algorithm was given an explicit description of our function $\tilde{f}$, finding a ReLU function $g$ that approximately maximizes $\mathbf{E}[\tilde{f}(\mathbf{x})g(\mathbf{x})]$ is essentially equivalent to approximately optimizing a homogeneous degree-5 polynomial of the sphere, which is SSE-hard.

### 9.1.5 Related Work

Here we discuss prior and related work that was not already discussed in the introduction.

**Comparison to Prior Work** Our algorithmic template involves two steps to agnostically learn multi-index models under the Gaussian distribution. First, we use queries to "smooth" the label function without adding a lot of noise to the instance.

We then use PCA on the expected gradient outer-product of the "smoothed" concept $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\nabla f(\mathbf{x})\nabla f(\mathbf{x})^T]$ to find a low-dimensional space containing an (nearly) optimal hypothesis.

Using PCA on the expected gradient outer-product is a well-known dimension reduction technique that has been applied in many supervised learning settings, see, e.g., [XTLZ02, MW06, MZST06, WGMM10]. We emphasize that prior results of this type focus on (i) the noiseless (realizable) setting, and (ii) the case of differentiable target functions. In comparison, we perform agnostic learning with non-differentiable functions by crucially exploiting query access. Using sample access only, estimating the gradient of $f(\mathbf{x})$ requires exponentially many examples in the dimension, see, e.g., [MZST06].

[GKK08a] developed an efficient agnostic query learner for decision trees under the uniform distribution on the Boolean hypercube. The approach of [GKK08a] crucially relies on the fact that the target hypothesis can be represented as a sparse polynomial. The class of functions we consider (Definition 9.1.3) — and in particular even a single halfspace or ReLU — does not have this property, and therefore methods relying on sparsity [KM93, GKK08a] are not applicable.

In the context of property testing, [DMN21] used a similar approach based on PCA on the expected outer gradient product to test whether the observed label is close to a smooth low-dimensional junta (similarly to Definition 9.1.3). An important difference with the current work is that in many interesting applications the link function may assumed to be known, e.g., agnostically learning a ReLU or a halfspace, and the goal is to *learn* a good hypothesis — a task that information-theoretically requires $\Omega(d)$ samples. In contrast, [DMN21] focuses on the semi-parametric task of only testing the unknown link function (and not identifying the underlying low-dimensional subspace)

while avoiding a poly($d$) dependence in the sample complexity.

Finally, related to our setting is the more recent work of [DKK+21a], where a combination of polynomial regression and PCA on the average outer product of the gradient was employed for proper, agnostic learning of a single halfspace with runtime and sample complexity $d^{\mathrm{poly}(1/\epsilon)}$. In this work, we crucially exploit the query access to bypass the polynomial regression step and significantly improve the runtime to $\mathrm{poly}(d)2^{\mathrm{poly}(1/\epsilon)}$ (for the special case of a single halfspace).

**Agnostically Learning Boolean Functions with Queries** It is known (see, e.g., [Fel08]) that the availability of queries does *not* help computationally in the distribution-free agnostic setting. Specifically, Feldman [Fel08] showed that every concept class that is agnostically learnable with queries is also agnostically learnable from random samples only (while preserving computational efficiency within a polynomial factor). This simple yet powerful fact has motivated the study of agnostic query learning *with respect to specific natural distributions*, such as the uniform distribution on the hypercube or the Gaussian distribution.

In the context of learning Boolean functions, the study of distribution-specific agnostic learning with queries has a rich history. One of the earliest results in this vein is the classical algorithm of Goldreich and Levin [GL89] that uses queries to efficiently agnostically learn parity functions under the uniform distribution. (Recall that the problem of learning parities with noise is conjectured to be computationally hard with random samples only.) Kushilevitz and Mansour [KM93], building on the ideas of [GL89], developed an efficient (non-agnostic) query learner for decision trees under the uniform distribution. As already mentioned, [GKK08a] subsequently gave a polynomial-time agnostic query learner for decision trees under the uniform distribution.

### 9.1.6 Roadmap

In Section (9.2.1), we show that we can use queries to simulate gradient access to the Ornstein–Uhlenbeck smoothing $T_\rho y$. In Sections (9.2.2) and (9.2.3), we show that the noise operator we use does not affect the agnostic learning task for real-valued functions and Boolean concepts. In Section 9.3, we show our result for learning real-valued functions and prove Theorem 9.1.2. In Section (9.3.3), we show how Theorem 9.1.2 implies agnostic learning for linear combinations of ReLU activations and deep networks. In Section 9.4, we give our agnostic learner for Boolean concepts with bounded surface area and establish Theorem 9.1.4 and the associated applications. In Section 9.5, we show that under the SSE hypothesis, no polynomial-time proper query learner for agnostically learning ReLUs or LTFs exists. In Section (9.4.4) and Section (9.3.4), we give our result for proper agnostic learning of LTFs and ReLUs.

## 9.2 From Zero- to First-Order: Derivative Queries via Oracle Queries

In this section, we show that we can efficiently simulate gradient access to a smoothed version of the label $y$ using queries. In Section (9.2.1) we show how to use the Ornstein–Uhlenbeck operator to get acecss to gradient queries of $y$. In Section (9.2.3) and Section (9.2.2) we show that the noise that we introduce in order to simulate the gradient queries does not affect the agnostic learning task for Boolean and real valued concepts as long as the Gaussian surface area (for Boolean concepts) and the expected gradient norm (for real-valued functions) are bounded.

### 9.2.1 Gradient Queries via Oracle Queries

We first formally define the Ornstein–Uhlenbeck smoothing operator.

**Definition 9.2.1** (Ornstein–Uhlenbeck Operator)**.** *Let $\rho \in (0,1)$. We denote as $T_\rho$ the linear operator that maps a function $g \in L^2(\mathcal{N})$ to the function $T_\rho g$ defined as:*

$$(T_\rho g)(\mathbf{x}) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}\left[g(\sqrt{1-\rho^2}\mathbf{x}+\rho\mathbf{z})\right] .$$

*To simplify notation, we often write $T_\rho g(\mathbf{x})$ instead of $(T_\rho g)(\mathbf{x})$.*

The Ornstein–Uhlenbeck operator is well studied (see, e.g., [Bog98, KOS08] and references therein) and has several structural properties that enable the analysis of our algorithm. Its crucial property is that regardless of how complex the initial function $g$ is, $T_\rho g$ is always everywhere differentiable and also the norm of the gradient of $T_\rho g$ only depends on the maximum value of the function $g$. In the next fact we collect the properties that we use.

**Fact 9.2.2** (see, e.g., [Bog98])**.** *Let $g : \mathbb{R}^d \mapsto \mathbb{R}$. For the function $T_\rho g(\mathbf{x})$ the following properties hold*

1. *$T_\rho g(\mathbf{x})$ is differentiable at every point $\mathbf{x}$.*
2. *$T_\rho g(\mathbf{x})$ is $1/\rho$-Lipschitz, i.e., $\|\nabla T_\rho g(\mathbf{x})\|_2 \leq \|g\|_\infty/\rho$.*
3. *For any $p \geq 1$, $T_\rho$ is a contraction with respect the $\|\cdot\|_p$, i.e., it holds $\|T_\rho g\|_{L^p} \leq \|g\|_{L^p}$.*

Using it allows the gradient of the smoothed function $T_\rho g(\mathbf{x})$ to be computed directly given value access to the underlying function $g$. We now present the main result of

this section showing that given query access to the label $y(\cdot)$ we can efficiently simulate gradient queries to the smoothed label $T_\rho y(\cdot)$ with roughly $\widetilde{O}(d/\epsilon)$ queries.

**Lemma 9.2.3** (Gradient Queries from Oracle Queries)**.** *Fix $\epsilon, \delta, \rho > 0$. Let $y(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ be a function in $L_2^2(\mathcal{N})$ with $|y(\mathbf{x})| \leq M$. There exists an algorithm (see Algorithm 20) that given a point $\mathbf{x} \in \mathbb{R}^d$ makes $N = \widetilde{\Omega}(dM/\epsilon)\log(1/\delta)$ queries to $y(\mathbf{x})$ and, in polynomial time, returns a vector $\widetilde{\xi}$ such that, with probability at least $1-\delta$, it holds $\|\widetilde{\xi} - \nabla T_\rho y(\mathbf{x})\|_2 \leq \epsilon$.*

*Proof.* To show the lemma, we first need to show that for any point $\mathbf{x} \in \mathbb{R}^d$, we can use enough queries to estimate $D_\rho y(\mathbf{x})$ accurately, meaning that we need to estimate the random variable $\mathbf{Z} = \frac{\sqrt{1-\rho^2}}{\rho}\mathbf{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[y(\sqrt{1-\rho^2}\mathbf{x} + \rho\mathbf{z})\mathbf{z}\right]$ accurately. Note that by definition the random variable $\mathbf{Z}$ is $1/\rho^2$ sub-gaussian, therefore from a simple application of the Hoefding inequality, we get that with $O(dM/(\rho\epsilon)^2\log(1/\delta_1))$ queries, we can find a $\widetilde{\mathbf{Z}}$ such that $\|\widetilde{\mathbf{Z}} - \mathbf{E}[\mathbf{Z}]\|_2 \leq \epsilon$ with probability at least $1-\delta_1$. □

**Lemma 9.2.4** (Gradient of Smoothed Label)**.** *Let $\rho \in (0,1)$. We denote as $D_\rho$ the linear operator that maps a function $g \in L^2(\mathcal{N})$ to the function $D_\rho g$ defined as: $(D_\rho g)(\mathbf{x}) \stackrel{\text{def}}{=} \nabla(T_\rho g)(\mathbf{x})$. It holds that*

$$(D_\rho g)(\mathbf{x}) = \frac{\sqrt{1-\rho^2}}{\rho}\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}\left[g(\sqrt{1-\rho^2}\mathbf{x} + \rho\mathbf{z})\mathbf{z}\right] .$$

*To simplify notation, we often write $D_\rho g(\mathbf{x})$ instead of $(D_\rho g)(\mathbf{x})$.*

*Proof.* We first observe that for any fixed $\mathbf{x}$ the random variable $\sqrt{1-\rho^2}\mathbf{x} + \rho\mathbf{z}$ is distributed according to $\mathcal{N}(\sqrt{1-\rho^2}\mathbf{x}, \rho^2\mathbf{I})$. Therefore, we have

$$T_\rho g(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}[g(\sqrt{1-\rho^2}\mathbf{x} + \rho\mathbf{z})] = \mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{N}(\sqrt{1-\rho^2}\mathbf{x},\rho^2\mathbf{I})}[g(\mathbf{u})]$$

We can now directly compute the gradient of the smoothed function $T_\rho g$:

$$\nabla_{\mathbf{x}}(T_\rho g)(\mathbf{x}) = \nabla_{\mathbf{x}} \mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{N}(\sqrt{1-\rho^2}\mathbf{x},\rho^2\mathbf{I})}[g(\mathbf{u})] = \frac{\sqrt{1-\rho^2}}{\rho^2} \mathop{\mathbf{E}}_{\mathbf{u}\sim\mathcal{N}(\sqrt{1-\rho^2}\mathbf{x},\rho^2\mathbf{I})}\left[g(\mathbf{u})(\mathbf{u}-\sqrt{1-\rho^2}\mathbf{x})\right]$$
$$= \frac{\sqrt{1-\rho^2}}{\rho} \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}\left[g(\sqrt{1-\rho^2}\mathbf{x}+\rho\mathbf{z})\mathbf{z}\right] .$$

□

**Algorithm 20** Simulating Gradient Queries with Queries

**Input:** $\epsilon > 0$, $\delta > 0$, $\rho > 0$, location $\mathbf{x} \in \mathbb{R}^d$.
**Requires:** Sample & query access to labeled-example distribution $\mathcal{D}$.
**Output:** Estimate $\widetilde{\xi} = \nabla T_\rho y(\mathbf{x})$ with $\|\widetilde{\xi} - \nabla T_\rho y(\mathbf{x})\|_2 \le \epsilon$.
Sample $N \leftarrow \widetilde{O}\left(\frac{d}{\epsilon}\log\frac{1}{\delta}\right)$ points $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
**for** $j = 1$ **to** $N$ **do**
    Query at $\mathbf{q}^{(j)} \leftarrow \sqrt{1-\rho^2}\,\mathbf{x} + \rho\,\mathbf{z}^{(j)}$, receive label $y^{(j)}$.
**end for**
**Return:** $\widetilde{\xi} \leftarrow \frac{\sqrt{1-\rho^2}}{N\rho}\sum_{j=1}^{N} y^{(j)}\,\mathbf{z}^{(j)}$ .

### 9.2.2 Smoothing the Labels for Learning Real-valued Functions

In this section we show that adding noise to the label $y(\mathbf{x})$ in order to make it smooth and compute its gradients does not "change" the agnostic learning task significantly. Assume that there exists a learning algorithm that can learn a hypothesis $h(\cdot)$ that achieves $\epsilon$-excess error compared to a class of concepts $C$, given access to the smooth labels $T_\rho y(\mathbf{x})$. In other words, assume that we are given a learner that finds a hypothesis $h(\cdot)$ that satisfies

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(h(\mathbf{x}) - T_\rho y(\mathbf{x}))^2] \le \inf_{f\in C} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - T_\rho y(\mathbf{x}))^2] + \epsilon\,.$$

Then, can we say that $h(\cdot)$ will perform well compared to the same class $C$ under the original (non-smooth) label $y(\cdot)$? We show that this is true when (i) the hypothesis $h(\cdot)$ produced by the learner is not very complicated in the sense that it has bounded variation and (ii) the hypothesis class $C$ that we are comparing $h(\cdot)$ against has also bounded variation.

In particular, we show that a hypothesis $h(\cdot)$ achieves $\epsilon$-excess error compared to some concept class $C$ in the *smoothed* instance, achieves $(\epsilon + O(\sqrt{\rho})$-excess error with respect to the original instance. In other words, as long as the variation and $L_2^2$ norms of the target concept class and the hypothesis produced by the learner are bounded, smoothing the noisy label $y(\mathbf{x})$ does not introduce significantly more noise to the instance. To simplify notation, we first define the excess error, i.e., the error of a classifier minus the error of the best-in-class classifier of some class $C$.

**Definition 9.2.5** (Excess Error)**.** *Given hypotheses $h, f : \mathbb{R}^d \mapsto \mathbb{R}$ we define the $L_1$-excess error of $h(\cdot)$ compared to $f(\cdot)$ with respect to the label $y(\cdot)$ to be $\mathcal{E}_1(h, f; y) = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|h(\mathbf{x}) - y(\mathbf{x})|] - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x}) - y(\mathbf{x})|]$. Moreover, for a class of concepts $C$ we define the excess error of $h(\cdot)$ compared to $C$ with respect to $y(\cdot)$ as $\sup_{f\in C} \mathcal{E}_1(h, f; y)$. Similarly, we define the $L_2^2$-excess error as $\mathcal{E}_2(h, f; y) = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(h(\mathbf{x}) - y(\mathbf{x}))^2] - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - y(\mathbf{x}))^2]$ and $\mathcal{E}_2(h, C; y) = \sup_{f\in C} \mathcal{E}_2(h, f; y)$.*

We now show that that the Ornstein–Uhlenbeck noise operator also preserves the $L_2^2$-excess error of a classifier $h : \mathbb{R}^d \mapsto \mathbb{R}$ as long as the target class and the classifier $h$ have bounded expected gradient.

**Proposition 9.2.6** (Smoothing the Noisy Labels)**.** *Fix $f \in \mathcal{R}(M, L, k)$. Let $y : \mathbb{R}^d \mapsto \mathbb{R}$ be a function in $L^2(\mathcal{N})$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[y^2(\mathbf{x})] \leq M$. Moreover, let $p(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ be an*

*almost everywhere differential function in* $L_2(\mathcal{N})$ *with* $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla p(\mathbf{x})\|_2^2] \leq L$. *It holds that*

$$\mathcal{E}_2(p, C; y) \leq \mathcal{E}_2(p, C; T_\rho y) + O(\sqrt{\rho M L}) \,.$$

*Proof of Proposition 9.2.6.* We first prove the following lemma that connects the excess error of a real-valued function $h(\cdot)$ with respect to the smoothed label $T_\rho y(\cdot)$ to its excess error with respect to the original label $y(\cdot)$. If the operator $T_\rho$ preserves the correlation of all concepts $f \in C$, i.e., $|\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})] - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})T_\rho y(\mathbf{x})]| \leq \epsilon$ for all $f \in C$ and it also preserves the correlation of the hypothesis $h(\cdot)$, i.e., $|\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[h(\mathbf{x})y(\mathbf{x})] - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[h(\mathbf{x})T_\rho y(\mathbf{x})]| \leq \epsilon$, then the excess error of $h(\cdot)$ with respect to $y(\cdot)$ is at most $2\epsilon$ worse than its excess error with respect to the smoothed label $T_\rho y(\cdot)$. In the following lemma, we show that we can connect the $L_2$-excess error with the correlation of concepts.

**Lemma 9.2.7** (From Excess Error to Correlation Preservation)**.** *Let* $h : \mathbb{R}^d \mapsto \mathbb{R}$ *be a real-valued hypotheses and* $C$ *be a class of real-valued hypotheses. It holds*

$$\mathcal{E}_2(h, C; T_\rho y) - \mathcal{E}_2(h, C; y) \leq 2 \sup_{f\in C} \left| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})T_\rho y(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})] \right| + 2 \left| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[h(\mathbf{x})T_\rho y(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[h(\mathbf{x})y(\mathbf{x})] \right| .$$

*Proof.* We first note that $\mathcal{E}_2(h, C; T_\rho y) - \mathcal{E}_2(h, C; y) = \sup_{f\in C} \mathcal{E}_2(h, f; T_\rho y) - \sup_{f\in C} \mathcal{E}_2(h, f; y) \leq \sup_{f\in C} \left|\mathcal{E}_2(h, f; T_\rho y) - \mathcal{E}_2(h, f; y)\right|$. For some fixed concept $f \in C$, we have

$$\mathcal{E}_2(h, f; T_\rho y) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[h^2(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f^2(\mathbf{x})] + 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - h(\mathbf{x}))(T_\rho y)] \,.$$

Therefore, we have

$$\begin{aligned}\mathcal{E}_2(h, f; T_\rho y) - \mathcal{E}_2(h, f; y) \\ = 2\left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})(T_\rho y(\mathbf{x}) - y(\mathbf{x}))] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[h(\mathbf{x})(T_\rho y(\mathbf{x}) - y(\mathbf{x}))]\right).\end{aligned}$$

By taking the supremum over the $f$, we complete the proof. □

Note that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})(T_\rho y(\mathbf{x}) - y(\mathbf{x}))] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[y(\mathbf{x})(T_\rho f(\mathbf{x}) - f(\mathbf{x}))]$. Therefore, using Cauchy-Schwarz inequality we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[y(\mathbf{x})(T_\rho f(\mathbf{x}) - f(\mathbf{x}))] &\le \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[y^2(\mathbf{x})] \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(T_\rho f(\mathbf{x}) - f(\mathbf{x}))^2]\right)^{1/2} \\ &\le \sqrt{M}\left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(T_\rho f(\mathbf{x}) - f(\mathbf{x}))^2]\right)^{1/2},\end{aligned}$$

where we used that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[y^2(\mathbf{x})] \le \sqrt{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[y^4(\mathbf{x})]} \le M$. To bound the remaining term, we prove the following claim.

**Claim 9.2.8.** *Let $f \in L^2(\mathcal{N})$ be a continuous and (almost everywhere) differentiable function. Then, $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(T_\rho f(\mathbf{x}) - f(\mathbf{x}))^2] \le 2\rho^2 \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2]$.*

*Proof.* We will use the following result from [KTZ19].

**Fact 9.2.9** (Correlated Differences, (Lemma 7 in [KTZ19]))**.** *Let $f \in L^2(\mathcal{N})$ be an (almost everywhere) differentiable function. Denote by*

$$D_\tau = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{I} & (1-\tau)\mathbf{I} \\ (1-\tau)\mathbf{I} & \mathbf{I} \end{pmatrix}\right).$$

*It holds $\mathbf{E}_{(\mathbf{x},\mathbf{z})\sim D_\tau}[(f(\mathbf{x}) - f(\mathbf{z}))^2] \le 2\tau \ \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2]$.*

Therefore, using Jensen's inequality, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(T_\rho f(\mathbf{x}) - f(\mathbf{x}))^2] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}[f(\sqrt{1-\rho^2}\mathbf{x} + \rho\mathbf{z})] - f(\mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{z}')\sim D_\tau}[(f(\mathbf{z}') - f(\mathbf{x}))^2]\,,$$

for $\tau = 1 - \sqrt{1-\rho^2}$. Therefore, using Fact 9.2.9, we obtain

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(T_\rho f(\mathbf{x}) - f(\mathbf{x}))^2] \leq 2(1-\sqrt{1-\rho^2}) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2 \leq 2\rho^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2\,,$$

where we used the fact that $\sqrt{1-\rho^2} \geq 1-\rho^2$ which holds for all $\rho \in [0,1]$ and implies that $1 - \sqrt{1-\rho^2} \leq \rho^2$. □

Therefore, from Claim 9.2.8, we have that

$$\mathcal{E}_2(p, C; y) \leq \mathcal{E}_2(p, C; T_\rho y) + O(\sqrt{\rho M})\left(\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2]} + \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla p(\mathbf{x})\|_2^2]}\right).$$

Using that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2], \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla p(\mathbf{x})\|_2^2] \leq L$, we complete the proof of Proposition 9.2.6. □

### 9.2.3 Smoothing Labels for Learning Boolean Concepts

The following proposition shows that the $L_1$-excess error of a hypothesis $h$ with respect to the original label $y$ is close to its $L_1$-excess error with respect to the smoothed label $T_\rho y$ as long as (i) the class $C$ contains concepts with bounded surface area and (ii) the classifier $h$ also has bounded surface area.

**Proposition 9.2.10** (Smoothing the Noisy Labels Preservs $L_1$-Excess Error)**.** *Fix*

*$y : \mathbb{R}^d \mapsto \{\pm 1\}$ and let $C$ be a class of Boolean concepts. It holds*

$$\mathcal{E}_1(h, C; y) \leq \mathcal{E}_1(h, C; T_\rho y) + O(\rho)\ (\Gamma(C) + \Gamma(h))\,,$$

*where $\mathcal{E}(\cdot, \cdot; \cdot)$ is the excess error defined in Definition 9.2.5*

*Proof.* We first prove the following lemma showing that connects the excess error of a classifier $h(\cdot)$ with respect to the smoothed label $T_\rho y(\cdot)$ to its excess error with respect to the original label $y(\cdot)$. This is analogous to the real-valued case (Lemma 9.2.7). In the following lemma we show that we can connect the $L_1$-excess error with the correlation of concepts (which basically relies on the identity $|t - s| = 1 - ts$ when $t \in [-1, 1]$ and $s \in \{\pm 1\}$.

**Lemma 9.2.11** (From Excess Error to Correlation Preservation: Boolean Concepts)**.** *Let $h : \mathbb{R}^d \mapsto \{\pm 1\}$ and $C$ be a class of Boolean hypotheses. It holds*

$$\mathcal{E}_1(h, C; T_\rho y) - \mathcal{E}_1(h, C; y) \leq \sup_{f \in C} \left| \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[f(\mathbf{x}) T_\rho y(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[f(\mathbf{x}) y(\mathbf{x})] \right| + \left| \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[h(\mathbf{x}) T_\rho y(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[h(\mathbf{x}) y(\mathbf{x})] \right|.$$

*Proof.* We first note that $\mathcal{E}_1(h, C; T_\rho y) - \mathcal{E}_1(h, C; y) = \sup_{f \in C} \mathcal{E}_1(h, f; T_\rho y) - \sup_{f \in C} \mathcal{E}_1(h, f; y) \leq \sup_{f \in C} \left| \mathcal{E}_1(h, f; T_\rho y) - \mathcal{E}_1(h, f; y) \right|$. Using the fact that $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}}[|f_1(\mathbf{x}) - f_2(\mathbf{x})|] = 1 - \mathbf{E}_{\mathbf{x} \sim \mathcal{N}}[f_1(\mathbf{x}) f_2(\mathbf{x})]$, for any functions $f_1 : \mathbb{R}^d \mapsto [-1, 1]$ and $f_2 : \mathbb{R}^d \mapsto \{\pm 1\}$, we have that

$$\begin{aligned} \mathcal{E}_1(h, f; T_\rho y) &= \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[|T_\rho y(\mathbf{x}) - h(\mathbf{x})|] - \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[|T_\rho y(\mathbf{x}) - f(\mathbf{x})|] \\ &= \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[T_\rho y(\mathbf{x}) f(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[T_\rho y(\mathbf{x}) h(\mathbf{x})]\,. \end{aligned}$$

Therefore, for some concept $f \in C$, we have that

$$\left|\mathcal{E}_1(h, f; T_\rho y) - \mathcal{E}_1(h, f; y)\right| = \left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(T_\rho y(\mathbf{x}) - y(\mathbf{x}))f(\mathbf{x})]\right| + \left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(T_\rho y(\mathbf{x}) - y(\mathbf{x}))h(\mathbf{x})]\right| .$$

Taking the supremum over the $C$ completes the proof. □

First, note that since $|y(\mathbf{x})| \leq 1$, it also holds that $|T_\rho y(\mathbf{x})| \leq 1$. Using Lemma 9.2.11, we have that Proposition 9.2.10 is equivalent to showing that for a Boolean function $f : \mathbb{R}^d \mapsto \{\pm 1\}$ it holds $|\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(T_\rho y(\mathbf{x}) - y(\mathbf{x}))f(\mathbf{x})]| \leq O(\rho)\ \Gamma(f)$ . We do this in the following lemma.

**Lemma 9.2.12** ($T_\rho$ Preserves Correlation)**.** *Let $y : \mathbb{R}^d \mapsto \{\pm 1\}$ and let $f : \mathbb{R}^d \mapsto \{\pm 1\}$ be a (Borel) Boolean function. It holds that*

$$\left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})T_\rho y(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})]\right| \leq O(\rho)\ \Gamma(f)\,.$$

*Proof.* Using the fact that the Ornstein–Uhlenbeck noise operator $T_\rho$ is a symmetric linear operator on $L^2(\mathcal{N})$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})T_\rho y(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[y(\mathbf{x})T_\rho f(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[y(\mathbf{x})f(\mathbf{x})] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[y(\mathbf{x})(T_\rho f(\mathbf{x}) - f(\mathbf{x}))]\,.$$

Therefore,

$$\left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})T_\rho y(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})]\right| = \left|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[y(\mathbf{x})(T_\rho f(\mathbf{x}) - f(\mathbf{x}))]\right| \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|T_\rho f(\mathbf{x}) - f(\mathbf{x})|]\,,$$

where, for the inequality we used the fact that the label $y(\mathbf{x}) \in \{\pm 1\}$. We next bound the term $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|T_\rho f(\mathbf{x}) - f(\mathbf{x})|]$. We will use the following result from Ledoux and

Pisier as stated in [KOS08].

**Fact 9.2.13** (Ledoux-Pisier [Led94])**.** *Let $f : \mathbb{R}^d \mapsto \{\pm 1\}$ be a Boolean function. It holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})T_\rho f(\mathbf{x})] \geq 1 - 2\sqrt{\pi}\ \Gamma(f)\ \rho$.*

In what follows, we denote by $K$ the set labeled as positive by the LTF $f(\mathbf{x})$. Using the fact that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|T_\rho f(\mathbf{x}) - f(\mathbf{x})|] = 1 - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[T_\rho f(\mathbf{x}) f(\mathbf{x})]$, which holds because $|T_\rho f(\mathbf{x})| \leq 1$ and $f(\mathbf{x}) \in \{\pm 1\}$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|T_\rho f(\mathbf{x}) - f(\mathbf{x})|] = 1 - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})T_\rho f(\mathbf{x})] \leq O(\rho\Gamma(f))\,,$$

where the inequality follows from Fact 9.2.13.

□

Applying Lemma 9.2.12 on $f$ and $g$ gives the result. □

## 9.3 Agnostically Learning Real-valued Multi-index Models

In this section we present our algorithmic result Theorem 9.1.2 for learning real-valued function classes in the $L_2^2$ norm. For convenience, we first restate the class of bounded variation concepts that we consider.

**Definition 9.3.1** (Bounded Variation, Low-Dimensional Concepts)**.** *Fix $L, M > 0$ and $k \in \mathbb{Z}_+$. We define the class $\mathfrak{R}(M, L, k)$ of continuous, (almost everywhere) differentiable real-valued functions with the following properties:*

1. *For every* $f \in \mathfrak{R}(M, L, k)$*, it holds* $(\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f^4(\mathbf{x})])^{1/2} \leq M$ *and* $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[\|\nabla f(\mathbf{x})\|_2^2] \leq L$*.*

2. *There exists a subspace* $U$ *of* $\mathbb{R}^d$ *of dimension at most* $k$ *such that* $f$ *depends only on* $U$*, i.e., for every* $\mathbf{x} \in \mathbb{R}^d$*,* $f(\mathbf{x}) = f(\mathrm{proj}_U \mathbf{x})$*.*

We now state the main result of this section (the formal version of Theorem 9.1.2).

**Theorem 9.3.2** (Improper Learner for Real-valued Functions)**.** *Fix* $k \in \mathbb{N}$ *and* $M, L \in \mathbb{R}^+$*. Let* $D$ *be a distribution on* $\mathbb{R}^d \times \mathbb{R}^+$ *such that the* $\mathbf{x}$*-marginal of* $D$ *is standard* $d$*-dimensional normal. There exists an algorithm that makes* $N_q = \mathrm{poly}(d/\epsilon)$ *queries, draws* $N_s = \mathrm{poly}(d) + \mathrm{poly}((kM/\epsilon)^{L^2/\epsilon^4}, 1/\epsilon, \log(1/\delta))$ *samples from* $D$*, runs in time* $\mathrm{poly}(N_s, N_q, d)$ *and outputs a polynomial* $p : \mathbb{R}^d \mapsto \mathbb{R}$ *so that with probability at least* $1 - \delta$ *it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(p(\mathbf{x}) - y)^2] \leq \inf_{f\in\mathfrak{R}(M,L,k)} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - y)^2] + \epsilon \,.$$

Before we proceed to the proof we define the Hermite expansion operator that maps a function $f$ to its degree $m$ Hermite polynomial.

**Definition 9.3.3** (Hermite Expansion Operator)**.** *Given a function* $f \in L^2(\mathcal{N})$*, we denote by* $P_m(f)(\mathbf{x})$*, the linear operator that maps* $f$ *to the Hermite polynomial of degree* $m$ *of* $f$*, i.e.,*

$$(\mathrm{P}_m f)(\mathbf{x}) = \sum_{|I|\leq m} \widehat{f}(I)\mathrm{He}_I(\mathbf{x}),$$

*where* $\mathrm{He}_I$ *is the multivariate Hermite polynomial of degree* $I \in \mathbb{N}^d$ *and* $\widehat{f}(I) = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})\mathrm{He}_I(\mathbf{x})]$ *is the corresponding Hermite coefficient of* $f(\mathbf{x})$*.*

The following lemma bounds the error of the polynomial approximation of degree $m$ for "smooth" functions. Its proof is implicit in [KTZ19]; we provide a short proof for completeness.

**Lemma 9.3.4** (Polynomial Approximation of Smooth Functions)**.** *Let $f(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ be an (almost everywhere) differentiable function and $m \in \mathbb{N}$. It holds*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \mathrm{P}_m f(\mathbf{x}))^2] \leq O\Big(\frac{1}{m}\Big) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2] \,.$$

*Proof.* We denote as $\mathrm{P}_{>m} f$ the Hermite expansion of $f$, which contains the terms with degrees higher than $m$. We have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \mathrm{P}_m f(\mathbf{x}))^2] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_{>m} f(\mathbf{x}))^2] = \sum_{I:|I|>m} (\widehat{f}(I))^2 \leq \frac{1}{m} \sum_{I:|I|>m} |I|(\widehat{f}(I))^2 \,,$$

where in the last inequality, we used that $1 \leq |I|/m$. Furthermore, (see, e.g., the proof of Lemma 6 in [KTZ19]) we have that for a continuous and (almost everywhere) differentiable function $f$, it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2] = \sum_{I\in\mathbb{N}^d} |I|(\widehat{f}(I))^2 \,.$$

Combining the above, the result follows. □

As we discussed in Section (1.3.8) to show that an approximately optimal, low-dimensional concept exists we will use the Gaussian Marginalization Operator defined below.

**Definition 9.3.5** (Gaussian Marginalization Operator)**.** *Let $U$ be a subspace of $\mathbb{R}^d$. Denote by $D_{U^\perp}$ the standard normal distribution on the subspace $U^\perp$ (we assume that a*

*vector* $\mathbf{z} \sim D_{U^\perp}$ *is a d-dimensional vector that lies in* $U^\perp$*). Given a function* $f \in L^2(\mathcal{N})$*, we denote by* $\Pi_U f$ *the linear operator defined by*

$$(\Pi_U f)(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{z}\sim D_{U^\perp}}[f(\mathrm{proj}_U(\mathbf{x}) + \mathbf{z})] \,.$$

**Motivation about the Gaussian Marginalization Operator,** $\Pi_V$ By the assumption of Proposition 9.1.8, all directions in the orthogonal complement $V^\perp$ are low-influence, i.e., for $\mathbf{h} \in V^\perp$ it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{h}\cdot\nabla\widetilde{y}(\mathbf{x}))^2] \le O(\epsilon^2/k)$. In words, the function $\widetilde{y}$ is "approximately constant" along some low-influence direction $\mathbf{h}$. Let us first assume that $\widetilde{y}$ is exactly constant on all directions of $V^\perp$. Then, in order to preserve the correlation of $\widetilde{y}$ with $f$, *we only need to match the expected value of* $f$ *over* $V^\perp$. This motivates the following "Gaussian Marginalization Operator" of Definition 9.3.5. Indeed, if $\widetilde{y}$ was constant on $V^\perp$, using the fact that $\mathrm{proj}_V\mathbf{x}$ and $\mathrm{proj}_{V^\perp}\mathbf{x}$ are independent standard Gaussians, we would obtain that

$$\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}}[\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}} f(\mathrm{proj}_V(\mathbf{x}) + \mathrm{proj}_{V^\perp}(\mathbf{z}))\widetilde{y}(\mathbf{x})]] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})\widetilde{y}(\mathbf{x})] = 0 \,.$$

We observe that since $\Pi_V f$ is a convex combination of different translations of $f$ and $\mathfrak{B}(\Gamma, k)$ is closed under translations, we obtain that the Gaussian surface area of $f$ is also bounded above by $\Gamma$.

In the next lemma, we collect some useful properties of the Gaussian Marginalization Operator.

**Lemma 9.3.6.** *Let* $g \in L^2(\mathcal{N})$ *and* $V \subseteq \mathbb{R}^d$*. We have the following properties for the operator* $\Pi_V$*.*

- $\Pi_V$ *are contractions, i.e.,* $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_V g(\mathbf{x}))^2] \le \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[g^2(\mathbf{x})]$.

- *Let $U, V \subseteq \mathbb{R}^d$, it holds that $\Pi_V \Pi_{U+V} g = \Pi_{V+U} \Pi_{V+U^\perp} g = \Pi_V g$.*

*Proof.* To show that $\Pi_V$ is a contraction, note that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_V g(\mathbf{x}))^2] \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}} \mathbf{E}_{\mathbf{z}\sim\mathcal{N}_{V^\perp}}[g^2(\mathbf{x}_V + \mathbf{z})] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[g^2(\mathbf{x})]$, where we used Jensen's inequality. For the second part, let $H = (U+V)^\perp$ and note that

$$
\begin{aligned}
&\Pi_V \Pi_{U+V} g \\
&= \Pi_V \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_H}[g(\mathbf{x}_{U+V} + \mathbf{z})] = \Pi_V \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_H}[g(\mathbf{x}_V + \mathbf{x}_{U/V} + \mathbf{z})] = \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_H}[\Pi_V g(\mathbf{x}_V + \mathbf{x}_{U/V} + \mathbf{z})] \\
&= \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_H}[\mathop{\mathbf{E}}_{\mathbf{z}'\sim\mathcal{N}_{U/V}}[g(\mathbf{x}_V + \mathbf{z}' + \mathbf{z})] = \Pi_{U+V} \Pi_{U^\perp+V} g \ ,
\end{aligned}
$$

where we used Fubini's theorem. □

**Lemma 9.3.7.** *Let $g \in L^2(\mathcal{N})$, $m \in \mathbb{N}$ and $V \subseteq \mathbb{R}^d$. We have the following properties for the operators $\mathrm{P}_m$ and $\Pi_V$.*

- *$\mathrm{P}_m$ is a contraction, i.e., $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_m g(\mathbf{x}))^2] \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[g^2(\mathbf{x})]$.*
- *$\mathrm{P}_m$ and $\Pi_V$ commute, i.e., $\mathrm{P}_m \Pi_V g = \Pi_V \mathrm{P}_m g$.*

*Proof.* First, we show that $\mathrm{P}_m$ is a contraction. Using the $g \in L^2(\mathcal{N})$, we have that $g$ admits a Hermite expansion. We denote as $\mathrm{P}_{>m} g$ the Hermite expansion of $g$, which contains the terms with degrees higher than $m$. We have that

$$
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[g^2(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_m g(\mathbf{x}) + P_{>m} g(\mathbf{x}))^2] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_m g(\mathbf{x}))^2 + (P_{>m} g(\mathbf{x}))^2] \ ,
$$

where in the last equality, we used that the Hermite basis is orthogonal, hence $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\mathrm{P}_m g(\mathbf{x}) P_{>m} g(\mathbf{x})] = 0$. Therefore, $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[g^2(\mathbf{x})] \geq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_m g(\mathbf{x}))^2]$.

Next, we show that $\Pi_U$ and $\mathrm{P}_m$ commute.

**Claim 9.3.8** ($\mathrm{P}_m$ and $\Pi_U$ commute)**.** *Let $g \in L^2(\mathcal{N})$, $m \in \mathbb{N}$, and $V$ be a subspace of $\mathbb{R}^d$. It holds that $\mathrm{P}_m \Pi_V g = \Pi_V \mathrm{P}_m g$.*

*Proof.* Because $\mathrm{P}_m$ and $\Pi_V$ are linear operators, it suffices to show the above each term of the Hermite basis, i.e.,

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V g(\mathbf{x})\mathrm{He}_I(\mathbf{x})]\mathrm{He}_I(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_{V^\perp}}[\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[g(\mathbf{x})\mathrm{He}_I(\mathbf{x})]\mathrm{He}_I(\mathbf{x}_V+\mathbf{z})] \ .$$

Note that if $\mathrm{He}_I(\mathbf{x})$ does not depend on $V^\perp$, then $\mathbf{E}_{\mathbf{z}\sim\mathcal{N}_{V^\perp}}[\mathrm{He}_I(\mathbf{x}_V+\mathbf{z})] = \mathrm{He}_I(\mathbf{x})$. Therefore, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_{V^\perp}}[\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[g(\mathbf{x})\mathrm{He}_I(\mathbf{x})]\mathrm{He}_I(\mathbf{x}_V+\mathbf{z})] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[g(\mathbf{x})\mathrm{He}_I(\mathbf{x})]\mathrm{He}_I(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_V}[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_{V^\perp}}[g(\mathbf{x}_V+\mathbf{z})\mathrm{He}_I(\mathbf{x}+\mathbf{z})]]\mathrm{He}_I(\mathbf{x}) \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_V}[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_{V^\perp}}[g(\mathbf{x}_V+\mathbf{z})]\mathrm{He}_I(\mathbf{x})]\mathrm{He}_I(\mathbf{x}) \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V g(\mathbf{x})\mathrm{He}_I(\mathbf{x})]\mathrm{He}_I(\mathbf{x}) \ .
\end{aligned}$$

In the case where $\mathrm{He}_I(\mathbf{x})$ depends on $V^\perp$, we have that $\mathbf{E}_{\mathbf{z}\sim V^\perp}[\mathrm{He}_I(\mathbf{x}_V+\mathbf{z})] = 0$. Therefore, it suffices to prove that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V g(\mathbf{x})\mathrm{He}_I(\mathbf{x})]\mathrm{He}_I(\mathbf{x}) = 0$. Note that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V g(\mathbf{x})\mathrm{He}_I(\mathbf{x})]\mathrm{He}_I(\mathbf{x}) &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_V}[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_{V^\perp}}[\Pi_V g(\mathbf{x}+\mathbf{z})]\mathrm{He}_I(\mathbf{x}_V+\mathbf{z})]\mathrm{He}_I(\mathbf{x}) \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_V}[\Pi_V g(\mathbf{x}) \mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_{V^\perp}}[\mathrm{He}_I(\mathbf{x}_V+\mathbf{z})]]\mathrm{He}_I(\mathbf{x}) = 0 \ .
\end{aligned}$$

□

This completes the proof of Lemma 9.3.7. □

**Algorithm 21** Estimation of the influence matrix $\mathbf{M}$ with Queries

**Input:** $\epsilon > 0$, $\delta > 0$; sample & query access to distribution $\mathcal{D}$.
**Output:** Estimate of $\mathbf{M} = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[D_\rho y(\mathbf{x})\, D_\rho y(\mathbf{x})^\intercal\right]$.
Set $\rho \leftarrow C\,\epsilon^2$, $\eta \leftarrow C\,\epsilon^2$ for a sufficiently small constant $C > 0$.
Draw $N$ samples $S_N = \{\mathbf{x}^{(i)}\}_{i=1}^N \sim \mathcal{D}$.
**for** each $\mathbf{x}^{(i)} \in S_N$ **do**
    Compute $\widehat{D_\rho y}(\mathbf{x}^{(i)})$ by invoking Algorithm 20 on $\mathbf{x}^{(i)}$.
**end for**
**Return:** $\widehat{\mathbf{M}} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \widehat{D_\rho y}(\mathbf{x}^{(i)})\, \widehat{D_\rho y}(\mathbf{x}^{(i)})^\intercal$.

Having access to the gradient, enables us to calculate the influence matrix of the function which captures the sensitivity of the function in different directions. We formally define the influence matrix of a function $g$.

**Definition 9.3.9** (Influence Matrices)**.** *Given a differentiable $g \in L^2(\mathcal{N})$, we define the influence matrix as*

$$\mathbf{Inf}_g \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\nabla g(\mathbf{x}) \nabla g(\mathbf{x})^\intercal].$$

*Fix $\rho \in (0,1)$. Given $g \in L^2(\mathcal{N})$ (not necessarily differentiable), we define its $\rho$-smoothed influence matrix as*

$$\mathbf{Inf}_g^\rho \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[D_\rho g(\mathbf{x})(D_\rho g(\mathbf{x}))^\intercal]\,.$$

### 9.3.1 Influence PCA for Learning in $L_2^2$

In this section we show that for learning real-valued concepts of bounded variation in $L_2^2$ we can effectively reduce the dimension of the problem via PCA in the influence of the smoothed label $T_\rho y$. We show that the low-degree polynomial approximation of the smoothed label $T_\rho y$ can be projected down to the subspace $V$ via the Gaussian Marginalization Operator. In other words, we construct an explicit polynomial approxi-

mation of the label $T_\rho$ that depends only on the low-dimensional subspace $V$. We now state our dimension-reduction result.

**Proposition 9.3.10.** *Fix $\epsilon, M, L, Q > 0$ and let $\psi : \mathbb{R}^d \mapsto \mathbb{R}$ with $|\psi(\mathbf{x})| \leq Q$ and $\|\nabla\psi(\mathbf{x})\|_2 \leq \Psi$. Let $\eta$ be sufficiently small multiple of $\epsilon^2/(kM)$ and $m$ be sufficiently large multiple of $(Q^2L)/\epsilon^2$. Let $\widehat{\mathbf{M}}$ so that $\|\mathbf{Inf}_\psi - \widehat{\mathbf{M}}\|_2 \leq \eta/2$ and let $V$ be the subspace spanned by all the eigenvectors of $\widehat{\mathbf{M}}$ whose corresponding eigenvalues are at least $\eta$. Then, it holds*

1. $$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_m\Pi_V\psi(\mathbf{x}) - \psi(\mathbf{x}))^2] \leq \inf_{f\in\mathfrak{R}(M,L,k)} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[((\psi(\mathbf{x}) - f(\mathbf{x}))^2] + \epsilon\ .$$

2. *The dimension of $V$ is at most $O(\Psi^2/\eta)$.*

*Proof of Proposition 9.3.10.* Fix $f \in \mathfrak{R}(M, L, k)$. By assumption, there exists a subspace $U$ of dimension at most $k$, so that $f$ depends only on $U$, i.e., $f(\mathbf{x}) = f(\mathrm{proj}_U\mathbf{x})$. Therefore, $\Pi_{U+V}f(\mathbf{x}) = f(\mathbf{x})$.

**Lemma 9.3.11** (Excess $L_2^2$ Error Decomposition)**.** *We have*

$$\mathcal{E}_2(P_m\Pi_V\psi, f; \psi) \leq Q\underbrace{(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - P_mf(\mathbf{x}))^2])^{1/2}}_{\textit{Polynomial Approximation Error}} + 2\underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\psi(\mathbf{x}) - \Pi_V\psi(\mathbf{x}))f(\mathbf{x})]}_{\textit{Correlation Error}}\ .$$

*Proof.* We have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\psi(\mathbf{x}) - \mathrm{P}_m\Pi_V\psi(\mathbf{x}))^2] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\psi(\mathbf{x}) - f(\mathbf{x}))^2] \\
&\quad = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_m\Pi_V\psi(\mathbf{x}))^2 - f^2(\mathbf{x})] + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\psi(\mathbf{x})(f(\mathbf{x}) - \mathrm{P}_m\Pi_V\psi(\mathbf{x}))] \\
&\quad = \underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_m\Pi_V\psi(\mathbf{x}))^2 - f^2(\mathbf{x})] + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V\psi(\mathbf{x})(f(\mathbf{x}) - \mathrm{P}_m\Pi_V\psi(\mathbf{x}))]}_{I} \\
&\quad + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\psi(\mathbf{x}) - \Pi_V\psi(\mathbf{x}))f(\mathbf{x})] \,,
\end{aligned}$$

where we used that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\psi(\mathbf{x}) - \Pi_V\psi(\mathbf{x}))\mathrm{P}_m\Pi_V\psi(\mathbf{x})] = 0$. Furthermore, note that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V\psi(\mathbf{x})\mathrm{P}_m\Pi_V\psi(\mathbf{x})] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_m\Pi_V\psi(\mathbf{x}))^2]$, therefore, we have that

$$\begin{aligned}
I &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[-(\mathrm{P}_m\Pi_V\psi(\mathbf{x}))^2 - f^2(\mathbf{x})] + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V\psi(\mathbf{x})f(\mathbf{x})] \\
&\leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[-(\mathrm{P}_m\Pi_V\psi(\mathbf{x}))^2 - (\mathrm{P}_m f(\mathbf{x}))^2] + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V\psi(\mathbf{x})f(\mathbf{x})] \\
&= -\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_m\Pi_V\psi(\mathbf{x}) - \mathrm{P}_m f(\mathbf{x}))^2] + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V\psi(\mathbf{x})(f(\mathbf{x}) - \mathrm{P}_m f(\mathbf{x}))] \\
&\leq 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V\psi(\mathbf{x})(f(\mathbf{x}) - \mathrm{P}_m f(\mathbf{x}))] \,.
\end{aligned}$$

Using that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_V\psi(\mathbf{x}))^2] \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\psi(\mathbf{x}))^2] \leq Q^2$ and Cauchy-Schwarz inequality we get that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\Pi_V\psi(\mathbf{x})(f(\mathbf{x}) - \mathrm{P}_m f(\mathbf{x}))] \leq Q\,\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - P_m f(\mathbf{x}))^2]^{1/2}$. This completes the proof of Lemma 9.3.11. □

**Lemma 9.3.12** (Correlation Error Bound)**.** *It holds*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\psi(\mathbf{x}) - \Pi_V\psi(\mathbf{x}))f(\mathbf{x})] \leq O(\epsilon) \,. \tag{9.1}$$

*Proof.* Note that $f(\mathbf{x})$ depends only on the subspace $U$, therefore, $\Pi_{U+V}f(\mathbf{x}) = f(\mathbf{x})$.

Therefore, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\psi(\mathbf{x})-\Pi_V\psi(\mathbf{x}))f(\mathbf{x})] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{V+U}\psi(\mathbf{x})-\Pi_{V+U}\Pi_V\psi(\mathbf{x}))f(\mathbf{x})] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{V+U}\psi(\mathbf{x})-\Pi_V\Pi_{V+U}\psi(\mathbf{x}))f(\mathbf{x})] \\ &\leq \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{V+U}\psi(\mathbf{x})-\Pi_V\Pi_{V+U}\psi(\mathbf{x}))^2]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f^2(\mathbf{x})]\right)^{1/2},\end{aligned}$$

where in the last equality we used Lemma 9.3.7 and in the last inequality we used the Cauchy-Schwarz inequality. Note that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f^2(\mathbf{x})] \leq M$. To bound the other term we show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{V+U}\psi(\mathbf{x})-\Pi_V\Pi_{U+V}\psi(\mathbf{x}))^2]$ is small. For that, we prove a generalization of Lemma 9.4.14.

**Lemma 9.3.13** (Generalized Gaussian Marginalization Error)**.** *Let $g:\mathbb{R}^d\mapsto\mathbb{R}$ be a function in $L^2(\mathcal{N})$ such that $\nabla g\in L^2(\mathcal{N})$ and let $V,U$ be subspaces of $\mathbb{R}^d$. It holds*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{V+U}g(\mathbf{x})-\Pi_V\Pi_{V+U}g(\mathbf{x}))^2] \leq \dim(V^\perp\cap U)\max_{\mathbf{v}\in V^\perp\cap U,\|\mathbf{v}\|_2=1}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\nabla g(\mathbf{x})\cdot\mathbf{v})^2].$$

*Proof.* Assume that $\dim(V^\perp\cap U)=k\leq d$. Using the rotation invariance of the Gaussian distribution, without loss of generality, we may assume that $\mathbf{e}_1,\ldots,\mathbf{e}_k$ is a basis of $V^\perp\cap U$. Note that it holds $\Pi_V\Pi_{U+V}g(\mathbf{x})=\Pi_{V+U^\perp}\Pi_{V+U}g(\mathbf{x})=\Pi_{V+U}\Pi_{V+U^\perp}g(\mathbf{x})$. We

have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{U+V}g(\mathbf{x}) - \Pi_V\Pi_{U+V}g(\mathbf{x}))^2] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{U+V}g(\mathbf{x}) - \Pi_{V+U}\Pi_{V+U^\perp}g(\mathbf{x}))^2] \\
&\le \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(g(\mathbf{x}) - \Pi_{V+U^\perp}g(\mathbf{x}))^2] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}_{k+1},\ldots,\mathbf{x}_d\sim\mathcal{N}}[\mathbf{Var}_{\mathbf{x}_1,\ldots\mathbf{x}_k\sim\mathcal{N}}[g(\mathbf{x}_1,\ldots,\mathbf{x}_d)]] \\
&\le \frac{1}{2}\mathop{\mathbf{E}}_{\mathbf{x}_{k+1},\ldots,\mathbf{x}_d\sim\mathcal{N}}\left[\sum_{i=1}^{k}\mathop{\mathbf{E}}_{\mathbf{x}_1,\ldots,\mathbf{x}_k\sim\mathcal{N}}[\mathbf{Var}_{\mathbf{x}_i\sim\mathcal{N}}[g(\mathbf{x}_1,\ldots,\mathbf{x}_i,\ldots\mathbf{x}_d)]]\right],
\end{aligned}$$

where in the inequality, we used Efron-Stein's inequality. Using Fact 9.4.15, for each $i \in [k]$ we have $\mathbf{Var}_{\mathbf{x}_i\sim\mathcal{N}}[g(\mathbf{x}_1,\ldots,\mathbf{x}_i,\ldots\mathbf{x}_d)] \le \mathbf{E}_{\mathbf{x}_i\sim\mathcal{N}}[(\nabla g(\mathbf{x}_1,\ldots,\mathbf{x}_d)\cdot\mathbf{e}_i)^2]$, and therefore we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(g(\mathbf{x}) - r(\mathbf{x}))^2] \le \sum_{i=1}^{k}\mathop{\mathbf{E}}_{\mathbf{x}_1,\ldots,\mathbf{x}_d\sim\mathcal{N}}[(\nabla g(\mathbf{x})\cdot\mathbf{e}_i)^2] \le k\max_{\mathbf{v}\in H^\perp,\|\mathbf{v}\|_2=1}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\nabla g(\mathbf{x})\cdot\mathbf{v})^2].$$

This completes the proof of Lemma 9.3.13. □

From Lemma 9.3.13, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{V+U}\psi(\mathbf{x}) - \Pi_V\Pi_{U+V}\psi(\mathbf{x}))^2] \le \dim(U\cap V^\perp)\max_{\mathbf{v}\in U\cap V^\perp,\|\mathbf{v}\|_2=1}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[((\nabla\psi(\mathbf{x}))\cdot\mathbf{v})^2].$$

Furthermore, $\max_{\mathbf{v}\in U\cap V^\perp,\|\mathbf{v}\|_2=1}\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[((\nabla\psi(\mathbf{x}))\cdot\mathbf{v})^2] \le \eta/2 + \max_{\mathbf{v}\in U\cap V^\perp,\|\mathbf{v}\|_2=1}\mathbf{v}^\intercal\widehat{\mathbf{M}}\mathbf{v} \le 2\eta$ because the subspace $U\cap V^\perp$ contains vectors with influence at most $\eta$. Note that $\dim(U\cap V^\perp) \le \dim(U) \le k$ and noting $\eta = O(\epsilon^2/(Mk))$ completes the proof of Lemma 9.3.12. □

Combining Lemmas 9.3.11 and 9.3.12 and using that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \mathrm{P}_m f(\mathbf{x}))^2] \le L/m$

from Lemma 9.3.4, we get that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{P}_m\Pi_V\psi(\mathbf{x})-\psi(\mathbf{x}))^2] \le \inf_{f\in\mathfrak{R}(M,L,k)} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[((\psi(\mathbf{x})-f(\mathbf{x}))^2] + \epsilon \ .$$

To show that the subspace $V$ has small dimension, we show the following lemma

**Lemma 9.3.14.** *Fix $\eta > 0, \rho \in (0,1)$. Let $\psi$ be a function from $\mathbb{R}^d$ to $\mathbb{R}$ such that $\|\nabla\psi(\mathbf{x})\|_2 \le \Psi$ and let $V$ be the subspace spanned by all the eigenvectors of $\mathbf{Inf}_g$ with eigenvalue at least $\eta$. Then the dimension of the subspace $V$ is $\dim(V) = O(\Psi^2/\eta)$.*

*Proof.* Let $m = \dim(V)$. $V$ is spanned by the eigenvectors of $\mathbf{Inf}_g = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\nabla\psi(\mathbf{x})(\nabla\psi(\mathbf{x}))^\intercal]$ with eigenvalue at least $\eta$, hence,

$$m\eta \le \mathrm{tr}(\mathbf{Inf}_g) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\mathrm{tr}(\nabla\psi(\mathbf{x})(\nabla\psi(\mathbf{x}))^\intercal)] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla\psi(\mathbf{x})\|_2^2] \ .$$

From the assumption, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla\psi(\mathbf{x})\|_2^2] = O(\Psi^2)$. Therefore, we have that $m \le O(\Psi^2/\eta)$. □

An application of the lemma above (Lemma 9.3.14) gives, which gives that the subspace it at most $O(\Psi^2/\eta)$. This completes the proof of Proposition 9.3.10 □

### 9.3.2 Proof of Theorem 9.3.2

We will use the following fact about the $L_2$ polynomial regression.

**Fact 9.3.15** (see, e.g., Theorem D.7 [DKK+21a])**.** *Let $D$ be a distribution on $\mathbb{R}^d \times \mathbb{R}$ such that the $\mathbf{x}$-marginal of $D$ is standard $d$-dimensional normal and the labels $y$ are bounded by $M$. The $L_2$-regression algorithm draws $N = \mathrm{poly}((dm)^{m^2}, 1/\epsilon, M, \log(1/\delta))$*

*samples from $D$, runs in time* $\mathrm{poly}(N, d)$*, and outputs a polynomial* $p : \mathbb{R}^d \mapsto \mathbb{R}$ *such that with probability at least* $1 - \delta$ *it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(p(\mathbf{x}) - y)^2] \leq \min_{p\in\mathcal{P}_m} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(p(\mathbf{x}) - y)^2] + \epsilon \ ,$$

*where $\mathcal{P}_m$ is the class of polynomials with degree at most $m$.*

We first show that we can truncate the labels with $|y(\mathbf{x})| \geq M' = M^{1/2}/\epsilon^{1/2}$ without increasing the error by a lot. From Markov's inequality, we have that

$$\mathbf{Pr}[|f(\mathbf{x})| \geq M'] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f^2(\mathbf{x})]/(M')^2 \leq \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f^4(\mathbf{x})]}/(M')^2 \leq \epsilon \ .$$

Let $\mathrm{trunc}(y(\mathbf{x})) = \mathrm{sign}(y(\mathbf{x}))\min(|y(\mathbf{x})|, M')$. We have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \mathrm{trunc}(y(\mathbf{x})))^2] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \mathrm{trunc}(y(\mathbf{x})))^2(\mathbb{1}\{|f(\mathbf{x})| \leq M'\} + \mathbb{1}\{|f(\mathbf{x})| > M'\})] \\
&\leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - y(\mathbf{x}))^2] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \mathrm{trunc}(y(\mathbf{x})))^2\mathbb{1}\{|f(\mathbf{x})| > M'\}] \\
&\leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - y(\mathbf{x}))^2] + 2(\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f^4(\mathbf{x})]} + (M')^2)\sqrt{\mathbf{Pr}[|f(\mathbf{x})| \geq M']} \\
&\leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - y(\mathbf{x}))^2] + \epsilon \ .
\end{aligned}$$

For the rest of the proof, we assume that $y(\mathbf{x})$ is truncated at $M'$. Let $\psi(\mathbf{x}) = T_\rho y$ for $\rho = \mathrm{poly}(\epsilon/(ML))$. Note that $\|\nabla\psi(\mathbf{x})\|_2 \leq M'$. From Lemma 9.2.3, with $N = \mathrm{poly}(d/\epsilon)\log(1/\delta)$ queries, we get that with probability $1 - \delta/2$ a matrix $\mathbf{M}$, so that $\|\mathbf{M} - \mathbf{Inf}_\psi\|_F \leq \epsilon$. Applying Proposition 9.3.10 to the matrix $\mathbf{M}$, we get that in the subspace $V$ spanned by the eigenvectors of the matrix $\mathbf{M}$ with eigenvalues larger

than $\eta = \text{poly}(\epsilon/Mk))$ with dimension at most $O(\text{poly}(M', 1/\eta, 1/\epsilon))$, there exists a polynomial $p : V \mapsto \mathbb{R}$ of degree $m = \text{poly}(M_2/\epsilon)$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[p^2(\mathbf{x})] \leq \mathbf{E}[\psi^2(\mathbf{x})] \leq (M')^2$, so that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(p(\mathbf{x}) - \psi(\mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \psi(\mathbf{x}))^2] + \epsilon/2 \; .$$

From Proposition 9.2.6, we get that for the same polynomial and using that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla p(\mathbf{x})\|_2 \leq m\, \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[p^2(\mathbf{x})]$, it also holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(p(\mathbf{x}) - y(\mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - y(\mathbf{x}))^2] + \epsilon/2 \; .$$

Let $\mathbf{P} : \mathbb{R}^d \mapsto V$ be the projection matrix to the subspace $V$. Let $(\mathbf{P}\mathbf{x}, y) \sim D'$, where $(\mathbf{x}, y) \sim D$. We use the $L_2$-regression algorithm on $D'$ and from Fact 9.3.15, using $\text{poly}((kM/\epsilon)^{L^2/\epsilon^4}, 1/\epsilon, \log(1/\delta))$ samples from $D'$, we get a polynomial $p' : V \mapsto \mathbb{R}$ so that with probability at least $1 - \delta$, it holds

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(p'(\mathbf{P}\mathbf{x}) - y(\mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(p(\mathbf{x}) - y(\mathbf{x}))^2] + \epsilon/2 \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - y(\mathbf{x}))^2] + \epsilon \; .$$

This completes the proof of Theorem 9.3.2.

### 9.3.3 Applications of Theorem 9.3.2

In this section, we apply Theorem 9.3.2 for several real-valued activations. We start by applying our theorem for the class of ReLU activations.

**Theorem 9.3.16** (Improper Learner for ReLUs Activations)**.** *Fix $M \in \mathbb{R}_+$. Let $\mathcal{C}$ be the concept class containing all the ReLU activations with normal vectors bounded in $\ell_2$ norm by $M$. Let $D$ be a distribution on $\mathbb{R}^d \times \mathbb{R}$ such that the $\mathbf{x}$-marginal of $D$ is the standard*

*d-dimensional normal. There exists an algorithm that makes* $N_q = \mathrm{poly}(dM/\epsilon)$ *queries, draws* $N_s = \mathrm{poly}(d/\epsilon) + 2^{\mathrm{poly}(M/\epsilon)} \log(1/\delta)$ *samples from* $D$*, runs in time* $\mathrm{poly}(N_s, N_q, d)$ *and outputs a polynomial* $p : \mathbb{R}^d \mapsto \mathbb{R}$ *so that with probability at least* $1 - \delta$ *it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(p(\mathbf{x}) - y)^2] \leq \inf_{f\in\mathcal{C}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - y)^2] + \epsilon \, .$$

*Proof.* To prove the above theorem it suffices to show that $\mathcal{C} \subseteq \mathfrak{R}(\sqrt{3}M^2, M^2, 1)$. Note that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{ReLU}(\mathbf{w}\cdot\mathbf{x}))^4] \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{w}\cdot\mathbf{x})^4] \leq 3M^4$. Furthermore, we bound the derivative of the activation. We have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla_{\mathbf{x}}\mathrm{ReLU}(\mathbf{w}\cdot\mathbf{x})\|_2^2] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \geq 0\}\mathbf{w}\|_2^2] \leq M^2 \, .$$

Therefore, it follows that $\mathcal{C} \subseteq \mathfrak{R}(\sqrt{3}M^2, M^2, 1)$. An application of Theorem 9.3.2 gives the result. □

We next consider learning Single-index models (SIMs) with an unknown Lipschitz link function $g : \mathbb{R} \mapsto \mathbb{R}$, i.e., $f(\mathbf{x}) = g(\mathbf{w}\cdot\mathbf{x})$.

**Definition 9.3.17.** *We define the class of L-Lipschitz SIMs on* $\mathbb{R}^d$ *denoted* $\mathrm{SIM}(L, M)$ *as follows. For each* $f \in \mathrm{SIM}(L, M)$*,* $f(\mathbf{x}) = g(\mathbf{w}\cdot\mathbf{x})$*, for L-Lipschitz* $g : \mathbb{R} \mapsto \mathbb{R}$ *and* $\|\mathbf{w}\|_2 \leq M$*.*

**Theorem 9.3.18** (Improper Learner for SIMs)**.** *Fix* $L, M \in \mathbb{R}_+$*. Let* $D$ *be a distribution on* $\mathbb{R}^d \times \mathbb{R}$ *such that the* $\mathbf{x}$*-marginal of* $D$ *is the standard d-dimensional normal. There exists an algorithm that makes* $N_q = \mathrm{poly}(dL/\epsilon)$ *queries, draws* $N_s = \mathrm{poly}(d/\epsilon) + 2^{\mathrm{poly}(LM/\epsilon)} \log(1/\delta)$ *samples from* $D$*, runs in time* $\mathrm{poly}(N_s, N_q, d)$ *and outputs*

*a polynomial* $p : \mathbb{R}^d \mapsto \mathbb{R}$ *so that with probability at least* $1 - \delta$ *it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(p(\mathbf{x}) - y)^2] \leq \inf_{f \in \mathrm{SIM}(L,M)} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - y)^2] + \epsilon \; .$$

*Proof.* Note that for any $f \in \mathrm{SIM}(L)$ by definition if holds that $\|\nabla f(\mathbf{x})\|_2 \leq L$ and also that $\mathbf{E}[f^4(\mathbf{x})] \leq L^4 \, \mathbf{E}[(\mathbf{w} \cdot \mathbf{x})^4] \lesssim M^4 L^4$. Therefore, we have that $f \in \mathrm{SIM}(L, M) \subseteq \mathfrak{R}(M^2 L^2, L, 1)$. An application of Theorem 9.3.2 gives the result. □

We define the class of linear combinations of ReLU networks.

**Definition 9.3.19** (ReLU Networks)**.** *We define the class* $\mathfrak{Re}(M, k)$ *of ReLU networks as follows. For each* $f \in \mathfrak{Re}(M, k)$*,* $f(\mathbf{x}) = \mathbf{W}_2 \mathrm{ReLU}(\mathbf{W}_1 \mathbf{x})$*, for matrices* $\mathbf{W}_1 \in \mathbb{R}^{k \times d}, \mathbf{W}_2 \in \{\pm 1\}^{k \times 1}$*, with* $\|\mathbf{W}_1\|_{op} \leq M$*.*

We give our result for learning linear combinations of ReLUs, i.e., real-valued functions of the form $f(\mathbf{x}) = \sum_{i=1}^{k} a_i \mathrm{ReLU}(\mathbf{w}^{(i)} \cdot \mathbf{x})$, where $a_i \in \mathbb{R}$.

**Theorem 9.3.20** (Improper Learner for Linear Combinations of ReLUs)**.** *Fix* $k \in \mathbb{N}$ *and* $M \in \mathbb{R}_+$*. Let* $D$ *be a distribution on* $\mathbb{R}^d \times \mathbb{R}$ *such that the* $\mathbf{x}$*-marginal of* $D$ *is the standard* $d$*-dimensional normal. There exists an algorithm that makes* $N_q = \mathrm{poly}(dM/\epsilon)$ *queries, draws* $N_s = \mathrm{poly}(d/\epsilon) + (kM/\epsilon)^{\mathrm{poly}(kM/\epsilon)} \log(1/\delta)$ *samples from* $D$*, runs in time* $\mathrm{poly}(N_s, N_q, d)$ *and outputs a polynomial* $p : \mathbb{R}^d \mapsto \mathbb{R}$ *so that with probability at least* $1 - \delta$ *it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(p(\mathbf{x}) - y)^2] \leq \inf_{f \in \mathfrak{Re}(M,k)} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - y)^2] + \epsilon \; .$$

*Proof.* We show that $\mathfrak{Re}(M, k) \subseteq \mathfrak{R}(M', L', k)$ for appropriate parameters $M', L'$. We show the following

**Lemma 9.3.21.** *Let* $f(\mathbf{x}) = \sum_{i=1}^{k} a_i \mathrm{ReLU}(\mathbf{w}^{(i)} \cdot \mathbf{x})$ *where* $a_i \in \{\pm 1\} \in \mathbb{R}$ *and* $\mathbf{w}^{(i)} \in \mathbb{R}^d$ *with* $\|\mathbf{w}^{(i)}\|_2 \leq M$ *for all* $i \in [k]$*. Then, we have that* $f \in \mathfrak{R}(kM^2, kM^2, k)$*.*

*Proof.* We have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f^4(\mathbf{x})] \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\sum_{i=1}^{k} \mathrm{ReLU}(\mathbf{w}^{(i)} \cdot \mathbf{x}))^4]$. From the Cauchy-Schwarz inequality we have that $(\sum_{i=1}^{k} z_i)^2 \leq k \sum_{i=1}^{k} z_i^2$. Therefore, applying this inequality twice, we get that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f^4(\mathbf{x})] \leq k^3 \sum_{i=1}^{k} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{ReLU}(\mathbf{w}^{(i)} \cdot \mathbf{x}))^4] \leq O(k^3 M^4)$. We then bound the derivative of $f$. We have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla_{\mathbf{x}} f(\mathbf{x})\|_2^2] = k \sum_{i=1}^{k} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\mathbb{1}\{\mathbf{w}^{(i)} \cdot \mathbf{x} \geq 0\}\mathbf{w}^{(i)}\|_2^2] \leq O(kM^2) \; .$$

□

Then the proof follows from Lemma 9.3.21 along with Theorem 9.3.2. □

We now give an improved result for learning sums of ReLUs, i.e., real-valued functions of the form $f(\mathbf{x}) = \sum_{i=1}^{k} \mathrm{ReLU}(\mathbf{w}^{(i)} \cdot \mathbf{x})$. We first define the class of sum of ReLUs.

**Definition 9.3.22** (Sums of ReLU Networks)**.** *We define the class* $\mathfrak{Re}_+(M, k)$ *of ReLU networks as follows. For each* $f \in \mathfrak{Re}_+(M, k)$*,* $f(\mathbf{x}) = \mathrm{ReLU}(\mathbf{W}\mathbf{x})$*, for matrices* $\mathbf{W} \in \mathbb{R}^{k\times d}$*, with* $\mathbf{E}[f^2(\mathbf{x})] \leq M$*.*

**Theorem 9.3.23** (Improper Learner for Sums of ReLUs)**.** *Fix* $k \in \mathbb{N}$ *and* $M \in \mathbb{R}_+$*. Let* $D$ *be a distribution on* $\mathbb{R}^d \times \mathbb{R}_+$ *such that the* $\mathbf{x}$*-marginal of* $D$ *is the standard* $d$*-dimensional normal. There exists an algorithm that makes* $N_q = \mathrm{poly}(dM/\epsilon)$ *queries, draws* $N_s = \mathrm{poly}(d/\epsilon) + (kM/\epsilon)^{\mathrm{poly}(M/\epsilon)} \log(1/\delta)$ *samples from* $D$*, runs in time* $\mathrm{poly}(N_s, N_q, d)$ *and outputs a polynomial* $p : \mathbb{R}^d \mapsto \mathbb{R}$ *so that with probability at least*

$1-\delta$ *it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(p(\mathbf{x})-y)^2] \leq \inf_{f\in \mathfrak{R}e_+(M,k)} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x})-y)^2] + \epsilon \ .$$

*Proof.* Note that from Lemma 9.3.21 we have that for $f(\mathbf{x}) = \sum_{i=1}^k \mathrm{ReLU}(\mathbf{w}^{(i)}\cdot\mathbf{x})$, where $\mathbf{w}^{(i)} \in \mathbb{R}^d$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f^2(\mathbf{x})] \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\sum_{i=1}^k \mathrm{ReLU}(\mathbf{w}^{(i)}\cdot\mathbf{x}))^4] \leq M^2$. We show that $f \in \mathfrak{R}(kM^2, M^2, k)$. Similar to Theorem 9.3.20, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f^4(\mathbf{x})] \leq O(kM^2)$. The proof differs from Theorem 9.3.20 on the fact that we can bound the gradient of $f$ by the $L_2^2$ norm of $f$ yielding a bound independent of $k$ in the exponent. We show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla_{\mathbf{x}} f(\mathbf{x})\|_2^2] \leq O(M)$. We have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla_{\mathbf{x}} f(\mathbf{x})\|_2^2] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\|\sum_{i=1}^k \mathbb{1}\{\mathbf{w}^{(i)}\cdot\mathbf{x}\geq 0\}\mathbf{w}^{(i)}\|_2^2] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\sum_{i,j=1}^k \mathbb{1}\{\mathbf{w}^{(i)}\cdot\mathbf{x}\geq 0\}\mathbb{1}\{\mathbf{w}^{(j)}\cdot\mathbf{x}\geq 0\}\mathbf{w}^{(i)}\cdot\mathbf{w}^{(j)}] \\
&\leq 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[\sum_{i,j=1}^k \mathrm{ReLU}(\mathbf{w}^{(i)}\cdot\mathbf{x})\mathrm{ReLU}(\mathbf{w}^{(j)}\cdot\mathbf{x})] \\
&\leq 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f^2(\mathbf{x})] \leq O(M) \ .
\end{aligned}$$

Therefore, we have that $\mathcal{C} \subseteq \mathfrak{R}(M, 2M, k)$. An application of Theorem 9.3.2 gives the result. □

Next we show our result for a general ReLU network. We first define the clas of Deep ReLU networks.

**Definition 9.3.24** (Deep ReLU Networks)**.** *We define the class* $\mathfrak{D}(M, L, k, S)$ *of depth-*$(L+1)$ *ReLU networks as follows. For each* $f \in \mathfrak{D}(M, L, k)$, $f(\mathbf{x}) = \mathbf{W}_L\mathrm{ReLU}(\mathbf{W}_{L-1}\cdots \mathrm{ReLU}(\mathbf{W}_1\mathbf{x}))$*, for matrices* $\mathbf{W}_1 \in \mathbb{R}^{k\times d}, \ldots, \mathbf{W}_L \in \mathbb{R}^{k_L\times 1}$*, with* $\|\mathbf{W}_i\|_{op} \leq M$ *and*

$k_i \leq S$.

We show the following theorem.

**Theorem 9.3.25** (Agnostic Learner for Deep ReLU Networks)**.** *Fix $k, S, L \in \mathbb{N}$ and $M \in \mathbb{R}_+$. Let $D$ be a distribution on $\mathbb{R}^d \times \mathbb{R}^+$ such that the $\mathbf{x}$-marginal of $D$ is the standard $d$-dimensional normal. There exists an algorithm that makes $N_q = \mathrm{poly}(dM/\epsilon)$ queries, draws $N_s = \mathrm{poly}(d/\epsilon) + 2^{\mathrm{poly}(kSM/\epsilon)} \log(1/\delta)$ samples from $D$, runs in time $\mathrm{poly}(N_s, N_q, d)$ and outputs a polynomial $p : \mathbb{R}^d \mapsto \mathbb{R}$ so that with probability at least $1 - \delta$ it holds*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(p(\mathbf{x}) - y)^2] \leq \inf_{f \in \mathfrak{D}(M,L,k,S)} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - y)^2] + \epsilon \; .$$

*Proof.* We show that $\mathfrak{D}(M, L, k, S) \subseteq \mathfrak{R}(M', L', k)$ for appropriate parameters $M', L'$. We first calculate for each $f \in \mathfrak{D}(M, L, k, S)$, the $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2]$. Denote as $D_i(\mathbf{x}) = \mathbf{W}_i \mathrm{ReLU}(\mathbf{W}_{i-1} \cdots \mathrm{ReLU}(\mathbf{W}_1 \mathbf{x}))$ the sub-network of $f(\mathbf{x})$. From the product rule, we have that

$$\nabla f(\mathbf{x}) = \mathbf{W}_L \mathrm{diag}(\mathbb{1}\{D_{L-1} \geq 0\}) \mathbf{W}_{L-1} \cdots \mathrm{diag}(\mathbb{1}\{\mathbf{W}_1 \mathbf{x} \geq 0\}) \mathbf{W}_1 \; .$$

Therefore, we have that $\|\nabla f(\mathbf{x})\|_2 \leq \prod_{i=1}^{L} \|\mathbf{W}_i\|_{op} \sqrt{k_i} \leq (MS)^L$. Using the Poincare inequality, we can show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f^2(\mathbf{x})] \leq k \, \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[\|\nabla f(\mathbf{x})\|_2^2] \leq k(MS)^L$. Therefore, $\mathfrak{D}(M, L, k, S) \subseteq \mathfrak{R}((kMS)^{O(L)}, (kMS)^{O(L)}, k)$. Then the proof follows from Theorem 9.3.2. □

### 9.3.4 Proper Agnostic Query Learner for ReLUs

In this section we present our algorithmic result for agnostic proper learning of ReLU functions with queries. Our goal is to show Theorem 9.3.26 which we state below.

**Theorem 9.3.26** (Proper Agnostic Query Learner for ReLUs)**.** *Let $\mathcal{C}$ be the class of ReLUs on $\mathbb{R}^d$ with normal vectors bounded by $M > 0$ and denote by $y(\mathbf{x}) \in \mathbb{R}$ the label (chosen by an adversary) of $\mathbf{x} \in \mathbb{R}^d$. There exists an algorithm that makes $N_s = \mathrm{poly}(d/\epsilon)\log(1/\delta)$ sample queries, $N_q = \mathrm{poly}(d/\epsilon)\log(1/\delta)$ queries, and, with runtime $\mathrm{poly}(d)2^{\mathrm{poly}(1/\epsilon)}$, computes a ReLU activation $f(\mathbf{x})$ such that, with probability at least $1-\delta$, it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(h(\mathbf{x}) - y(\mathbf{x}))^2] \leq \inf_{c\in\mathcal{C}} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(c(\mathbf{x}) - y(\mathbf{x}))^2] + \epsilon$.*

Our algorithm is presented in Algorithm 22. It uses membership queries to estimate a matrix $M$ corresponding to the influence matrix of the appropriately smoothed label function $y(\mathbf{x})$. It then restricts attention to a small subspace $V$ given by the large eigenvectors of $\mathbf{M}$ and exhaustively searches for a near-optimal halfspace with a normal vector in $V$.

**Algorithm 22** Agnostic Proper Learning of ReLU Activations with Queries

1: **Input:** $\epsilon > 0$, $\delta > 0$; sample and query access to distribution $\mathcal{D}$.
2: **Output:** Hypothesis $h \in \mathcal{C}$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\Big[(h(\mathbf{x}) - y(\mathbf{x}))^2\Big] \leq \inf_{c\in\mathcal{C}} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\Big[(c(\mathbf{x}) - y(\mathbf{x}))^2\Big] + \epsilon$ with probability $1-\delta$.
3: $\rho \leftarrow C\,\epsilon^2, \quad \eta \leftarrow C\,\epsilon^2$, for sufficiently small constant $C > 0$.
4: Estimate $\mathbf{M} = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\Big[D_\rho y(\mathbf{x})\, D_\rho y(\mathbf{x})^{\intercal}\Big]$ using $\mathrm{poly}(d/\epsilon)$ queries via Algorithm 21.
5: Let $V$ be the subspace spanned by eigenvectors of $\mathbf{M}$ with eigenvalues $\geq \eta$.
6: Let $\mathcal{H}_V$ be the set of ReLUs with normal vectors in $V$. Compute the ERM hypothesis $h$ using $m = \Theta\Big(\frac{\dim(V)}{\epsilon^2}\log\frac{1}{\delta}\Big)$ i.i.d. samples from $\mathcal{D}$ in time $O\Big(m^{\dim(V)}\Big)$.
7: **Return:** $h$

### 9.3.5 Reducing the Dimension via Influence PCA

**Proposition 9.3.27** (Dimension Reduction via Influence PCA: ReLU)**.** *Fix $\epsilon > 0, M > 0$ and let $\psi : \mathbb{R}^d \mapsto \mathbb{R}$ be a differentiable function with $\|\nabla\psi(\mathbf{x})\|_2 \leq \Psi$ for all $\mathbf{x} \in \mathbb{R}^d$. Let $\eta$ be a sufficiently small multiple of $\epsilon^2/M$ and let $\widehat{\mathbf{M}} \in \mathbb{R}^{d\times d}$ be such that $\|\widehat{\mathbf{M}} - \mathbf{Inf}_\psi\|_2 \leq \eta/2$. Moreover, let $V$ be the subspace spanned by all the eigenvectors of $\widehat{\mathbf{M}}$ whose corresponding eigenvalues are at least $\eta$. The following hold true:*

1. *There exists $\mathbf{v} \in V$ with $\|\mathbf{v}\|_2 \leq M$ and $t \in \mathbb{R}$ such that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{ReLU}(\mathbf{v}\cdot\mathbf{x}+t)-\psi(\mathbf{x}))^2] \leq \inf_{\mathbf{w}\in\mathbb{R}^d,\|\mathbf{w}\|_2\leq M, t\in\mathbb{R}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{ReLU}(\mathbf{w}\cdot\mathbf{x}+t)-\psi(\mathbf{x}))^2]+\epsilon\,.$$

2. *The dimension of $V$ is at most $O(\Psi^2/\eta)$.*

*Proof.* Suppose for the sake of contradiction that there exists a $\mathbf{w} \in \mathbb{R}^d, t \in \mathbb{R}$ such that for every $\mathbf{v} \in V, t' \in \mathbb{R}$, it holds

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{ReLU}(\mathbf{v}\cdot\mathbf{x}+t')-\psi(\mathbf{x}))^2] \geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{ReLU}(\mathbf{w}\cdot\mathbf{x}+t)-\psi(\mathbf{x}))^2]+\epsilon\,,$$

the above is equivalent to

$$\begin{aligned} &2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{ReLU}(\mathbf{w}\cdot\mathbf{x}+t)-\mathrm{ReLU}(\mathbf{v}\cdot\mathbf{x}+t'))\psi(\mathbf{x}))] \\ &\geq \epsilon + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{ReLU}(\mathbf{w}\cdot\mathbf{x}+t))^2] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{ReLU}(\mathbf{v}\cdot\mathbf{x}+t'))^2]\,, \end{aligned} \tag{9.2}$$

Let $f(\mathbf{x}) = \mathrm{ReLU}(\mathbf{w}\cdot\mathbf{x}+t)$ and $f_V(\mathbf{x}) = \Pi_V f(\mathbf{x})$. Note that $\mathbf{w}_{V^\perp} \neq \mathbf{0}$, since otherwise we would have that the vector of $f$ would be inside $V$. Note that $f_V$ is a

convex combination of ReLUs with vectors in $V$, hence Equation (9.2) becomes

$$2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \Pi_V f(\mathbf{x}))\psi(\mathbf{x}))] \geq \epsilon \,, \tag{9.3}$$

where we used that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x})^2] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathrm{ReLU}(\mathbf{w}_V \cdot \mathbf{x} + \mathbf{w}_{V^\perp} \cdot \mathbf{x} + t))^2]$, (by adding all the convex combinations of $\Pi_V$). Moreover, we define $\psi_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}_\mathbf{h}}[\psi(\mathbf{z} + \mathbf{x}_V)]$. By adding and subtracting $\psi_V(\mathbf{x})$, we get the following

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f_V(\mathbf{x}))\psi(\mathbf{x})] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(\psi(\mathbf{x}) - \psi_V(\mathbf{x}))] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f_V(\mathbf{x}))\psi_V(\mathbf{x})] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(\psi(\mathbf{x}) - \psi_V(\mathbf{x}))] \,,\end{aligned}$$

where the last equality holds because the term $\psi_V(\mathbf{x})$ does not depend on the directions inside $V^\perp$. It remains to bound the term $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(\psi(\mathbf{x}) - \psi_V(\mathbf{x}))]$. From Cauchy-Schwarz inequality, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(\psi(\mathbf{x}) - \psi_V(\mathbf{x}))] \leq \sqrt{2M \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\psi(\mathbf{x}) - \psi_V(\mathbf{x}))^2]} \,.$$

From Lemma 9.4.14, we get that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(\psi(\mathbf{x}) - \psi_V(\mathbf{x}))] \leq \sqrt{2M \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\nabla\psi(\mathbf{x}) \cdot \mathbf{h})^2]} = \sqrt{2M\mathbf{h}^\intercal \mathbf{Inf}_\psi \mathbf{h}} \,.$$

Furthermore, using that $\|\widehat{\mathbf{M}} - \mathbf{Inf}_\psi\|_2 \leq \eta/2$, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(\psi(\mathbf{x}) - \psi_V(\mathbf{x}))] \leq \sqrt{2M}\sqrt{\eta/2 + \mathbf{h}^\intercal \widehat{\mathbf{M}}\mathbf{h}} \leq 2\sqrt{M\eta} \,.$$

where in the last inequality we used that $h$ lies in the $V^\perp$ and for any $\mathbf{v} \in V^\perp$, it holds

$\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{v}^\intercal\widehat{\mathbf{M}}\mathbf{v})^2] \leq \eta$. By choosing $\eta = \epsilon^2/(M32)$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - g(\mathbf{x}))\psi(\mathbf{x})] \leq 2\sqrt{M}\sqrt{\eta} \leq \epsilon/2 \ ,$$

which from Equation (9.3), we get a contradiction. An application of Lemma 9.3.14 completes the proof. □

We require the following standard fact showing the existence of a small $\epsilon$-cover of ReLU activations for of the set $V$.

**Fact 9.3.28** (see, e.g., Corollary 4.2.13 of [Ver18])**.** *For any $\epsilon, M > 0$, there exists an explicit $\epsilon$-cover $\mathcal{H}$ of ReLUs over the $\mathbb{R}^d$ over the $L_2$ norm with respect the Gaussian distribution of size* $\mathrm{poly}(M/\epsilon)^d$.

We are now ready to prove the main theorem of this section.

*Proof of Theorem 9.3.26.* We first show that there is a set $\mathcal{H}$ of size $(M/\epsilon)^{\mathrm{poly}(1/\epsilon)}$ which contains tuples $(\mathbf{u}, t)$ with $\mathbf{u} \in \mathbb{R}^d$ and $\|\mathbf{u}\|_2 \leq M$ and $t \in \mathbb{R}$, such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{ReLU}(\mathbf{u}\cdot\mathbf{x} + t) - y)^2] \leq \inf_{f\in\mathcal{C}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - y)^2] + \epsilon \ .$$

First note we can assume that $1/\epsilon^6 \leq d$, since otherwise one can directly do a brute-force search over an $\epsilon$-cover of the $d$-dimensional $M$-ball. The runtime to perform this brute-force search will be $(M/\epsilon)^{O(d)}\log(1/\delta)$ which, by the assumption that $\mathrm{poly}(1/\epsilon) > d$, is smaller than $(M/\epsilon)^{\mathrm{poly}(1/\epsilon)}\log(1/\delta)$.

Let $f \in \mathcal{C}$ be such that the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(f(\mathbf{x}) - y(\mathbf{x}))^2]$ is minimized.

Similar with the proof of Theorem 9.1.2, we can assume that $|y(\mathbf{x})| \leq M^{1/2}/\epsilon^{1/2}$ as this does not increase the error by a lot. Let $\psi(\mathbf{x}) = T_\rho y$ for $\rho = \mathrm{poly}(\epsilon/(M))$. Note

that $\|\nabla\psi(\mathbf{x})\|_2 \leq M'$. From Lemma 9.2.3, with $N = \mathrm{poly}(d/\epsilon)\log(1/\delta)$ queries, we get that with probability $1-\delta/2$ a matrix $\mathbf{M}$, so that $\|\mathbf{M} - \mathbf{Inf}_\psi\|_F \leq \epsilon$. Applying Proposition 9.3.27 to the matrix $\mathbf{M}$, we get that in the subspace $V$ spanned by the eigenvectors of the matrix $\mathbf{M}$ with eigenvalues larger than $\eta = \mathrm{poly}(\epsilon/M))$ with dimension at most $O(\mathrm{poly}(M', 1/\eta, 1/\epsilon))$, there exists a ReLU activation $h$ with vector lying in the subspace $V$ so that

$$\min_{t\in\mathbb{R}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\psi(\mathbf{x}) - \mathrm{ReLU}(\mathbf{v}\cdot\mathbf{x}+t))^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\psi(\mathbf{x}) - f(\mathbf{x}))^2] + \epsilon\ .$$

Applying Fact 9.3.28, we get that there exists an $\epsilon$-cover $\mathcal{H}$ of halfspaces in $V$ of size $(M/\epsilon)^{\mathrm{poly}(1/\epsilon)}$, so that for any $f(\mathbf{x}) = \mathrm{ReLU}(\mathbf{w}\cdot\mathbf{x}+t)$ with $\mathbf{w}\in V$ there exists $f'\in\mathcal{H}$ so that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f'(\mathbf{x}))^2] \leq \epsilon$. Hence, there exists $f'\in\mathcal{H}$ so that

$$\min_{f'\in\mathcal{H}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\psi(\mathbf{x}) - f'(\mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\psi(\mathbf{x}) - f(\mathbf{x}))^2] + 2\epsilon\ .$$

Furthermore, from Proposition 9.2.6 we have that it also holds

$$\min_{f'\in\mathcal{H}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y(\mathbf{x}) - f'(\mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y(\mathbf{x}) - f(\mathbf{x}))^2] + O(\epsilon)\ .$$

To complete the proof, we show that Step 6 of Algorithm 22 outputs the correct hypothesis. From ERM it follows that $O(\frac{M}{\epsilon^2}\log(\mathcal{H}/\delta))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most $\epsilon$ with probability at least $1-\delta/2$. To bound the runtime of the algorithm, we note that the exhaustive search over an $\epsilon$-cover takes time $(M/\epsilon)^{\mathrm{poly}(1/\epsilon)}\log(1/\delta)$. Thus, the total runtime of our algorithm

in the case where $\mathrm{poly}(1/\epsilon) \leq d$ is

$$\left(\mathrm{poly}(d/\epsilon) + d(1/\epsilon)^{\mathrm{poly}(1/\epsilon)}\right)\log(1/\delta)\,.$$

This completes the proof of Theorem 9.3.26. □

## 9.4 Agnostically Learning Boolean Multi-index Models

In this section, we present our results for Boolean multi-index models of bounded surface area. For convenience, we restate the class of concepts that we consider.

**Definition 9.4.1** (Bounded Surface Area, Low-Dimensional Boolean Concepts)**.** *We define the class $\mathfrak{B}(\Gamma, k)$ of Boolean concepts with the following properties:*

1. *For every $f \in \mathfrak{B}(\Gamma, k)$, it holds $\Gamma(f) \leq \Gamma$.*
2. *For every $f \in \mathfrak{B}(\Gamma, k)$, there exists a subspace $U$ of $\mathbb{R}^d$ of dimension at most $k$ such that $f$ depends only on $U$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$, $f(\mathbf{x}) = f(\mathrm{proj}_U \mathbf{x})$.*
3. *$\mathfrak{B}(\Gamma, k)$ is closed under translations, i.e., if $f(\mathbf{x}) \in \mathfrak{B}(\Gamma, k)$ then $f(\mathbf{x}+\mathbf{t}) \in \mathfrak{B}(\Gamma, k)$ for all $\mathbf{t} \in \mathbb{R}^d$.*

We remark that $\mathfrak{B}(\Gamma, k)$ is a general, *non-parametric class*. For example $\mathfrak{B}(\Omega(k), k)$ contains LTFs, intersections of $k$ LTFs, and Polynomial Threhsold Functions (PTFs) of degree at most $k$ (that depend on a $k$-dimensional subspace). Our learner is able to learn a hypothesis of low excess error when compared against all concepts of $\mathfrak{B}(\Gamma, k)$ with roughly $\mathrm{poly}(d/\epsilon) + k^{\mathrm{poly}(\Gamma/\epsilon)}$ runtime.

**Theorem 9.4.2.** *Fix $k \in \mathbb{N}$ and $M \in \mathbb{R}^+$. Let $D$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that the $\mathbf{x}$-marginal of $D$ is standard $d$-dimensional normal. There exists an algorithm that makes $N_q = \mathrm{poly}(d/\epsilon)$ queries and draws $N_s = \mathrm{poly}(d/\epsilon) + \mathrm{poly}((k\Gamma/\epsilon)^{\Gamma^2/\epsilon^4}, 1/\epsilon, \log(1/\delta))$ samples from $D$ and runs in time $\mathrm{poly}(N_s, N_q, d)$ and outputs a polynomial $p : \mathbb{R}^d \mapsto \mathbb{R}$ so that with probability at least $1 - \delta$ it holds*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(p(\mathbf{x})) \neq y] \leq \inf_{f \in \mathfrak{B}(\Gamma,k)} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] + \epsilon \,.$$

### 9.4.1 Influence PCA for Learning in $L_1$-norm

In this section we show our main dimension-reduction tool for the concepts of bounded surface of Definition 9.1.3. Our dimension-reduction tool establishes that: given (an approximation of) the influence matrix of the smooth function $T_\rho y$, we can perform PCA and learn a low-dimensional subspace $V$ so that a bounded surface area concept that depends only on $V$ can achieve $\epsilon$ excess error with respect to $T_\rho y$ in $L_1$-norm. We now state our result.

**Proposition 9.4.3** (Dimension Reduction)**.** *Fix $\epsilon > 0, k \in \mathbb{N}$ and let $\psi : \mathbb{R}^d \mapsto [-1, 1]$ be a differentiable function with $\|\nabla\psi(\mathbf{x})\|_2 \leq \Psi$ for all $\mathbf{x} \in \mathbb{R}^d$. Let $\eta$ be sufficiently small multiply of $\epsilon^2/k$ let $\widehat{\mathbf{M}} \in \mathbb{R}^{d\times d}$ be such that $\|\widehat{\mathbf{M}} - \mathbf{Inf}_\psi\|_2 \leq \eta/2$. Let $V$ be the subspace spanned by all the eigenvectors of $\widehat{\mathbf{M}}$ whose corresponding eigenvalues are at least $\eta$. The following hold true:*

1. *There exists $g$ with $\Gamma(g) \leq \Gamma$ so that $g(\mathbf{x}) = g(\mathrm{proj}_V \mathbf{x})$ such that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|g(\mathbf{x}) - \psi(\mathbf{x})|] \leq \inf_{f \in \mathfrak{B}(\Gamma,k)} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x}) - \psi(\mathbf{x})|] + \epsilon \,.$$

*2. The dimension of $V$ is at most $O(\Psi^2/\eta)$.*

Before we proceed to the proof of Proposition 9.4.3 we give some intuition behind the choice of the Gaussian Marginalization Operator defined above. We first give the following simple lemma showing that in order to show that a concept class $C_2$ (think of this as the class of concepts that depend on the subspace $V$ of Proposition 9.4.3) has not much worse approximation error (to some label $y$) than some other class $C_1$ (think of this as the original concept class $\mathfrak{B}(\Gamma, k)$) as long as for every concept $f$ of $C_1$, we can construct a distribution over concepts of $C_2$ that (on expectation) achieves at most $\epsilon$ worse correlation with the label $y$ than the original concept $f$. Its proof relies on the simple fact that for $t \in [-1, 1]$ and $s \in \{\pm 1\}$, it holds that $|t - s| = 1 - ts$.

**Lemma 9.4.4** (Correlating Convex Combinations)**.** *Fix a function $y : \mathbb{R}^d \mapsto [-1, 1]$ and $\epsilon > 0$. Let $C_1, C_2$ be classes of Boolean concepts on $\mathbb{R}^d$. Assume that for every $f \in C_1$ there exists a distribution $Q$ over hypotheses of the class $C_2$ such that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\mathop{\mathbf{E}}_{g\sim Q}[g(\mathbf{x})\ y(\mathbf{x})]\right] \leq \epsilon\,.$$

*Then* $\inf_{g\in C_2} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|g(\mathbf{x}) - y(\mathbf{x})|] - \inf_{f\in C_1} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x}) - y(\mathbf{x})|] \leq \epsilon$.

*Proof.* Fix $f \in C_1$. By assumption, we have that there exists a distribution $Q_f$ over $C_2$ so that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\mathop{\mathbf{E}}_{g\sim Q_f}[g(\mathbf{x})\ y(\mathbf{x})]\right] \leq \epsilon\,.$$

Note that $g \in \{\pm 1\}$ and $|y(\mathbf{x})| \leq 1$, therefore the expectation is bounded and hence from Fubini's theorem, we have that

$$\mathop{\mathbf{E}}_{g\sim Q_f}\left[\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[g(\mathbf{x})\ y(\mathbf{x})]\right] \leq \epsilon\,.$$

That means that there exists a $r_f \in C_2$ so that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[r_f(\mathbf{x})\ y(\mathbf{x})] \leq \epsilon\,.$$

Because $f(\mathbf{x}), r_f(\mathbf{x})$ are Boolean functions we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})y(\mathbf{x})] = 1 - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x})y(\mathbf{x})]|$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[r_f(\mathbf{x})y(\mathbf{x})] = 1 - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|r_f(\mathbf{x})y(\mathbf{x})]|$. Therefore, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|r_f(\mathbf{x}) - y(\mathbf{x})|] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x}) - y(\mathbf{x})|] \leq \epsilon\,.$$

Furthermore, because $r_f \in C_2$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|r_f(\mathbf{x}) - y(\mathbf{x})|] \geq \inf_{g\in C_2} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|g(\mathbf{x}) - y(\mathbf{x})|]$. The proof is completed by taking the supremium over all the $f \in C_1$. This completes the proof of Lemma 9.4.4. □

**Proof of Proposition 9.4.3**

We define that set of hypotheses $\mathfrak{B}_V(\Gamma, k) = \{f \in \mathfrak{B}(\Gamma, k) : f(\mathrm{proj}_V(\mathbf{x})) = f(\mathbf{x})\}$. We are going to show that

$$\inf_{g\in\mathfrak{B}_V(\Gamma,k)} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|g(\mathbf{x}) - \psi(\mathbf{x})|] \leq \inf_{f\in\mathfrak{B}(\Gamma,k)} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x}) - \psi(\mathbf{x})|] + \epsilon\,. \tag{9.4}$$

To prove Equation (9.4), by Lemma 9.4.4 it suffices to construct, for each $f \in \mathfrak{B}(\Gamma, k)$, a distribution $Q$ over the set $\mathfrak{B}_V(\Gamma, k)$ and show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})\psi(\mathbf{x})] - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}\left[\mathbf{E}_{g\sim Q}[g(\mathbf{x})\ \psi(\mathbf{x})]\right] \leq \epsilon$. To this end, we first show that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \Pi_V f(\mathbf{x}))\psi(\mathbf{x})] \leq \epsilon\,. \tag{9.5}$$

Note that $\Pi_V f(\mathbf{x})$ is a distribution over $\mathfrak{B}_V(\Gamma, k)$. To see that note that $\Pi_V f(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}_{V^\perp}}[f(\mathbf{x}_V + \mathbf{z})]$ and note that for each $\mathbf{z} \in \mathbb{R}^d$, we have that $f(\mathbf{x}_V + \mathbf{z}) \in \mathfrak{B}_V(\Gamma, k)$. To prove Equation (9.4), we prove the following lemma:

**Lemma 9.4.5** (Correlation via Gaussian Marginalization)**.** *Let $y : \mathbb{R}^d \mapsto \mathbb{R}$ be some function in $L^2(\mathcal{N})$. Fix some concept $f \in \mathfrak{B}(\Gamma, k)$ and denote by $U$ the subspace of $\mathbb{R}^d$ that $f$ depends on (i.e., $f(\mathbf{x}) = f(\mathrm{proj}_U(\mathbf{x}))$. Moreover, let $V$ be some other subspace of $\mathbb{R}^d$. Then it holds*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \Pi_V f(\mathbf{x}))y(\mathbf{x})] \leq 2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{U+V}y(\mathbf{x}) - \Pi_V\Pi_{U+V}y(\mathbf{x}))^2]}\,.$$

*Proof.* Note that by our assumption $f(\mathbf{x}) = f(\mathbf{x}_U) = \Pi_U f(\mathbf{x})$, therefore, $\Pi_V\Pi_U f(\mathbf{x}) = \Pi_V f(\mathbf{x})$. Observe that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \Pi_V f(\mathbf{x}))\Pi_V\Pi_{U+V}y(\mathbf{x})] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_V\Pi_U f(\mathbf{x}) - \Pi_V f(\mathbf{x}))\Pi_V\Pi_{U+V}y(\mathbf{x})] = 0$, which gives that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \Pi_V f(\mathbf{x}))y(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \Pi_V f(\mathbf{x}))(\Pi_{U+V}y(\mathbf{x}) - \Pi_V\Pi_{U+V}y(\mathbf{x}))]\,.$$

Using that $f$ is a Boolean function, we have that $|\Pi_V f(\mathbf{x})| \leq 1$, hence, $|f(\mathbf{x}) - \Pi_V f(\mathbf{x})| \leq 2$, which gives

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \Pi_V f(\mathbf{x}))y(\mathbf{x})] &\leq 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|\Pi_{U+V}y(\mathbf{x}) - \Pi_V\Pi_{U+V}y(\mathbf{x})|] \\ &\leq 2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{U+V}y(\mathbf{x}) - \Pi_V\Pi_{U+V}y(\mathbf{x}))^2]}\,,\end{aligned}$$

where we used Cauchy-Schwarz inequality. This completes the proof of Lemma 9.4.5. □

It remains to bound the term $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{U+V}\psi(\mathbf{x}) - \Pi_V\Pi_{U+V}\psi(\mathbf{x}))^2]$. Observe that

$\dim(U \cap V^\perp) \leq \dim(U) \leq k$, by applying Lemma 9.3.13 we get that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\Pi_{U+V}\psi(\mathbf{x}) - \Pi_V\Pi_{U+V}\psi(\mathbf{x}))^2] &\leq \dim(U \cap V^\perp) \max_{\mathbf{v}\in U\cap V^\perp, \|\mathbf{v}\|_2=1} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\nabla\psi(\mathbf{x})\cdot\mathbf{v})^2] \\ &= k \max_{\mathbf{v}\in U\cap V^\perp, \|\mathbf{v}\|_2=1} \mathbf{v}^\intercal \mathbf{Inf}_\psi \mathbf{v} \,.\end{aligned}$$

Furthermore, using that $\|\widehat{\mathbf{M}} - \mathbf{Inf}_\psi\|_2 \leq \eta/2$, we have that

$$\max_{\mathbf{v}\in U\cap V^\perp, \|\mathbf{v}\|_2=1} \mathbf{v}^\intercal \mathbf{Inf}_\psi \mathbf{v} \leq \eta/2 + \max_{\mathbf{v}\in U\cap V^\perp, \|\mathbf{v}\|_2=1} \mathbf{v}^\intercal \widehat{\mathbf{M}} \mathbf{v} \leq 2\eta \,,$$

where in the last inequality we used that $\mathbf{v}$ lies in the $V^\perp$ and for any $\mathbf{v} \in V^\perp$, it holds $\mathbf{v}^\intercal\widehat{\mathbf{M}}\mathbf{v} \leq \eta$. By choosing $\eta = \epsilon^2/(32k)$, we have shown that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - \Pi_V f(\mathbf{x}))\psi(\mathbf{x})] \leq \epsilon \,.$$

The proof then follows from Lemma 9.4.4. It remains to bound the dimension of the subspace $V$. To this end, we use Lemma 9.3.14 which gives that $\dim(V) = O(\Psi^2/\eta^2)$. This completes the proof of Proposition 9.4.3.

### 9.4.2 Proof of Theorem 9.4.2

For learning geometric concepts, we use the standard $L_1$-regression algorithm from [KKMS08].

**Fact 9.4.6** (Theorem 9 [KOS08])**.** *Let $\mathcal{C}$ be a class of Boolean functions in $\mathbb{R}^d$. Let $D$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that the $\mathbf{x}$-marginal of $D$ is the standard $d$-dimensional normal. The $L_1$-regression algorithm draws $N = \mathrm{poly}(d^{\Gamma(\mathcal{C})^2/\epsilon^4}, 1/\epsilon, \log(1/\delta))$ samples from $D$ and runs in time $\mathrm{poly}(N, d)$ and outputs a polynomial $p : \mathbb{R}^d \mapsto \mathbb{R}$ so that with*

*probability at least* $1-\delta$ *it holds*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(p(\mathbf{x}))\neq y]\leq \min_{f\in\mathcal{C}}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f(\mathbf{x})\neq y]+\epsilon\,.$$

Let $\psi(\mathbf{x})=T_\rho y(\mathbf{x})$ with $\rho$ be less than a sufficiently small constant multiple of $\epsilon/\Gamma(f)$. Using $\mathrm{poly}(d,1/\epsilon,\log(1/\delta)$ queries, we calculate an estimation $\mathbf{M}$ of the influence matrix $\mathbf{Inf}_\psi$ (Lemma 9.2.3). Let $V$ be the subspace spanned with the eigenvectors of $\mathbf{M}$ with eigenvalue at least $\eta$, where $\eta$ is a sufficiently small constant multiply of $\epsilon^2/k$. Using Proposition 9.4.3, we have that $V$ has dimension at most $O(\Gamma^2k/\epsilon^4)$ and furthermore there exists $g$ with $\Gamma(g)\leq\Gamma$ so that $g(\mathbf{x})=g(\mathrm{proj}_V\mathbf{x})$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|g(\mathbf{x})-\psi(\mathbf{x})|]\leq\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x})-\psi(\mathbf{x})|]+\epsilon\,.$$

From Proposition 9.2.10, we have that it also holds

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|g(\mathbf{x})-y(\mathbf{x})|]\leq\inf_{f\in\mathfrak{B}(\Gamma,k)}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x})-y(\mathbf{x})|]+O(\epsilon)\,.$$

Equivalently, we have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[g(\mathbf{x})\neq y(\mathbf{x})]\leq\inf_{f\in\mathfrak{B}(\Gamma,k)}\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})\neq y(\mathbf{x})]+O(\epsilon)$. Let $\mathbf{P}:\mathbb{R}^d\mapsto V$ be the projection matrix to the subspace $V$. Let $(\mathbf{Px},y)\sim D'$, where $(\mathbf{x},y)\sim D$. We the $L_1$-regression algorithm on $D'$ and from Fact 9.4.6, using $\mathrm{poly}((k\Gamma/\epsilon)^{\Gamma^2/\epsilon^4},1/\epsilon,\log(1/\delta))$ samples from $D'$, we get a polynomial $p:V\mapsto\mathbb{R}$ so that with probability at least $1-\delta$, it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(p(\mathbf{Px}))\neq y]\leq\inf_{f\in\mathfrak{B}(\Gamma,k)}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f(\mathbf{x})\neq y]+\epsilon\,.$$

This completes the proof of Theorem 9.4.2.

### 9.4.3 Corollaries for Intersections of Halfspaces and PTFs

Using Theorem 9.4.2, we can show the following corollary for intersections of $k$ halfspaces:

**Corollary 9.4.7.** *Let $\mathcal{C}$ be the class of intersections $k$ halfspaces in $\mathbb{R}^d$. Let $D$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that the $\mathbf{x}$-marginal of $D$ is the standard $d$-dimensional normal. There exists an algorithm that makes $N_q = \mathrm{poly}(d/\epsilon)$ queries and draws $N_s = \mathrm{poly}(d/\epsilon) + \mathrm{poly}((k/\epsilon)^{\log(k)/\epsilon^4}, 1/\epsilon, \log(1/\delta))$ samples from $D$ and runs in time $\mathrm{poly}(N_s, N_q, d)$ and outputs a polynomial $p : \mathbb{R}^d \mapsto \mathbb{R}$ so that with probability at least $1 - \delta$ it holds*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(p(\mathbf{x})) \neq y] \leq \min_{f \in \mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] + \epsilon \, .$$

*Proof of Corollary 9.4.7.* For the proof, we need the following fact about the Gaussian surface area of the intersection of $k$ halfspaces.

**Fact 9.4.8** (Theorem 20 of [KOS08])**.** *The surface area $\Gamma(f)$ of the intersection of $k$ halfspaces is at most $O(\sqrt{\log k})$.*

The proof follows from Theorem 9.4.2 and Fact 9.4.8. □

We show that we can use Theorem 9.4.2 to learn low-degree polynomial threshold functions (PTFs) that depend only on a small dimensional subspace.

**Corollary 9.4.9.** *Let $\mathcal{C}$ be the class of degree-$\ell$ PTFs in $\mathbb{R}^d$ that depend on an unknown $k$-dimensional subspace. Let $D$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that the $\mathbf{x}$-marginal of $D$ is the standard $d$-dimensional normal. There exists an algorithm that makes $N_q = \mathrm{poly}(d/\epsilon)$ queries, draws $N_s = \mathrm{poly}(d/\epsilon) + \mathrm{poly}((k/\epsilon)^{\ell/\epsilon^4}, 1/\epsilon, \log(1/\delta))$ samples*

*from $D$, runs in time* $\mathrm{poly}(N_s, N_q, d)$ *and outputs a polynomial* $p : \mathbb{R}^d \mapsto \mathbb{R}$ *so that with probability at least* $1 - \delta$ *it holds*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(p(\mathbf{x})) \neq y] \leq \min_{f\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] + \epsilon \; .$$

*Proof of Corollary 9.4.9.* For the proof, we need the following fact about the Gaussian surface area of degree-$\ell$ PTFs.

**Fact 9.4.10** (Gaussian Surface Area of PTFs, [Kan11])**.** *The surface area $\Gamma(f)$ of $\ell$-degree polynomial threshold functions is at most $O(\ell)$.*

The proof follows from Theorem 9.4.2 and Fact 9.4.10. □

Finally, we show that we can use Theorem 9.4.2 to learn arbitrary functions of $\ell$ halfspaces.

**Corollary 9.4.11.** *Let $\mathcal{C}$ be the class of functions of $\ell$ halfspaces in $\mathbb{R}^d$. Let $D$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that the $\mathbf{x}$-marginal of $D$ is the standard $d$-dimensional normal. There exists an algorithm that makes $N_q = \mathrm{poly}(d/\epsilon)$ queries, draws $N_s = \mathrm{poly}(d/\epsilon) + \mathrm{poly}((\ell/\epsilon)^{\ell/\epsilon^4}, 1/\epsilon, \log(1/\delta))$ samples from $D$, runs in time $\mathrm{poly}(N_s, N_q, d)$ and outputs a polynomial $p : \mathbb{R}^d \mapsto \mathbb{R}$ so that with probability at least $1 - \delta$ it holds*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\mathrm{sign}(p(\mathbf{x})) \neq y] \leq \min_{f\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] + \epsilon \; .$$

*Proof of Corollary 9.4.11.* We note that the Gaussian surface area of functions of $\ell$ halfspaces is bounded above by $\ell$. From [KOS08] (see, e.g., Fact 17), we have that the surface area of a Boolean function $f$ that depends on $\ell$ halfspaces, is bounded above

by the sum of the surface area of the individual halfspaces; therefore, we have that $\Gamma(f) \leq O(\ell)$. The proof follows from Theorem 9.4.2. □

### 9.4.4 Proper Agnostic Query Learner for LTFs

In this section we present our algorithmic result for agnostic proper learning of linear threshold functions with membership queries. Our goal is to show Theorem 9.4.12 which we state below.

**Theorem 9.4.12** (Proper Agnostic Query Learner for LTFs)**.** *Let $\mathcal{C}$ be the class of LTFs on $\mathbb{R}^d$ and denote by $y(\mathbf{x}) \in \{\pm 1\}$ the label (chosen by an adversary) of $\mathbf{x} \in \mathbb{R}^d$. There exists an algorithm that makes $N_s = \mathrm{poly}(d/\epsilon)$ sample queries, $N_q = \mathrm{poly}(d/\epsilon)$ queries, and, with runtime $\mathrm{poly}(d/\epsilon) + 2^{\mathrm{poly}(1/\epsilon)}$, computes an LTF $f(\mathbf{x}) : \mathbb{R}^d \mapsto \{\pm 1\}$ such that, with probability at least $1 - \delta$, it holds $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[h(\mathbf{x}) \neq y(\mathbf{x})] \leq \inf_{c\in\mathcal{C}} \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[c(\mathbf{x}) \neq y(\mathbf{x})] + \epsilon$.*

Our algorithm is presented in Algorithm 23. It uses membership queries to estimate a matrix $M$ corresponding to the influence matrix of the appropriately smoothed label function $y(\mathbf{x})$. It then restricts attention to a small subspace $V$ given by the large eigenvectors of $\mathbf{M}$ and exhaustively searches for a near-optimal halfspace with a normal vector in $V$.

### 9.4.5 Reducing the Dimension via Influence PCA

In this section we show our main dimension-reduction tool. We prove that the subpspace corresponding to eigvenvectors with large (i.e., larger than $\mathrm{poly}(\epsilon)$) eigenvalues contains the normal-vector of an approximately optimal halfspace.

**Algorithm 23** Agnostic Proper Learning of LTFs with Membership Queries

1: **Input:** $\epsilon > 0$, $\delta > 0$; sample and query access to distribution $\mathcal{D}$.
2: **Output:** Hypothesis $h \in \mathcal{C}$ such that $\mathrm{err}^{\mathcal{D}}_{0\text{-}1}(h) \leq \min_{f \in \mathcal{C}} \mathrm{err}^{\mathcal{D}}_{0\text{-}1}(f) + \epsilon$ with probability $1 - \delta$.
3: $\rho \leftarrow C\,\epsilon^2, \quad \eta \leftarrow C\,\epsilon^2$, for a sufficiently small constant $C > 0$.
4: Estimate $\mathbf{M} = \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} \left[ D_\rho y(\mathbf{x})\, D_\rho y(\mathbf{x})^\intercal \right]$ using $\mathrm{poly}(d/\epsilon)$ queries via Algorithm 21.
5: Let $V$ be the subspace spanned by those eigenvectors of $\mathbf{M}$ whose eigenvalues are at least $\eta$.
6: Let $\mathcal{H}_V$ be the set of LTFs with normal vectors in $V$. Compute the ERM hypothesis

$$h \;=\; \arg\min_{g \in \mathcal{H}_V} \mathrm{err}^{\mathcal{D}}_{0\text{-}1}(g)$$

using $m = \Theta\Big(\frac{\dim(V)}{\epsilon^2} \log \frac{1}{\delta}\Big)$ i.i.d. samples from $\mathcal{D}$ in time $O\Big(m^{\dim(V)}\Big)$.
7: **Return:** $h$.

**Proposition 9.4.13** (Dimension Reduction via Influence PCA: LTFs)**.** *Fix $\epsilon > 0$ and let $\psi : \mathbb{R}^d \mapsto [-1, 1]$ be a differentiable function with $\|\nabla \psi(\mathbf{x})\|_2 \leq \Psi$ for all $\mathbf{x} \in \mathbb{R}^d$. Let $\eta$ be a sufficiently small multiple of $\epsilon^2$ and let $\widehat{\mathbf{M}} \in \mathbb{R}^{d \times d}$ be such that $\|\widehat{\mathbf{M}} - \mathbf{Inf}_\psi\|_2 \leq \eta/2$. Moreover, let $V$ be the subspace spanned by all the eigenvectors of $\widehat{\mathbf{M}}$ whose corresponding eigenvalues are at least $\eta$. The following hold true:*

1. *There exists $\mathbf{v} \in V$ and $t \in \mathbb{R}$ such that*

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[|\mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t) - \psi(\mathbf{x})|] \leq \inf_{\mathbf{w} \in \mathbb{R}^d, t \in \mathbb{R}} \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[|\mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t) - \psi(\mathbf{x})|] + \epsilon\,.$$

2. *The dimension of $V$ is at most $O(\Psi^2/\eta)$.*

*Proof.* Suppose for the sake of contradiction that there exists a halfspace $f \in \mathcal{C}$ such that for every halfspace $g \in \mathcal{C}_V$, it holds

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[(f(\mathbf{x}) - g(\mathbf{x}))\psi(\mathbf{x})] \geq \epsilon\,. \tag{9.6}$$

We can write $f(\mathbf{x}) = \text{sign}(\mathbf{w}\cdot\mathbf{x}+t) = \text{sign}(\mathbf{w}_V\cdot\mathbf{x}+\mathbf{w}_{V^\perp}\cdot\mathbf{x}+t)$. Note that $\mathbf{w}_{V^\perp} \neq \mathbf{0}$, since otherwise we would have $f \in \mathcal{C}_V$. For simplicity, we denote $\mathbf{h} = \mathbf{w}_{V^\perp}/\|\mathbf{w}_{V^\perp}\|_2$. Notice that the direction $\mathbf{h}$ has low influence, since $\mathbf{h} \in V^\perp$. Recall that by $\mathcal{N}_\mathbf{h}$ we denote the projection of $\mathcal{N}$ onto the (one-dimensional) subspace spanned by $\mathbf{h}$. We define $f_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}_\mathbf{h}}[f(\mathbf{z}+\mathbf{x}_V)]$ to be a convex combination of halfspaces in $\mathcal{C}_V$. In particular, $f_V(\mathbf{x})$ is a smoothed version of the halfspace $\text{sign}(\mathbf{w}_V\cdot\mathbf{x}+t)$ whose normal vector belongs in $V$. Moreover, we define $\psi_V(\mathbf{x}) = \mathbf{E}_{\mathbf{z}\sim\mathcal{N}_\mathbf{h}}[\psi(\mathbf{z}+\mathbf{x}_V)]$. By adding and subtracting $\psi_V(\mathbf{x})$, we get the following

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x})-f_V(\mathbf{x}))\psi(\mathbf{x})] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x})-f_V(\mathbf{x}))(\psi(\mathbf{x})-\psi_V(\mathbf{x}))] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x})-f_V(\mathbf{x}))\psi_V(\mathbf{x})] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x})-f_V(\mathbf{x}))(\psi(\mathbf{x})-\psi_V(\mathbf{x}))]\,,\end{aligned}$$

where the last equality holds because the term $\psi_V(\mathbf{x})$ does not depend on the direction $h$, therefore,

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x})-f_V(\mathbf{x}))\psi_V(\mathbf{x})] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_{\mathbf{h}^\perp}}\left[\mathop{\mathbf{E}}_{\mathbf{z}\sim\mathcal{N}_\mathbf{h}}[(f(\mathbf{x}+\mathbf{z})-f_V(\mathbf{x}))\psi_V(\mathbf{x})]\right] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_{\mathbf{h}^\perp}}[(f_V(\mathbf{x})-f_V(\mathbf{x}))\psi_V(\mathbf{x})] = 0\,.\end{aligned}$$

It remains to bound the term $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x})-f_V(\mathbf{x}))(\psi(\mathbf{x})-\psi_V(\mathbf{x}))]$. From Cauchy-Schwarz inequality, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x})-f_V(\mathbf{x}))(\psi(\mathbf{x})-\psi_V(\mathbf{x}))] \leq \sqrt{2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\psi(\mathbf{x})-\psi_V(\mathbf{x}))^2]}\,.$$

First, we show the following to bound this term

**Lemma 9.4.14** (Gaussian Marginalization Error)**.** *Let $g : \mathbb{R}^d \mapsto \mathbb{R}$ be a function*

*in* $L^2(\mathcal{N})$ *such that* $\nabla g \in L^2(\mathcal{N})$. *Fix some direction* $\mathbf{v}$ *with* $\|\mathbf{v}\|_2 = 1$ *and define* $r(\mathbf{x}) = \mathbf{E}_{z\sim\mathcal{N}}[g(\mathrm{proj}_{\mathbf{v}^\perp}(\mathbf{x}) + z\mathbf{v})]$. *Then it holds*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(g(\mathbf{x}) - r(\mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\nabla g(\mathbf{x})\cdot\mathbf{v})^2]\,.$$

*Proof.* Using the rotation invariance of the Gaussian distribution, without loss of generality, we may assume that $\mathbf{v} = \mathbf{e}_1$. We have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(g(\mathbf{x})-r(\mathbf{x}))^2] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(g(\mathbf{x})-\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[r(z,\mathbf{x}_2,\ldots,\mathbf{x}_d)])^2] = \mathop{\mathbf{E}}_{\mathbf{x}_2,\ldots,\mathbf{x}_d}[\mathbf{Var}_{\mathbf{x}_1\sim\mathcal{N}}[g(\mathbf{x}_1,\ldots,\mathbf{x}_d)]]$$

To bound the variance in the above expression, we can use Poincare's inequality.

**Fact 9.4.15** (Gaussian Poincare Inequality)**.** *Let* $g : \mathbb{R} \mapsto \mathbb{R}$ *be a function in* $L^2(\mathcal{N})$ *such that* $g' \in L^2(\mathcal{N})$. *Then it holds* $\mathbf{Var}_{z\sim\mathcal{N}}[g(z)] \leq \mathbf{E}_{z\sim\mathcal{N}}[|g'(z)|^2]\,.$

Using Fact 9.4.15, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(g(\mathbf{x}) - r(\mathbf{x}))^2] \leq \mathbf{E}_{\mathbf{x}_2,\ldots,\mathbf{x}_d\sim\mathcal{N}^{d-1}}[\mathbf{E}_{\mathbf{x}_1\sim\mathcal{N}}[(\nabla g(\mathbf{x})\cdot \mathbf{e}_1)^2] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[(\nabla g(\mathbf{x})\cdot\mathbf{e}_1)^2]$. □

Therefore, from Lemma 9.4.14, we get that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(\psi(\mathbf{x}) - \psi_V(\mathbf{x}))] \leq 2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(\nabla\psi(\mathbf{x})\cdot\mathbf{h})^2]} = 2\sqrt{\mathbf{h}^\intercal\mathbf{Inf}_\psi\mathbf{h}}\,.$$

Furthermore, using that $\|\widehat{\mathbf{M}} - \mathbf{Inf}_\psi\|_2 \leq \eta/2$, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - f_V(\mathbf{x}))(\psi(\mathbf{x}) - \psi_V(\mathbf{x}))] \leq 2\sqrt{\eta/2 + \mathbf{h}^\intercal\widehat{\mathbf{M}}\mathbf{h}} \leq 4\sqrt{\eta}\,.$$

where in the last inequality we used that $h$ lies in the $V^\perp$ and for any $\mathbf{v} \in V^\perp$, it holds

$\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(\mathbf{v}^\intercal\widehat{\mathbf{M}}\mathbf{v})^2] \leq \eta$. By choosing $\eta = \epsilon^2/32$, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(f(\mathbf{x}) - g(\mathbf{x}))\psi(\mathbf{x})] \leq 2\sqrt{\eta} \leq \epsilon/2 \ ,$$

which from Equation (9.6), we get a contradiction. An application of Lemma 9.3.14 completes the proof of Proposition 9.4.13. □

We require the following standard fact showing the existence of a small $\epsilon$-cover of halfspaces for of the set $V$.

**Fact 9.4.16** (see, e.g., Corollary 4.2.13 of [Ver18])**.** *For any $\epsilon > 0$, there exists an explicit $\epsilon$-cover $\mathcal{H}$ of halfspaces over the $\mathbb{R}^d$ over the $L_1$ norm with respect the Gaussian distribution of size* $\mathrm{poly}(1/\epsilon)^d$*, i.e., for any $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} + t)$ there exists $f' \in \mathcal{H}$ so that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x}) - f'(\mathbf{x})|] \leq \epsilon$.*

We are now ready to prove the main theorem of this section.

*Proof of Theorem 9.4.12.* We first show that there is a set $\mathcal{H}$ of size $(1/\epsilon)^{O(1/\epsilon^6)}$ which contains tuples $(\mathbf{u}, t)$ with $\mathbf{u} \in \mathbb{R}^d$ and $t \in \mathbb{R}$, such that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{u}\cdot\mathbf{x} + t) \neq y] \leq \inf_{f\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y] + \epsilon \ .$$

First note we can assume that $1/\epsilon^6 \leq d$, since otherwise one can directly do a brute-force search over an $\epsilon$-cover of the $d$-dimensional unit ball: we do not need to perform our dimension-reduction process. The runtime to perform this brute-force search will be $(1/\epsilon)^{O(d)}\log(1/\delta)$ which, by the assumption that $1/\epsilon^6 > d$, is smaller than $(1/\epsilon)^{O(1/\epsilon^6)}\log(1/\delta)$.

Let $f \in \mathcal{C}$ be such that the $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[|f(\mathbf{x}) - y(\mathbf{x})|]$ is minimized. Let $\psi(\mathbf{x}) = T_\rho y(\mathbf{x})$ for $\rho = \Theta(\epsilon^2)$. Note that from Fact 9.2.2, we have that $\|\nabla\psi(\mathbf{x})\|_2 \leq 1/\rho$. From Lemma 9.2.3 for $N = \text{poly}(d/\epsilon)\log(1/\delta)$ queries, we get with probability at least $1-\delta/2$, a matrix $\mathbf{M}$ so that $\|\mathbf{M} - \mathbf{Inf}_\psi\|_F \leq \epsilon^2$. Applying Proposition 9.4.13 to the matrix $\mathbf{M}$, we get that the subspace $V$ spanned by the eigenvectors of the matrix $\mathbf{M}$ with eigenvalues larger than $\eta = \Theta(1/\epsilon^2)$ contains a vector $\mathbf{v} \in V$, such that

$$\min_{t\in\mathbb{R}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\psi(\mathbf{x}) - \text{sign}(\mathbf{v}\cdot\mathbf{x} + t)|] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\psi(\mathbf{x}) - f(\mathbf{x})|] + \epsilon\ .$$

Moreover, by Proposition 9.4.13, the dimension of $V$ is $O(1/\epsilon^6)$. Applying Fact 9.4.16, we get that there exists an $\epsilon$-cover $\mathcal{H}$ of halfspaces in $V$ of size $(1/\epsilon)^{O(1/\epsilon^6)}$, so that for any $f(\mathbf{x}) = \text{sign}(\mathbf{w}\cdot\mathbf{x} + t)$ with $\mathbf{w} \in V$ there exists $f' \in \mathcal{H}$ so that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[|f(\mathbf{x}) - f'(\mathbf{x})|] \leq \epsilon$. Hence, there exists $f' \in \mathcal{H}$ so that

$$\min_{f'\in\mathcal{H}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\psi(\mathbf{x}) - f'(\mathbf{x})|] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|\psi(\mathbf{x}) - f(\mathbf{x})|] + 2\epsilon\ .$$

Furthermore, from Proposition 9.2.10 and the fact that for halfspaces, $\Gamma(f) = O(1)$ (see, e.g., [KOS08]), we have that it also holds

$$\min_{f'\in\mathcal{H}} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|y(\mathbf{x}) - f'(\mathbf{x})|] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[|y(\mathbf{x}) - f(\mathbf{x})|] + O(\epsilon)\ .$$

To complete the proof, we show that Step 6 of Algorithm 23 outputs the correct hypothesis. From ERM for halfspaces, it follows that $O(\frac{1}{\epsilon^2}\log(\mathcal{H}/\delta))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most $\epsilon$ with probability at least $1 - \delta/2$. To bound the runtime of the algorithm, we note that

the exhaustive search over an $\epsilon$-cover takes time $(1/\epsilon)^{O(1/\epsilon^6)} \log(1/\delta)$. Thus, the total runtime of our algorithm in the case where $1/\epsilon^6 \leq d$ is

$$\left(\mathrm{poly}(d/\epsilon) + (1/\epsilon)^{O(1/\epsilon^6)}\right) \log(1/\delta)\,.$$

This completes the proof of Theorem 9.4.12. □

## 9.5 Hardness of Agnostic Proper Learning of Halfspaces and ReLUs with Queries

One might ask if the exponential dependence on $1/\epsilon$ in our upper bound (Corollaries 1.3.36 and 1.3.44) is necessary or just an artifact of our algorithmic approach. In this section, we provide some evidence that it is inherent. Unfortunately, there are very few circumstances where one can prove computational lower bounds against improper learners with query access to the function. So our bounds will apply only to proper learners. The basic idea of our argument is that if $f(\mathbf{x}) = \mathrm{sign}(\mathbf{v} \cdot \mathbf{x})$ is a linear threshold function or $f(\mathbf{x}) = \mathrm{ReLU}(\mathbf{v} \cdot \mathbf{x})$ with $\mathbf{v}$ a unit vector and $p(\mathbf{x})$ a polynomial, then $\mathbf{E}[f(\mathbf{x})p(\mathbf{x})]$ will be a polynomial in $\mathbf{v}$. As approximately optimizing low-degree polynomials over the unit sphere is conjectured to be computationally hard, this will prove hardness for proper learning of linear threshold functions. In particular, our hardness reduction starts from the small-set expansion problem [RS10]. We then rely on results of [BBH$^+$12] to reduce this problem to one about polynomial optimization. In particular we have:

**Theorem 9.5.1.** *If there is a polynomial-time algorithm that given* $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^d$

*outputs a constant factor approximation to* $\max_{\|\mathbf{x}\|_2=1} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a}_i \cdot \mathbf{x})^4$*, then there is a polynomial time algorithm for the small-set expansion problem.*

We note here that $\max_{\|\mathbf{x}\|_2=1} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a}_i \cdot \mathbf{x})^4$ is a homogeneous degree-4 polynomial. It will be important for our purposes that the polynomial in question have odd degree. Fortunately, we can reduce to this case.

**Corollary 9.5.2.** *If there is a polynomial-time algorithm that given a homogeneous degree-5 polynomial* $p$ *on* $\mathbb{R}^d$ *outputs a constant factor approximation to* $\max_{\|\mathbf{x}\|_2=1} p(\mathbf{x})$*, then there is a polynomial-time algorithm for the small-set expansion problem.*

*Proof.* We give a reduction to this problem from the problem in Theorem 9.5.1. In particular, given $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, we let $q(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{a}_i \cdot \mathbf{x})^4$. We then define the homogeneous degree-5 polynomial $p$ on $\mathbb{R}^{d+1}$ as $p(\mathbf{x}, y) = q(\mathbf{x})y$ (where $x$ here represents the first $d$ coordinates of the input and $y$ represents the last one). We note that if $\|(\mathbf{x}, y)\|_2 = 1$, then $\|\mathbf{x}\|_2 = a$ and $y = b$ for some $a^2 + b^2 = 1$. Letting $\mathbf{x}' = \mathbf{x}/a$ and using the homogeneity of $q$, we have that $p(\mathbf{x}, y) = a^4 b q(\mathbf{x}')$. For fixed $\mathbf{x}'$, the maximum of this over $a, b$ is obtained when $a = \sqrt{4/5}$ and $b = \sqrt{1/5}$. Thus, the maximum value of $p(\mathbf{x}, y)$ over the unit sphere equals the maximum value of $q(\mathbf{x}')$ over the unit sphere times $16/5^{2.5}$. Thus, finding a constant-factor approximation to the maximum value of one is equivalent to finding such an approximation of the other. This completes our proof. □

We are now ready to state our main theorem.

**Theorem 9.5.3** (Hardness of Proper Learning for LTFs)**.** *Suppose that there is an algorithm that given query access to a Boolean function* $f$ *on* $\mathbb{R}^d$ *runs in* $\mathrm{poly}(d)$ *time and approximates the minimum misclassification error between* $f$ *and a homogeneous*

*LTF (with respect to the standard Gaussian distribution) to additive error $\epsilon$ for some $\epsilon < d^{-10}$. Then there is a polynomial-time algorithm for the small set expansion problem.*

Before we prove Theorem 9.5.3, we note that any proper agnostic learner can be used to approximate this error merely by approximating the error between $f$ and the learned function. Thus, this result will imply a lower bound for learning.

*Proof.* We assume throughout that $d$ is sufficiently large, as otherwise there is nothing to prove. We proceed by a reduction from the problem in Corollary 9.5.2. In particular, let $p$ be a homogeneous degree-5 polynomial on $\mathbb{R}^d$. Let $\mathbf{T}$ be the unique symmetric tensor so that $p(\mathbf{x}) = \mathbf{T}(\mathbf{x}, \mathbf{x}, \mathbf{x}, \mathbf{x}, \mathbf{x})$. By scaling $\mathbf{T}$, we may assume that $\|\mathbf{T}\|_2 = 1$. Let $q(\mathbf{x}) = (\mathbf{T} \cdot H(\mathbf{x}))$, where $H(\mathbf{x})$ is the tensor whose entries are the degree-5 Hermite polynomials in $\mathbf{x}$.

Morally, we would like to take $f(\mathbf{x}) = q(\mathbf{x})$. Unfortunately, this does not work for two reasons.

First, $f(\mathbf{x})$ needs to be Boolean, while $q(\mathbf{x})$ distinctly is not. We can fix this by taking $f$ to be a random function, where the expected value of $f(\mathbf{x})$ equals $q(\mathbf{x})$.

Unfortunately, this cannot work because the expected value of $f(\mathbf{x})$ must still be in $[-1, 1]$, while $q$ is unbounded. To solve this, we first scale $q$ down substantially and then truncate its extreme values. To do this, we define:

$$t(x) = \begin{cases} 1 & \text{if } x > 1 \\ -1 & \text{if } x < -1 \\ x & \text{otherwise.} \end{cases}$$

We then divide $\mathbb{R}^d$ into tiny boxes of side length $\delta$ for some very small $\delta$. For each box $B$,

we pick an $\mathbf{x} \in B$ and then (independently for each box) let $f$ be 1 on $B$ with probability $(t(q(\mathbf{x})/d)+1)/2$ and $-1$ on $B$ otherwise. We note that the expected value of $f$ on $B$ is $t(q(\mathbf{x})/d)$, where $\mathbf{x}$ is the representative element. As the difference between $q$ at the representative element $\mathbf{x}$ of $B$ and at any other point in $B$ will be small if $\delta$ is (and if the box is not too far from the origin), it is not hard to see that the expectation over the randomness in defining $f$ of $|\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})\text{sign}(\mathbf{v}\cdot\mathbf{x})] - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[t(q(\mathbf{x})/d)\text{sign}(\mathbf{v}\cdot\mathbf{x})]|$ goes to 0 with $\delta$. As the variance of $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})\text{sign}(\mathbf{v}\cdot\mathbf{x})]$ also goes to 0 with $\delta$, if we take $\delta$ sufficiently small, then with high probability over the randomness in $f$, we have that $|\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})\text{sign}(\mathbf{v}\cdot\mathbf{x})] - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[t(q(\mathbf{x})/d)\text{sign}(\mathbf{v}\cdot\mathbf{x})]| < \epsilon/2$ for all unit vectors $\mathbf{v}$. Therefore, finding an $\epsilon$ additive approximation to the minimum misclassification error between $f$ and an LTF is equivalent to finding a $2\epsilon$-additive approximation to the maximum value of $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[f(\mathbf{x})\text{sign}(\mathbf{v}\cdot\mathbf{x})]$, which in turn is sufficient to find an $\epsilon$-additive approximation of $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[t(q(\mathbf{x})/d)\text{sign}(\mathbf{v}\cdot\mathbf{x})]$. We will show that this is computationally hard.

To start with, we note that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[q(\mathbf{x})^2] = \|\mathbf{T}\|_2 = 1$. Therefore, by standard concentration bounds, we have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}}[|q(\mathbf{x})| > d] = \exp(-\Omega(d^{2/5})) < \epsilon^3$. Therefore, by the Cauchy-Scwartz inequality, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[|q(\mathbf{x})/d - t(q(\mathbf{x}))/d|] \leq \sqrt{\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}}(|q(\mathbf{x})| > d) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[q(G\mathbf{x})^2]} \leq \epsilon/2 \; .$$

Thus, if one can approximate the maximum value of $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[t(q(\mathbf{x})/d)\text{sign}(\mathbf{v}\cdot\mathbf{x})]$ to additive error $\epsilon$, one can approximate the maximum value of $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(q(\mathbf{x})/d)\text{sign}(\mathbf{v}\cdot\mathbf{x})]$ to additive error $\epsilon/2$. However, we can compute this expectation by comparing the Hermite expansions for $q(\mathbf{x})/d$ and $\text{sign}(\mathbf{v}\cdot\mathbf{x})$. In particular, the former only has non-vanishing terms in degree 5, where they are given by the tensor $\mathbf{T}/d$. The latter has its degree-

5 Hermite tensor given by $c_5\mathbf{v}^{\otimes 5}$, where $c_5 = \mathbf{E}_{z\sim\mathcal{N}}[h_5(z)\text{sign}(z)] = (3/2)\sqrt{1/(15\pi)}$. Therefore, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[(q(\mathbf{x})/d)\text{sign}(\mathbf{v}\cdot\mathbf{x})] = (\mathbf{T}/d)\cdot(c_5\mathbf{v}^{\otimes 5}) = (c_5/d)\mathbf{T}(\mathbf{v},\mathbf{v},\mathbf{v},\mathbf{v},\mathbf{v}) = (c_5/d)p(\mathbf{v})\ .$$

Thus, finding an $\epsilon/2$-additive approximation to the maximum value of $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[(q(\mathbf{x})/d)\text{sign}(\mathbf{v}\cdot\mathbf{x})]$ for unit vectors $\mathbf{v}$ is equivalent to finding an $O(d^{-9})$-additive approximation to the maximum value of $p(\mathbf{v})$ over unit vectors $\mathbf{v}$. We claim that doing this would give a constant-factor multiplicative approximation to the maximum value of $p(\mathbf{v})$, finishing our reduction to the problem of Corollary 9.5.2. To do this, we need to show that the maximum value of $p(\mathbf{v})$ is much larger than $d^{-9}$.

To show this, we note that because $\|\mathbf{T}\|_2 = 1$, the sum of the squares of the entries of $\mathbf{T}$ is 1. Since $\mathbf{T}$ has only $d^5$ entries, this means that it must have some entry with norm at least $d^{-5}$. Therefore, there must be unit vectors $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_5$ so that $\mathbf{T}(\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3,\mathbf{v}_4,\mathbf{v}_5) \geq d^{-5}$. However, this value is proportional to $\sum_{\epsilon_1,\ldots,\epsilon_5 in\{\pm 1\}}\epsilon_1\epsilon_2\cdots\epsilon_5$ $p(\epsilon_1\mathbf{v}_1+\epsilon_2\mathbf{v}_2+\ldots+\epsilon_5\mathbf{v}_5)$. As each term here is proportional to $p$ of some unit vector (using the fact that $p$ is homogeneous), this implies that there is some unit vector $\mathbf{v}$ with $|p(\mathbf{v})| \gg d^{-5}$. Replacing $\mathbf{v}$ by its negation if necessary, we have that the maximum value of $p(\mathbf{v})$ over unit vectors $\mathbf{v}$ is $\Omega(d^{-5})$. This completes our proof. □

**Theorem 9.5.4** (Hardness of Proper Learning for ReLUs)**.** *Suppose that there is an algorithm that given query access to a real-valued function $f$ on $\mathbb{R}^d$ runs in* poly$(d)$ *time and approximates the minimum $L_2^2$ error between $f$ and a homogeneous ReLU (with respect to the standard Gaussian distribution) to additive error $\epsilon$ for some $\epsilon < d^{-4}$. Then there is a polynomial-time algorithm for the small set expansion problem.*

*Proof.* Let $p$ be a homogeneous degree-4 polynomial on $\mathbb{R}^d$. Let $\mathbf{T}$ be the unique symmetric tensor so that $p(\mathbf{x}) = \mathbf{T}(\mathbf{x}, \mathbf{x}, \mathbf{x}, \mathbf{x})$. By scaling $\mathbf{T}$, we may assume that $\|\mathbf{T}\|_2 = 1$. Let $f(\mathbf{x}) = (\mathbf{T} \cdot H(\mathbf{x}))$, where $H(\mathbf{x})$ is the tensor whose entries are the degree-4 Hermite polynomials in $\mathbf{x}$.

We can compute this expectation by comparing the Hermite expansions for $f(\mathbf{x})$ and $\mathrm{ReLU}(\mathbf{v} \cdot \mathbf{x})$. In particular, the former only has non-vanishing terms in degree 4, where they are given by the tensor $\mathbf{T}$. The latter has its degree-4 Hermite tensor given by $c_4 \mathbf{v}^{\otimes 4}$, where $c_4 = \mathbf{E}_{z \sim \mathcal{N}}[h_4(z)\mathrm{ReLU}(z)] = -(2\pi(24 + \sqrt{2}))^{-1}$. Therefore, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}}[f(\mathbf{x})\mathrm{ReLU}(\mathbf{v} \cdot \mathbf{x})] = \mathbf{T} \cdot (c_4 \mathbf{v}^{\otimes 4}) = c_4 \mathbf{T}(\mathbf{v}, \mathbf{v}, \mathbf{v}, \mathbf{v}) = c_4 p(\mathbf{v}) \ .$$

Thus, finding an $\epsilon$-additive approximation to the maximum value of $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}}[f(\mathbf{x})\mathrm{ReLU}(\mathbf{v} \cdot \mathbf{x})]$ for unit vectors $\mathbf{v}$ is equivalent to finding an $O(d^{-5})$-additive approximation to the maximum value of $p(\mathbf{v})$ over unit vectors $\mathbf{v}$. We claim that doing this would give a constant-factor multiplicative approximation to the maximum value of $p(\mathbf{v})$, finishing our reduction to the problem of Corollary 9.5.2. To do this, we need to show that the maximum value of $p(\mathbf{v})$ is much larger than $d^{-5}$.

To show this, we note that because $\|\mathbf{T}\|_2 = 1$, the sum of the squares of the entries of $\mathbf{T}$ is 1. Since $\mathbf{T}$ has only $d^4$ entries, this means that it must have some entry with norm at least $d^{-4}$. Therefore, there must be unit vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$ so that $\mathbf{T}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4) \geq d^{-4}$. However, this value is proportional to $\sum_{\epsilon_1, \ldots, \epsilon_4 \in \{\pm 1\}} \epsilon_1 \epsilon_2 \cdots \epsilon_4 p(\epsilon_1 \mathbf{v}_1 + \epsilon_2 \mathbf{v}_2 + \epsilon_3 \mathbf{v}_3 + \epsilon_4 \mathbf{v}_4)$. As each term here is proportional to $p$ of some unit vector (using the fact that $p$ is homogeneous), this implies that there is some unit vector $\mathbf{v}$ with $|p(\mathbf{v})| \gg d^{-4}$. Replacing $\mathbf{v}$ by its negation if necessary, we have that the maximum value of $p(\mathbf{v})$ over unit vectors $\mathbf{v}$ is $\Omega(d^{-4})$. This completes our proof. $\square$

# Chapter 10

# Agnostic PAC Learning Multi-Index Models

## 10.1 Introduction

This chapter develops a robust algorithmic framework for learning multi-index models (MIMs) from random examples, a more challenging setting than the query-based model of Chapter 9. We introduce an *iterative dimensionality reduction* technique that provides the first constant-factor approximate learners for a broad class of MIMs, including the important special case of multiclass linear classification (MLC), under adversarial label noise.

Before we describe our general setting and algorithm, we start with a specific basic model that was the original motivation for this work.

**Multiclass Linear Classification** A multiclass linear classifier is a function $f : \mathbb{R}^d \to [K]$ of the form $f(\mathbf{x}) = \mathrm{argmax}_{i\in[K]}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i)$ , where $\mathbf{w}^{(i)} \in \mathbb{R}^d$ and $t_i \in \mathbb{R}$ (ties are

resolved by selecting the smallest index). Multiclass linear classification (MLC) is a textbook machine learning problem [SSBD14], that has also been extensively studied in the online learning literature; see, e.g., [KSST08] and references therein.
The first question, we answer in this chapter is the following:

*Is there a* poly$(d, K)$ *time algorithm for multiclass linear classification in the presence of Random Classification Noise?*

Perhaps surprisingly, we provide strong evidence that the answer to the above question is negative. Specifically, we establish a Statistical Query (SQ) lower bound showing that the complexity of any SQ algorithm for this task is $d^{\Omega(K)}$.

**Theorem 10.1.1** (SQ Lower Bound for Learning $K$-MLC with RCN)**.** *For any $\epsilon \leq 1/K^C$, where $C$ a sufficiently large universal constant, any SQ algorithm that learns $K$-MLC under the Gaussian distribution in the presence of RCN to error* OPT $+ \epsilon$*, has complexity at least $d^{\Omega(K)}$.*

The next question we try to answer for MLCs is if there exists a fixed parameter tractability guarantee for the agnostic case:

*Is there a* constant-factor *approximate learner for MLC with complexity* poly$(d)f(\epsilon, K)$*?*

As a corollary of our approach, we answer this question in the affirmative (see Theorem 10.1.5).

### Learning Well-behaved Multi-Index Models with Noise

In fact the algorithm for MLC is a special case of a more general class of MIMs (Definition 1.3.30), namely *well-behaved* MIMs (Definition 10.1.2). We know present the formal definition of the properties we need our MIMs to have.

**Definition 10.1.2** (Well-Behaved Multi-Index Model)**.** *Fix $m \in \mathbb{Z}_+, \zeta, \tau \in (0,1)$, $\Gamma > 0$, and $\mathcal{Y} \subseteq \mathbb{Z}$ a set of finite cardinality. We say that a $K$-MIM $f : \mathbb{R}^d \to \mathcal{Y}$ is $(m, \zeta, \tau, \Gamma)$-well-behaved if $f$ has Gaussian surface area at most $\Gamma$, and for any random variable $y(\mathbf{x})$ supported on $\mathcal{Y}$ with $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) \neq y(\mathbf{x})] \leq \zeta$ and any subspace $V \subseteq \mathbb{R}^d$, one of the following conditions holds:*

*a) there exists a function $g : V \to \mathcal{Y}$ such that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}^V) \neq f(\mathbf{x})] \leq \tau$, or*

*b) with non-negligible probability[1] over $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, there exists a mean-zero, variance-one polynomial $p : U \to \mathbb{R}$ of degree at most $m$ and $z \in \mathcal{Y}$ such that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[p(\mathbf{x}^U)\mathbb{1}(y(\mathbf{x}) = z) \mid \mathbf{x}^V = \mathbf{x}_0^V] \geq \mathrm{poly}(\zeta/m) \ ,$$

*where $U = (V + W) \cap V^\perp$ is the projection of $W$ onto $V^\perp$ and $W$ is the subspace defining $f$.*

*We denote by $\mathcal{F}$ the class of $(m, \zeta, \tau, \Gamma)$-well-behaved $K$-MIMs over $\mathbb{R}^d$.*

Our main contribution is the following theorem:

**Theorem 10.1.3** (Learning Well-Behaved MIMs with Adversarial Label Noise)**.** *Let $\mathcal{F}$ be the class of $(m, \zeta, \tau, \Gamma)$-well-behaved $K$-MIMs on $\mathbb{R}^d$ of Definition 10.1.2. If $\zeta \geq \mathrm{OPT} + \epsilon$, there exists an agnostic algorithm that draws $N = d^m \, 2^{\mathrm{poly}(mK\Gamma/\epsilon)}$ i.i.d. samples, it runs in $\mathrm{poly}(d, N)$ time, and computes a hypothesis $h$ such that with high probability $\mathrm{err}_{0-1}^D(h) \leq \tau + \mathrm{OPT} + \epsilon$.*

In Table 10.1, we have several applications of Theorem 10.1.3 to well-studied concept classes, including MLC and intersections of halfspaces.

[1] By this we mean that the probability is not-too-small, specifically it is independent of the dimension $d$.

| Function Class | Prior Work | | Our Work | |
|---|---|---|---|---|
| | **Runtime** | **Error** | **Runtime** | **Error** |
| Agnostic $K$-MLC | $d^{\mathrm{poly}(K/\epsilon)}$ | $\mathrm{OPT}+\epsilon$ | (Theorem 10.1.5) $\mathrm{poly}(d)\,2^{\mathrm{poly}(K/\epsilon)}$ | $O(\mathrm{OPT})+\epsilon$ |
| Agnostic Intersections of $K$ halfspaces | $d^{O(\log(K)/\epsilon^4)}$ $\frac{\mathrm{poly}(d)}{\epsilon^{\mathrm{poly}(K)}}$ | $\mathrm{OPT}+\epsilon$ $\mathrm{poly}(K)\tilde{O}(\mathrm{OPT}^{\frac{1}{11}})$ | (Theorem 10.1.6) $\mathrm{poly}(d)\,2^{\mathrm{poly}(K/\epsilon)}$ | $K\tilde{O}(\mathrm{OPT})+\epsilon$ |
| Well-Behaved $K$-MIMs (Definition 10.1.2) | $d^{\mathrm{poly}(\Gamma/\epsilon)}$ | $\mathrm{OPT}+\epsilon$ | (Theorem 10.1.3) $d^{O(m)}\,2^{\mathrm{poly}(mK\Gamma/\epsilon)}$ | $\tau+\mathrm{OPT}+\epsilon$ |

Table 10.1: Comparison of runtime and error guarantees between state-of-the-art efficient algorithms (see [KOS08, DKS18] for intersections of $K$ halfspaces) and our subspace approximation procedure.

We complement our results with the following lower bound which shows that the complexity of Theorem 10.1.3 is essentially optimal in the SQ model.

**Theorem 10.1.4** (SQ Lower Bound for Learning Not-Well-behaved MIMs, Informal)**.** *Suppose that a function class $\mathcal{G}$ is rotationally invariant and contains a bounded surface area MIM not contained in the function class $\mathcal{F}$ of Definition 10.1.2 for parameters $m, \tau, \zeta$. Then, any SQ algorithm that learns $\mathcal{G}$ in the presence of $\zeta$ label noise to error smaller than $\tau - O(\zeta)$ has complexity at least $d^{\Omega(m)}$.*

### Concrete Learning Applications to Well-studied Concept Classes

In this subsection, we present algorithmic applications of our general MIM learner to two fundamental and well-studied concept classes.

**Multiclass Linear Classifiers** We denote by $\mathcal{L}_{d,K}$ the family of $K$-class linear classifiers in $\mathbb{R}^d$. For the adversarial noise setting, we obtain the following result (see Theorem 10.2.1):

**Theorem 10.1.5** (Learning $\mathcal{L}_{d,K}$ with Adversarial Label Noise)**.** *There exists an agnostic learning algorithm that draws $N = d\,2^{\mathrm{poly}(K/\epsilon)}$ i.i.d. labeled samples, runs in* $\mathrm{poly}(d, N)$ *time, and outputs a hypothesis $h$ such that with high probability* $\mathrm{err}^D_{0-1}(h) \leq O(\mathrm{OPT}) + \epsilon$*, where* OPT *is defined with respect to $\mathcal{L}_{d,K}$.*

**Intersections of Halfspaces** An intersection of $K$ halfspaces is any function $f : \mathbb{R}^d \to \{\pm 1\}$ for which there exist $K$ halfspaces $h_i(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i)$ such that $f(\mathbf{x}) = 1$ if and only if $h_i(\mathbf{x}) = 1$ for all $i \in [K]$. We denote by $\mathcal{F}_{K,d}$ the class of intersections of $K$ halfspaces on $\mathbb{R}^d$.

We obtain the following result for the class $\mathcal{F}_{K,d}$ (see Theorem 10.4.1 for a more detailed statement):

**Theorem 10.1.6** (Learning $\mathcal{F}_{K,d}$ with Adversarial Label Noise)**.** *There exists an agnostic learning algorithm that draws $N = O(d^2)\,2^{\mathrm{poly}(K/\epsilon)}$ i.i.d. labeled samples, runs in time* $\mathrm{poly}(d, N)$*, and computes a hypothesis $h$ such that with high probability* $\mathrm{err}^D_{0-1}(h) \leq K\widetilde{O}(\mathrm{OPT}) + \epsilon$*, where* OPT *is defined with respect to $\mathcal{F}_{K,d}$.*

### 10.1.1 Prior Work

The task of learning MIMs is fundamental in machine learning, with a long line of research in both statistics and computer science [FJS81, Hub85, Li91, HL93, XTLZ02, Xia08]. The majority of the literature focuses on learning various structured families of MIMs under natural distributional assumptions with clean labels [GLM18, JSA15b, DH18, BJW19, GKLW19, DKKZ20, DLS22, BBSS22, CDG+23]. A recent line of work has shown that achieving optimal error in the adversarial label setting under the Gaussian distribution cannot be achieved in fixed-degree polynomial time, even for very basic

MIM classes [DKPZ21, DKMR22a, Tie23]. These hardness results motivate the design of faster algorithms with relaxed error guarantees, which is the focus of this work. Despite this being a natural direction, no general algorithmic framework was previously known in this setting.

The recent work [DKK+24] developed agnostic learners for MIMs under the Gaussian distribution with fixed-parameter tractability guarantees similar to ours. The major difference is that, while our framework relies on learning from random examples, the latter work [DKK+24] assumes query access to the target function—a much stronger access model.

The task of efficiently learning intersections of halfspaces under Gaussian marginals is one of the most well-studied problems in the computational learning theory literature. A long line of work focused on achieving strong provable guarantees for this problem in the realizable setting [Vem10b, AV99, KLT09, KOS04]. However, the aforementioned work either fails in the presence of noise or scales superpolynomially with respect to the dimension (i.e., does not achieve fixed-degree polynomial runtime). An exception is the recent work [DKS18] that sacrifices the optimal error guarantee for a polynomial dependence in the input dimension.

Binary linear classification in the presence of noise under well-behaved distributions is a fairly well-understood problem, with a substantial body of work; see, e.g., [KKMS08, ABL17, DKS18, DKTZ20a, DKTZ20c, DKK+21b, DKK+22] and references therein. However, in the multiclass setting, most known results focus on online or bandit classification [KSST08, BPS+19]. That is, the subject of designing efficient algorithms in the multiclass setting that work in the presence of noise is rather poorly understood. It is worth pointing out that the multiclass regime offers several new interesting models of

noisy information that are challenging, even if the noise is random [PRKM+17, VRW17].

### 10.1.2 Technical Preliminaries

**Basic Notation** We denote by $2^S = \{T \mid T \subseteq S\}$ the power set of the set $S$ and $\overline{\mathbb{R}} \stackrel{\text{def}}{=} \mathbb{R} \cup \{\pm\infty\}$. Throughout this section, we will often omit the subscript and simply write $\|\mathbf{x}\|$ for the $\ell_2$-norm of $\mathbf{x}$. For vectors $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d$ and a subspace $V \subseteq \mathbb{R}^d$ denote by $\mathbf{x}^V$ the projection of $\mathbf{x}$ onto $V$, $\mathbf{x}^{\perp V}$ the projection of $\mathbf{x}$ onto the orthogonal complement of $V$ and by $\mathbf{x}^{\mathbf{v}}$ the projection of $\mathbf{x}$ onto the line spanned by $\mathbf{v}$. For two subspaces $V, W \subseteq \mathbb{R}^d$, we denote by $W^V = \{\mathbf{w}^V : \mathbf{w} \in W\}$ and by $V + W = \{\mathbf{w} + \mathbf{v} : \mathbf{w} \in W, \mathbf{v} \in V\}$, note that both $W^V$ and $V + W$ are subspaces. Furthermore, for a set of vectors $L \subseteq \mathbb{R}^d$, we denote by $\mathrm{span}(L)$ the subspace of $\mathbb{R}^d$ defined by their span. We slightly abuse notation and denote by $\mathbf{e}_i$ the $i$-th standard basis vector in $\mathbb{R}^d$. For a set $A \subseteq \mathbb{R}^d$ and a point $\mathbf{x} \in \mathbb{R}^d$, we define the distance from $\mathbf{x}$ to $A$ as $\mathrm{dist}(\mathbf{x}, A) = \min_{\mathbf{y} \in A} \|\mathbf{x} - \mathbf{y}\|$.

#### Gaussian Space Properties

**Fact 10.1.7** (Gaussian Density Properties)**.** *Let $\mathcal{N}$ be the standard one-dimensional normal distribution. Then, the following properties hold:*

1. *For any $t > 0$, it holds $e^{-t^2/2}/4 \leq \mathbf{Pr}_{z \sim \mathcal{N}}[z > t] \leq e^{-t^2/2}/2$.*
2. *For any $a, b \in \mathbb{R}$ with $a \leq b$, it holds $\mathbf{Pr}_{z \sim \mathcal{N}}[a \leq z \leq b] \leq (b - a)/\sqrt{2\pi}$.*

**Fact 10.1.8** (see, e.g., Lemma 6 in [KTZ19])**.** *Let $f \in L^2(\mathbb{R}^d, \mathcal{N}(0, I))$ with its $k$-degree Hermite expansion $f(\mathbf{x}) = \sum_{\alpha \in \mathbb{N}^d, \|\alpha\|_1 \leq k} \widehat{f}(\alpha) \mathrm{He}_\alpha(\mathbf{x})$. It holds that $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [(\nabla f(\mathbf{x}) \cdot \mathbf{e}_i)^2] = \sum_{\alpha \in \mathbb{N}^d \|\alpha\|_1 \leq k} \alpha_i (\widehat{f}(\alpha))^2$.*

**Definition 10.1.9** (Ornstein-Uhlenbeck Noise Operator)**.** *Let $k \in \mathbb{N}$ and $\rho \in [0,1]$. We define the Ornstein-Uhlenbeck operator $T_\rho : \{\mathbb{R}^d \to \mathbb{R}\} \to \{\mathbb{R}^d \to \mathbb{R}\}$ that maps $f : \mathbb{R}^d \to \mathbb{R}$ to the function $T_\rho f : \mathbb{R}^d \to \mathbb{R}$ with*

$$T_\rho f(x) = \mathop{\mathbf{E}}_{z \sim \mathcal{N}} \left[ f\left( \sqrt{1-\rho^2} \cdot x + \rho \cdot z \right) \right].$$

**Fact 10.1.10** (see, e.g., [Bog98])**.** *The following hold:*

1. *For any $f, g \in L_1$ and $\rho \in (0,1)$, it holds that $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}}[(T_\rho f) g] = \mathbf{E}_{\mathbf{x} \sim \mathcal{N}}[(T_\rho g(\mathbf{x})) f(\mathbf{x})]$ .*
2. *For any $g \in L_1$, it holds:*
   (a) *For any $\rho \in (0,1)$, $T_\rho g(\mathbf{x})$ is differentiable at every point $\mathbf{x}$.*
   (b) *For any $\rho \in (0,1)$ the $T_\rho g(\mathbf{x})$ is $\|g\|_\infty/\rho$-Lipschitz, i.e., $\|\nabla T_\rho g(\mathbf{x})\| \leq \|g\|_\infty/\rho$ for all $\mathbf{x} \in \mathbb{R}^d$.*
   (c) *For any $p \geq 1$, $T_\rho$ is a contraction with respect to the norm $\|\cdot\|_{L_p}$, i.e., it holds $\|T_\rho g\|_{L^p} \leq \|g\|_{L^p}$.*

**Fact 10.1.11** (Ledoux-Pisier [Pis86, Led94])**.** *Let $f : \mathbb{R}^d \mapsto \{\pm 1\}$ be a Boolean function. It holds*

$$\mathop{\mathbf{E}}_{x \sim \mathcal{N}}[f(x) T_\rho f(x)] \geq 1 - 2\sqrt{\pi}\, \Gamma(f)\, \rho.$$

**Fact 10.1.12.** *Let $\mathbf{c} \in \mathbb{R}^n$. Then the matrix*

$$\mathbf{M} = \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \cdots & c_{n-2} \\ c_{n-2} & c_{n-1} & c_0 & \cdots & c_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & c_3 & \cdots & c_0 \end{bmatrix}$$

*given by* $\mathbf{M}_{j,k} = c_{(k-j) \mod n}$ *has eigenvalues*

$$\lambda_j = \mathbf{c} \cdot \boldsymbol{\omega}^{(j)} \quad j = 0, 1, \ldots, n-1,$$

*where* $\boldsymbol{\omega}^{(j)} \in \mathbb{C}^n$ *a complex vector such that* $\boldsymbol{\omega}_k^{(j)} = e^{\frac{2\pi i}{n} jk}, k = 0, 1, \ldots, n-1.$

The following claim bounds the Gaussian noise sensitivity of a function when we re-randomize its input over a subspace rather than over the entire space $\mathbb{R}^d$. This claim is used in the proof of Claim 10.3.6 to show that a function with bounded Gaussian surface area remains nearly invariant when its input is re-randomized over sufficiently small cubes.

**Claim 10.1.13** (Gaussian Noise Sensitivity over a Subspace)**.** *Let* $f : \mathbb{R}^d \to [K]$ *be a a* $K$*-wise function and* $V$ *be a subspace of* $\mathbb{R}^d$

$$\mathbf{Pr}[f(\mathbf{x}) \neq f(\mathbf{x}^{\perp V} + \mathbf{x}'^{V})] \lesssim \sqrt{\epsilon}\Gamma(f) \,,$$

*where* $\mathbf{x}^V$ *and* $\mathbf{x}'^V$ *are* $(1-\epsilon)$*-correlated standard Gaussians.*

*Proof.* Without loss of generality, we assume that $f$ is a Boolean function, as we can use the union bound over the level sets to reduce to this case. Let the Hermite expansion

of $f$ be $f(\mathbf{x}) = \sum_{S\in\mathbb{N}^d} \widehat{f}(S)\mathrm{He}_S(\mathbf{x})$. Then we can express $f'(\mathbf{x}) = f(\mathbf{x}^{\perp V} + \mathbf{x}'^V)$ in the same basis as

$$f'(\mathbf{x}) = \sum_{S\in\mathbb{N}^d} \widehat{f}(S)\mathrm{He}_S(\mathbf{x}^{\perp V} + \mathbf{x}'^V) = \sum_{S\in\mathbb{N}^d} \widehat{f}(S)\mathrm{He}_S(\mathbf{x}^{\perp V})T_\epsilon[\mathrm{He}_S(\mathbf{x}'^V)] = \sum_{S\in\mathbb{N}^d} \epsilon^{|S_V|}\widehat{f}(S)\mathrm{He}_S(\mathbf{x})$$

where $S_V$ denotes subset of $S$ that corresponds to the coordinates in $V$. Hence the sensitivity over the subspace $V$ can be written as

$$\begin{aligned}\mathbf{Pr}[f(\mathbf{x}) \neq f(\mathbf{x}^{\perp V} + \mathbf{x}'^V)] &= \frac{1}{2} - \frac{1}{2}\mathbf{E}[f(\mathbf{x})f'(\mathbf{x})] = \frac{1}{2} - \frac{1}{2}\sum_{S\in\mathbb{N}^d} \epsilon^{|S_V|}(\widehat{f}(S))^2 \\ &\leq \frac{1}{2} - \frac{1}{2}\sum_{S\in\mathbb{N}^d} \epsilon^{|S|}(\widehat{f}(S))^2 = \frac{1}{2} - \frac{1}{2}\mathbf{E}[f(\mathbf{x})T_\epsilon[f(\mathbf{x})]] \,.\end{aligned}$$

Noting that the final expression is the Gaussian sensitivity, the result follows. □

**Legendre Polynomials** Legendre polynomials are orthogonal on the interval $[-1, 1]$. Their key properties are summarized in the following fact.

**Fact 10.1.14** ([Sze67])**.** *The Legendre polynomials $P_k$ for $k \in \mathbb{Z}$, satisfy the following properties:*

1. *$P_k$ is a $k$-degree polynomial and $P_0(x) = 1$ and $P_1(x) = x$.*
2. *$\int_{-1}^{1} P_i(x)P_j(x)\mathrm{d}x = 2/(2i+1)\mathbb{1}(i = j)$, for all $i, j \in \mathbb{Z}$.*
3. *$|P_k(x)| \leq 1$ for all $|x| \leq 1$.*
4. *$P_k(x) = (-1)^k P_k(-x)$.*
5. *$P_k(x) = 2^{-k}\sum_{i=1}^{\lceil k/2\rceil} \binom{k}{i}\binom{2k-2i}{k}x^{k-2i}$.*

### 10.1.3 Organization

In Section 10.2, we present a special case of our subspace approximation procedure for learning multiclass linear classifiers under agnostic noise, establishing Theorem 10.1.5. In Section 10.3, we generalize this approach, stating the necessary conditions (see Definition 10.1.2) for our algorithm to work for general classes of MIMs, establishing Theorem 10.1.3. In Section 10.4, we present additional applications of our algorithmic approach. In Section (10.4.1), we apply our general theorem to the class of intersections of halfspaces, establishing Theorem 10.1.6. Finally, Chapter 12 gives our SQ lower bounds. First, we demonstrate that the complexity of our algorithm is qualitatively optimal as a function of the dimension, establishing Theorem 10.1.4. Second, we give an SQ lower bound (Theorem 10.1.1) for learning multiclass linear classifiers under random classification noise.

## 10.2 Learning Multiclass Linear Classifiers with Adversarial Label Noise

In this section, we establish Theorem 10.1.5. Note that we introduce this special case as a preliminary step to build intuition for our general algorithm in Section 10.3. First, we state the formal version of the theorem and provide the algorithm pseudocode along with the necessary definitions.

**Theorem 10.2.1** (Learning Multiclass Linear Classifiers with Agnostic Noise)**.** *Let $\mathcal{L}_{d,K}$ be the class of $K$-class linear classifiers in $\mathbb{R}^d$. Let $D$ be a distribution over $\mathbb{R}^d \times [K]$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$, let* $\mathrm{OPT} = \inf_{f \in \mathcal{L}_{d,K}} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y]$ *and let $\epsilon, \delta \in (0,1)$. There exists a sufficiently large universal constant $C > 0$ such that*

*Algorithm 24 draws* $N = d2^{\text{poly}(1/\epsilon,K)}\log(1/\delta)$ *i.i.d. samples from* $D$*, runs in* $\text{poly}(N)$ *time, and returns a hypothesis* $h$ *such that, with probability at least* $1-\delta$*, it holds that* $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq C \cdot \text{OPT} + \epsilon$.

Before we present the algorithm, we provide the following definitions, which characterize how we discretize a low-dimensional subspace and approximate distributions over it.

To effectively approximate the distribution over a specific subspace $V$, we aim to preserve a substantial portion of its mass and partition it into fine regions of non-negligible mass. By selecting an orthonormal basis for $V$ and excluding regions where any coordinate exceeds $\sqrt{\log(k/\epsilon)}$, we ensure that at least a $1-\epsilon$ fraction of the mass remains. The remaining space is then partitioned into cubes aligned with $V$, each with edges of length $\epsilon$ and containing approximately $\epsilon^{\dim(V)}$ mass.

**Definition 10.2.2** ($\epsilon$-Approximating Partition)**.** *Let* $V$ *be a* $k$*-dimensional subspace of* $\mathbb{R}^d$ *with an orthonormal basis* $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}$*, and let* $\epsilon \in (0,1)$*. An* $\epsilon$*-approximating partition with respect to* $V$ *is a collection of sets* $\mathcal{S}$ *defined as follows: For each multi-index* $\mathbf{j} = (\mathbf{j}_1, \ldots, \mathbf{j}_k) \in [M_\epsilon]^k$*, define* $S_{\mathbf{j}} = \{\mathbf{x} \in \mathbb{R}^d : z_{\mathbf{j}_i - 1} \leq \mathbf{v}^{(i)} \cdot \mathbf{x} \leq z_{\mathbf{j}_i}, i \in [k]\}$ *where* $z_i$*'s are defined as* $z_i = -\sqrt{2\log(k/\epsilon)} + i\epsilon + t$*, for* $i \in \{0, \ldots, M_\epsilon\}, t \in (0, \epsilon/2)$ *and* $M_\epsilon = \left\lceil (2\sqrt{2\log(k/\epsilon)})/\epsilon \right\rceil$*. Furthermore, to each set* $S_{\mathbf{j}} \in \mathcal{S}$*, we associate a set of linear equations* $\mathcal{G}_{S_{\mathbf{j}}} = \{\mathbf{v}^{(1)} \cdot \mathbf{x} = t_1, \ldots, \mathbf{v}^{(k)} \cdot \mathbf{x} = t_k\}$*, where for each* $i \in [k]$*,* $t_i$ *is chosen so that* $z_{\mathbf{j}_i - 1} \leq t_i \leq z_{\mathbf{j}_i}$.

Furthermore, we define the piecewise constant approximation of a distribution over labeled examples as the function that minimizes the 0-1 loss with respect to the labels in each subset of a partition. Our algorithm will first estimate the hidden subspace of

the multiclass linear classifier and then return a piecewise constant approximation over it.

**Definition 10.2.3** (Piecewise Constant Approximation)**.** *Let $D$ be a distribution over $\mathbb{R}^d \times [K]$, let $V$ be a subspace of $\mathbb{R}^d$ and let $\epsilon \in (0,1)$. Let $\mathcal{S}$ be an $\epsilon$-approximating partition with respect to $V$ (see Definition 10.2.2). A piecewise constant approximation of the distribution $D$, with respect to $\mathcal{S}$, is any classifier $h_{\mathcal{S}} : \mathbb{R}^d \to [K]$ such that for each $S \in \mathcal{S}$ and $\mathbf{x} \in S$, $h_{\mathcal{S}}$ is defined as follows $h_{\mathcal{S}}(\mathbf{x}) \in \operatorname{argmin}_{i \in [K]} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[i \neq y \mid \mathbf{x} \in S]$.*

We now present our subspace approximation method for learning multiclass linear classifiers under agnostic noise (see Algorithm 24).

**Algorithm 24** Learning Multiclass Linear Classifiers under Agnostic Noise

**Input:** Accuracy $\epsilon > 0$, failure probability $\delta > 0$, sample access to distribution $D$ over $\mathbb{R}^d \times [K]$ with $\mathbf{x}$–marginal $\mathcal{N}(\mathbf{0}, \mathbf{I})$
**Output:** Hypothesis $h$ such that w.p. $\geq 1 - \delta$, $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[h(\mathbf{x}) \neq y] \leq C \cdot \mathrm{OPT} + \epsilon$, where $\mathrm{OPT} = \inf_{f \in \mathcal{L}_{d,K}} \mathbf{Pr}[f(\mathbf{x}) \neq y]$
Choose a sufficiently large constant $c > 0$, and let $T, p$ be sufficiently large constant-degree polynomials in $1/\epsilon$ and $K$
$N \leftarrow d\, 2^{T^c} \log \frac{1}{\delta}, \quad \epsilon' \leftarrow 1/p, \quad L_1 \leftarrow \emptyset, \quad t \leftarrow 1$
**while** $t \leq T$ **do**
    Draw $S_t$: $N$ i.i.d. samples from $D$
    $\mathcal{E}_t \leftarrow$ Alg. 25$(\epsilon, \epsilon', \delta, \mathrm{span}(L_t), S_t)$
    $L_{t+1} \leftarrow L_t \cup \mathcal{E}_t$
    $t \leftarrow t + 1$
**end while**
Construct an $\epsilon'$–approximating partition $\mathcal{S}$ w.r.t. $\mathrm{span}(L_{T+1})$ (Def. 10.2.2)
Draw $N$ i.i.d. samples from $D$ and build piecewise-constant classifier $h_{\mathcal{S}}$ by majority vote in each cell
**Return:** $h_{\mathcal{S}}$

**Algorithm 25** Estimating a Relevant Direction

**Input:** tolerances $\epsilon, \epsilon', \delta > 0$; subspace $V \subseteq \mathbb{R}^d$; samples $\{(\mathbf{x}^{(i)}, y_i)\}_{i=1}^N \sim D$ over $\mathbb{R}^d \times [K]$ with $\mathbf{x}$–marginal $\mathcal{N}(\mathbf{0}, \mathbf{I})$
**Output:** set of unit vectors $\mathcal{E}$
Choose $\sigma$ as a sufficiently small constant-degree polynomial in $\epsilon$ and $1/K$
Form empirical distribution $\widehat{D}$ from the samples
Construct $\mathcal{S}$, an $\epsilon'$–approximating partition w.r.t. $V$ (Def. 10.2.2)
**for all** $S \in \mathcal{S}$ **do**
    **for all** $i \in [K]$ **do**
        Compute $\widehat{\mathbf{u}}^{(S,i)} \leftarrow \mathbf{E}_{(\mathbf{x},y)\sim\widehat{D}}[\mathbf{x}\,\mathbf{1}(y = i) \mid \mathbf{x} \in S]$
    **end for**
**end for**
Compute $\widehat{\mathbf{U}} = \sum_{(S,i)\in\mathcal{S}\times[K]} \left(\widehat{\mathbf{u}}^{(S,i)}\right)^{\perp V} \left(\widehat{\mathbf{u}}^{(S,i)}\right)^{\perp V\,\intercal} \mathbf{Pr}_{(\mathbf{x},y)\sim\widehat{D}}[S]\,.$
**Return:** all unit–eigenvectors of $\widehat{\mathbf{U}}$ with eigenvalues $\geq \sigma$

Let $f^*$ be a $K$-multiclass linear classifier with error OPT, i.e., $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] = \mathrm{OPT}$. Algorithm 24 proceeds in a sequence of iterations: In each iteration $t$, it maintains and updates a set of unit vectors $L_t$ so that its span $V_t$ provides a better approximation to the span of the weight vectors of $f^*$, denoted by $W^*$. To this end, Algorithm 24 uses Algorithm 25 to construct a new list $\mathcal{E}_t$ that contains vectors with non-trivial correlation to $(W^*)^{\perp V_t}$.

First, in Section (10.2.1), we show that if $f^*$ is not $C \cdot \mathrm{OPT}$-close to any constant classifier, then there exists at least one label $i \in [K]$ for which the first moment $\mathbf{E}[\mathbf{x}\mathbb{1}(y = i)]$ correlates non-trivially (by at least $\mathrm{poly}(\epsilon/K)$) with a vector in $W^*$. This, in fact, suffices to initialize a list of vectors that correlate with a vector in $W^*$. In order to expand this list and at the same time improve the approximation of $W^*$, we need to show that, we can find vectors that not only correlate with $W^*$ but are also orthogonal to the current list of vectors $L_t$.

To this end, Algorithm 25 uses the current set $L_t$ of weight vectors to partition $\mathbb{R}^d$

into sufficiently small regions $S$. In each such region $S$, the classifier $f^*$ has a limited dependence on the vectors in $L_t$. Specifically, we show that $f^*$ is $\epsilon$-close to a function $f^*_S$ that does not depend on weight vectors contained in $L_t$. Thus, within each region $S$, we obtain a multiclass classifier that depends only on vectors in $(W^*)^{\perp V_t}$. Therefore, in every region $S$, either the classifier can be well-approximated by a constant function or the first moments will correlate non-trivially with a vector in $(W^*)^{\perp V_t}$, thus enabling further improvement of the approximation.

Furthermore, for $f^*$ to depend mostly on directions in $(W^*)^{\perp V_t}$, the cubes $S$ in the partition must have width on the order of $\mathrm{poly}(\epsilon/K)$. Such a fine partitioning would typically yield a number of cubes that is exponential in $\dim(V_t)$ and, hence could lead to a power-tower update in the dimension of $V_t$ when updating $L_t$. In order to avoid this power-tower dependence, we leverage the fact that the total probability mass of cubes with non-trivial correlation is at least $\mathrm{poly}(\epsilon/K)$ (this will be proved in the analysis of Algorithm 25 in Proposition 10.2.8). By weighting each moment vector by the probability mass of its corresponding cube (see Line 11 of Algorithm 25), we are guaranteed to find a vector with non-trivial correlation by examining only a $\mathrm{poly}(K/\epsilon)$-sized list of candidate vectors.

As follows from the above, our algorithm is not proper; that is, it does not return a multiclass linear classifier. Instead, it produces a piecewise constant function over a sufficiently fine partition of $\mathbb{R}^d$.

See Figure 10.1 for an illustration.

**Organization of Section 10.2** The body of this section is structured as follows: In Section (10.2.1), we state and prove the key structural result about the class of multiclass linear classifiers, Proposition 10.2.4, which will be exploited by our algorithm.

In Section (10.2.2), we provide the analysis of Algorithm 25 in Proposition 10.2.8. In Section (10.2.3), we analyze Algorithm 24, providing the full proof of Theorem 10.2.1. Finally, in Section (10.2.4), we provide the proof of Claim 10.2.19, a supporting lemma that is used in the proof of Proposition 10.2.4.

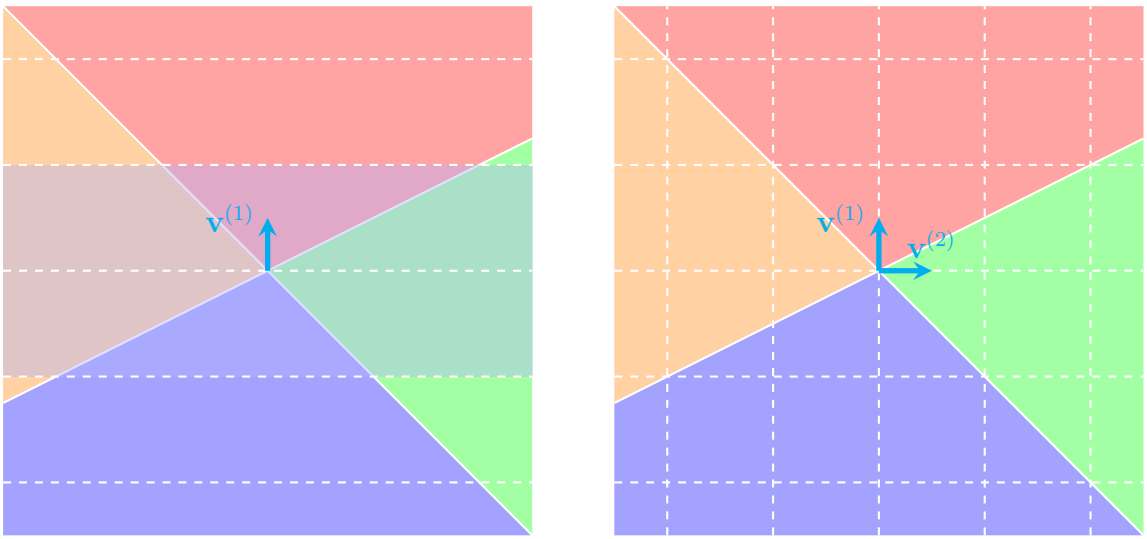



Figure 10.1: Illustration of two consecutive steps of our iterative subspace approximation procedure for 4-class Multiclass Linear Classification. In the left figure, the space $\mathbb{R}^d$ is partitioned according to one already discovered direction. The highlighted rectangular regions correspond to areas where a constant function over the current subspace $V$ yields an insufficient approximation, indicating that additional relevant directions can be found. In contrast, the non-highlighted regions already permit a sufficiently good constant approximation. The right figure illustrates that we discovered the second hidden direction so that partitioning the space to sufficiently small cubes leads to a good approximation.

### 10.2.1 Structural Result for Multiclass Linear Classification

In this section, we prove our main structural result, Proposition 10.2.4, which enables us to efficiently learn multiclass linear classifiers under adversarial noise. In particular, we show that for a multiclass linear classifier $f$, if $f$ is far from any constant function

by more than a constant multiple of its error with respect to the noisy label $y$, then the first moment of at least one of the noisy label regions has a non-trivial correlation with a vector in the subspace of weights defining $f$.

We prove this by finding carefully chosen vectors $\mathbf{v}^{(i)} \in \mathbb{R}^d$, $i \in [K]$, such that $\sum_{i=1}^K \mathbf{E}[\mathbf{v}^{(i)} \cdot \mathbf{x}\mathbb{1}(y = i)]$ is guaranteed to be non-trivially large. Note that this sum is equal to $\sum_{i=1}^K \mathbf{E}[\mathbf{v}^{(i)} \cdot \mathbf{x}\mathbb{1}(f(\mathbf{x}) = i)]$ plus an error term for instances when $y$ is not equal to $f(\mathbf{x})$. By carefully selecting these vectors, we ensure that even a small error probability does not significantly reduce the overall sum. Formally, we show the following:

**Proposition 10.2.4** (Structural Result)**.** *Let $d \in \mathbb{Z}_+$, $\epsilon \in (0, 1)$, $D$ be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times [K]$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Also, let $f$ be a multiclass linear classifier, $f(\mathbf{x}) = \mathrm{argmax}_{i \in [K]}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i)$ with $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)} \in \mathbb{R}^d$, and let $C$ be a sufficiently large positive universal constant. If $\epsilon \leq 1/C$ and $f(\mathbf{x})$ is not close to any constant classifier, i.e., $\min_{i \in [K]} \mathbf{Pr}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[f(\mathbf{x}) \neq i] \geq C \cdot \mathbf{Pr}_{(\mathbf{x},y) \sim D}[f(\mathbf{x}) \neq y] + \epsilon$, then for $j = \mathrm{argmax}_{i \in [K]} t_i$ there exists $i \in [K]$ with $\mathbf{w}^{(i)} - \mathbf{w}^{(j)} \neq \mathbf{0}$ such that*

$$\frac{\mathbf{w}^{(i)} - \mathbf{w}^{(j)}}{\|\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\|} \cdot \mathop{\mathbf{E}}_{(\mathbf{x},y) \sim D}[\mathbf{x}\mathbb{1}[y = i]] \geq \mathrm{poly}(1/K, \epsilon) \, .$$

*Proof.* Define $\mathrm{OPT} \stackrel{\mathrm{def}}{=} \mathbf{Pr}_{(\mathbf{x},y) \sim D}[f(\mathbf{x}) \neq y]$. Without loss of generality, we may assume that $\mathbf{w}^{(1)} = \mathbf{0}$ and $t_1 = 0$. This follows from the observation that subtracting the same function from every classifier does not change the output label. In particular, if we subtract $\mathbf{w}^{(j)} \cdot \mathbf{x} + t_j$, where $j \in \mathrm{argmax}_{i \in [K]} t_i$, from all affine functions and then rename the labels accordingly, one of the resulting affine functions becomes identically $\mathbf{0}$ and we obtain $t_i \leq 0$ for all $i \in [K]$.

We define $L_i : \mathbb{R}^d \to \mathbb{R}$, $L_i(\mathbf{x}) = \mathbf{w}^{(i)} \cdot \mathbf{x} + t_i$ the corresponding affine functions such

that $f(\mathbf{x}) = \text{argmax}_{i\in[K]} L_i(\mathbf{x})$ and we define the function $s(\mathbf{x}) \stackrel{\text{def}}{=} \max_{i\in[K]} L_i(\mathbf{x})$.

For some parameter $\gamma > 0$ which we will chose appropriately below, let $\mathbf{v}^{(i)} \stackrel{\text{def}}{=} \mathbf{w}^{(i)}/(\gamma - t_i)$ (this is well-defined as $\gamma > 0$ and $t_i \leq 0$) for all $i \in [K]$. Let $g(\mathbf{x}) \stackrel{\text{def}}{=} \max_{i\in[K]} \mathbf{v}^{(i)} \cdot \mathbf{x}$. Note that for all nonzero vectors $\mathbf{v}^{(i)}$, we have that $L_i(\mathbf{x}) \geq \gamma$ if and only if $\mathbf{v}^{(i)} \cdot \mathbf{x} \geq 1$ which means that $g(\mathbf{x}) \geq 1$. Furthermore, because $\mathbf{w}^{(1)} = \mathbf{0}$ and $t_1 = 0$, we have that $g(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$. We can decompose the first moment of label $i$ as follows:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{x}\mathbb{1}(y = i)] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{x}\mathbb{1}(f(\mathbf{x}) = i)] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{x}\mathbb{1}(y \neq i, f(\mathbf{x}) = i)] + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{x}\mathbb{1}(y = i, f(\mathbf{x}) \neq i)] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbf{x}\mathbb{1}(f(\mathbf{x}) = i)] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{x}S_i(\mathbf{x}, y)] + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{x}T_i(\mathbf{x}, y)] \,, \qquad (10.1)
\end{aligned}$$

where $S_i(\mathbf{x}, y) = \mathbb{1}(y \neq i, f(\mathbf{x}) = i)$ and $T_i(\mathbf{x}, y) = \mathbb{1}(y = i, f(\mathbf{x}) \neq i)$ for $i \in [K]$. Taking the inner product of $\mathbf{E}[x\mathbb{1}(y = i)]$ with $\mathbf{v}^{(i)}$ and summing Equation (10.1) over $i \in [K]$, we have that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{v}^{(y)} \cdot \mathbf{x}] = \underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbf{v}^{(f(\mathbf{x}))} \cdot \mathbf{x}]}_{I_1} \underbrace{- \sum_{i=1}^{K} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{v}^{(i)} \cdot \mathbf{x}S_i(\mathbf{x}, y)] + \sum_{i=1}^{K} \mathop{\mathbf{E}}_{(x,y)\sim D}[\mathbf{v}^{(i)} \cdot \mathbf{x}T_i(\mathbf{x}, y)]}_{I_2} \qquad (10.2)$$

Note that $I_1$ is the the contribution of each $\mathbf{v}^{(i)}$ over $\mathbf{x}$ in the region where $f(\mathbf{x}) = i$, and $I_2$ is the contribution that corresponds to the noise. We first bound $I_1$ from below.

**Claim 10.2.5.** *It holds that* $I_1 \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}) \geq 1]$.

*Proof of Claim 10.2.5.* Note that the random variable $\mathbf{v}^{(f(\mathbf{x}))} \cdot \mathbf{x}$ is non-negative since

$L_{f(\mathbf{x})}(\mathbf{x}) = \max_i L_i(\mathbf{x}) \geq L_1(\mathbf{x}) = 0$ which implies that $\mathbf{w}^{(f(\mathbf{x}))} \cdot \mathbf{x} \geq |t_i|$ and consequently, $\mathbf{v}^{(f(\mathbf{x}))} \cdot \mathbf{x} \geq 0$. Hence, by applying Markov's inequality, we have

$$I_1 = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbf{v}^{(f(\mathbf{x}))} \cdot \mathbf{x}] \geq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbf{v}^{(f(\mathbf{x}))} \cdot \mathbf{x} \geq 1] = \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[L_{f(\mathbf{x})}(\mathbf{x}) \geq \gamma] \,,$$

where the last equality holds by the definition of $\mathbf{v}^{(i)}$. It remains to show that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[L_{f(\mathbf{x})}(\mathbf{x}) \geq \gamma] \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}) \geq 1]$. To see this, note that the event $L_{f(\mathbf{x})}(\mathbf{x}) \geq \gamma$ contains the event $g(\mathbf{x}) \geq 1$. If $g(\mathbf{x}) \geq 1$ then by definition there exists some $i \in [K]$ such that $\mathbf{v}^{(i)} \cdot \mathbf{x} \geq 1$, which implies that $L_i(\mathbf{x}) \geq \gamma$ (recall that $\mathbf{v}^{(i)} = \mathbf{w}^{(i)}/(\gamma - t_i)$), and consequently, $L_{f(\mathbf{x})}(\mathbf{x}) \geq \gamma$. This gives that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[L_{f(\mathbf{x})}(\mathbf{x}) \geq \gamma] \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}) \geq 1]$ and completes the proof of Claim 10.2.5. □

It remains to bound the contribution from the noise, i.e., the quantity $I_2$.

**Claim 10.2.6.** *There exists an absolute constant $C' > 0$ and $\gamma > 0$ such that*

$$I_2 \geq -C' \left( \sqrt{\mathrm{OPT} \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}) \geq 1]} + \mathrm{OPT} \right) \,,$$

*and* $1/48 \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma] \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0]/96$.

*Proof.* We show how to bound the contribution of the terms involving $S_i$; similar arguments can be used to bound the contribution of the terms involving $T_i$. In order to

bound the contribution of the $S_i$ terms we can write

$$\begin{aligned}\sum_{i=1}^{K} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{v}^{(i)} \cdot \mathbf{x} S_i(\mathbf{x}, y)] &\leq \sum_{i=1}^{K} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\left[\max_{j\in[K]}(\mathbf{v}^{(j)} \cdot \mathbf{x}) S_i(\mathbf{x}, y)\right] \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\left[\max_{j\in[K]}(\mathbf{v}^{(j)} \cdot \mathbf{x}) S(\mathbf{x}, y)\right] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\max_{j\in[K]}(\mathbf{v}^{(j)} \cdot \mathbf{x}) \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[S(\mathbf{x}, y) \mid \mathbf{x}]\right] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[g(\mathbf{x}) \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[S(\mathbf{x}, y) \mid \mathbf{x}]\right] ,\end{aligned}$$

where $S = \sum_{i=1}^{K} S_i$. Since $f$ has misclassification error OPT, we have $\mathbf{E}_{(\mathbf{x},y)\sim D}[S(\mathbf{x}, y)] \leq$ OPT. Moreover, for each point $(\mathbf{x}, y)$, at most one among $\{S_i, T_i : i \in [K]\}$ can be nonzero which implies that $\mathbf{E}_{(\mathbf{x},y)\sim D}[S(\mathbf{x}, y) \mid \mathbf{x}] \leq 1$. Define $R(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{E}_{(\mathbf{x},y)\sim D}[S(\mathbf{x}, y) \mid \mathbf{x}]$. Then, by the law of total expectation, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[R(\mathbf{x})] = \mathbf{E}_{(\mathbf{x},y)\sim D}[S(\mathbf{x}, y)] \leq$ OPT and $R(\mathbf{x}) \in [0, 1]$.

Recall that $g(\mathbf{x}) \geq 0$ by construction, hence for any $t > 0$,

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x})R(\mathbf{x})] &\leq t \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[R(\mathbf{x})\mathbb{1}(g(\mathbf{x}) \leq t)] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x})R(\mathbf{x})\mathbb{1}(g(\mathbf{x}) > t)] \\ &\leq t\mathrm{OPT} + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x})\mathbb{1}(g(\mathbf{x}) > t)] ,\end{aligned}$$

where we use the fact that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[R(\mathbf{x})] \leq$ OPT and $R(\mathbf{x}) \leq 1$. We next bound the term $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x})\mathbb{1}(g(\mathbf{x}) > t)]$. Using that for any positive function $G(\mathbf{x})$ it holds that

$G(\mathbf{x}) = \int_0^\infty \mathbb{1}(G(\mathbf{x}) \geq z)\mathrm{d}z$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x})\mathbb{1}(g(\mathbf{x}) > t)] &= \int_t^\infty \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbb{1}(g(\mathbf{x}) \geq y)]\mathrm{d}y \\ &\leq \int_t^\infty \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbb{1}(g(\mathbf{x}) \geq \lfloor y \rfloor)]\mathrm{d}y = \sum_{n=\lfloor t \rfloor}^{\infty} \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[n \leq g(\mathbf{x})]\ ,\end{aligned} \tag{10.3}$$

where in the first equality, we exchange the order of the integrals using Tonelli's theorem and in the inequality, we used that $\mathbb{1}(g(\mathbf{x}) \geq y)$ is a decreasing function of $y$. In order to bound the terms $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[n \leq g(\mathbf{x})]$, we show that the tails of $g$ decay at least quadratically. We show the following:

**Lemma 10.2.7.** *Let $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)} \in \mathbb{R}^d$, $t_1, \ldots, t_K \in \mathbb{R}_-$ and $s : \mathbb{R}^d \to \mathbb{R}$ such that $s(\mathbf{x}) = \max_{i\in[K]} \max(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i, 0)$. There exists $\gamma > 0$ such that*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0]/96 \leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma] \leq 1/48$$

*and*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq n\gamma] \lesssim (1/n^2) \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma]$$

*for all $n \in \mathbb{Z}_+$.*

For the proof of Lemma 10.2.7, we refer the reader to Section (10.2.4). Now, we leverage Lemma 10.2.7 to complete the proof. Recall that the function $s(\mathbf{x})$ of Lemma 10.2.7 is defined as $s(\mathbf{x}) = \max_i L_i(\mathbf{x})$. We need to show that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}) \geq n] \leq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq n\gamma]$. First, note that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}) \geq 1] = \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma]$, by construction. Furthermore, recall that $g(\mathbf{x}) = \max_{i\in[K]} \mathbf{v}^{(i)} \cdot \mathbf{x}$ where $\mathbf{v}^{(i)} = \mathbf{w}^{(i)}/(\gamma - t_i)$ and that $L_i(\mathbf{x}) = \mathbf{w}^{(i)} \cdot \mathbf{x} + t_i$. Since $t_i \leq 0$ for all $i \in [K]$, it follows

that if $\mathbf{w}^{(i)} \cdot \mathbf{x} \geq n(\gamma - t_i)$, then $\mathbf{w}^{(i)} \cdot \mathbf{x} \geq n\gamma - t_i$. This, in turn, implies that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}) \geq n] \leq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq n\gamma]$. Therefore, using Lemma 10.2.7 into Equation (10.3), we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x})\mathbb{1}(g(\mathbf{x}) > t)] \lesssim \sum_{n=\lfloor t \rfloor}^{\infty} \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[1 \leq g(\mathbf{x})]/n^2 \lesssim \frac{\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[1 \leq g(\mathbf{x})]}{t-2} ,$$

where in the last inequality we used the approximation of the sum by an integral. Therefore, by choosing $t = 2 + (\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[1 \leq g(\mathbf{x})]/\mathrm{OPT})^{1/2}$, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x})R(\mathbf{x})] \lesssim t\mathrm{OPT} + \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[1 \leq g(\mathbf{x})]/(t-2) \lesssim \sqrt{\mathrm{OPT} \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[1 \leq g(\mathbf{x})]} + \mathrm{OPT}$$

For the noisy terms defined by $T_i$, we have that

$$\sum_{i=1}^{K} \mathop{\mathbf{E}}_{(x,y)\sim D}[\mathbf{v}^{(i)} \cdot \mathbf{x}T_i(\mathbf{x},y)] = -\sum_{i=1}^{K} \mathop{\mathbf{E}}_{(x,y)\sim D}[\mathbf{v}^{(i)} \cdot (-\mathbf{x})T_i(\mathbf{x},y)] = -\sum_{i=1}^{K} \mathop{\mathbf{E}}_{(x,y)\sim D'}[\mathbf{v}^{(i)} \cdot \mathbf{x}T_i(\mathbf{x},y)] ,$$

where $D'$ is a reflection of distribution $D$ with respect to $\mathbf{0}$. As a result, we can arrive at the same bound as for the term defined by the $S_i$'s. Therefore, we conclude that the total contribution of the noise is at most $O(\sqrt{\mathrm{OPT}\,\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[1 \leq g(\mathbf{x})]} + \mathrm{OPT})$, which completes the proof of Claim 10.2.6. □

Now we show that $I_2 \geq -\mathbf{Pr}_{x\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}) \geq 1]/2$. Recall that by the assumptions of Proposition 10.2.4 we have

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[f(\mathbf{x}) \neq i\right] \geq C\,\mathrm{OPT} + \epsilon \quad \text{for all } i \in [K].$$

We show that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[s(\mathbf{x}) \neq 0\right] \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[f(\mathbf{x}) \neq 1\right]$. To see this, observe that

for any $i \in [K]$ for which $L_i$ is not identically zero, we have that $L_i(\mathbf{x}) \neq 0$ almost surely (with respect to the Gaussian measure). Moreover, if for some $i \in [K]$ the function $L_i$ is identically zero and $i \neq 1$, then without loss of generality we may assume that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y = i] = 0$, since we could set 1 to be the label with minimum index in the set $\operatorname{argmax}_{i\in[K]} t_i$ (for the original biases $t_i$). Hence, in every case for any $i \neq 1$ we have $L_i(\mathbf{x}) \neq 0$ almost surely. Therefore, if $f(\mathbf{x}) \neq 1$ then, except for a measure zero set it holds that $L_i(\mathbf{x}) > 0$ for some $i \neq 1$, which implies that $s(\mathbf{x}) \neq 0$. As a result, it holds that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[s(\mathbf{x}) \neq 0\right] \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[f(\mathbf{x}) \neq 1\right]$. Hence, it follows that

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[s(\mathbf{x}) \neq 0\right] \geq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[f(\mathbf{x}) \neq 1\right] \geq C\,\mathrm{OPT} + \epsilon.$$

By the choice of $\gamma$ in Claim 10.2.6, we also have

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[s(\mathbf{x}) \geq \gamma\right] \geq \frac{1}{96} \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[s(\mathbf{x}) \neq 0\right].$$

Since $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[g(\mathbf{x}) \geq 1\right] = \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[s(\mathbf{x}) \geq \gamma\right]$, we get that that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}) \geq 1] \geq (C\,\mathrm{OPT} + \epsilon)/96$. Finally, by choosing $C$ sufficiently large (in particular, ensuring that $C \geq 256/((C')^2 + 1)$), from Claim 10.2.6, we conclude that

$$I_2 \geq -\frac{1}{2} \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[g(\mathbf{x}) \geq 1\right].$$

and that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[g(\mathbf{x}) \geq 1\right] \gtrsim \epsilon$. Combining the above with Equation (10.2) and Claim 10.2.5, we obtain

$$\sum_{i:\mathbf{w}^{(i)}\neq\mathbf{0}} \mathbf{w}^{(i)}/(\gamma - t_i) \cdot \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{x}\mathbb{1}(y = i)] \gtrsim \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[g(\mathbf{x}) \geq 1] \gtrsim \epsilon\,.$$

This implies that there exists some $j \in [K]$ such that

$$\frac{\mathbf{w}^{(j)}}{\gamma - t_j} \cdot \mathbf{E}[\mathbf{x}\, \mathbb{1}(y = j)] \gtrsim \epsilon / K.$$

To conclude the proof, we must show that $\gamma - t_j \geq \epsilon \left\|\mathbf{w}^{(j)}\right\|$. Suppose, for the sake of contradiction, that for some $i \in [K]$ we have $\gamma - t_i \leq \epsilon \left\|\mathbf{w}^{(i)}\right\|$. Then, by the definition of $L_i$, we have

$$\mathop{\mathbf{Pr}}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[ L_i(\mathbf{x}) \geq \gamma \right] = \mathop{\mathbf{Pr}}_{x \sim \mathcal{N}(0,1)} \left[ x \geq \frac{\gamma - t_i}{\|\mathbf{w}^{(i)}\|} \right] \geq \mathop{\mathbf{Pr}}_{x \sim \mathcal{N}(0,1)} [x \geq \epsilon] \,.$$

However, for $\epsilon < 2\sqrt{\log(48)}$, Fact 10.1.7 implies that

$$\mathop{\mathbf{Pr}}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[ \max_{i \in [K]} L_i(\mathbf{x}) \geq \gamma \right] > \frac{1}{48},$$

which leads to a contradiction from Lemma 10.2.7 (and the choice of $\gamma$). Thus, we must have $\gamma - t_i \geq \epsilon \left\|\mathbf{w}^{(i)}\right\|$, for all $i \in [K]$. As a result, we obtain

$$\left( \frac{\mathbf{w}^{(j)}}{\|\mathbf{w}^{(j)}\|} \right) \cdot \mathop{\mathbf{E}}_{(\mathbf{x}, y) \sim D} [\mathbf{x}\, \mathbb{1}(y = j)] \gtrsim \epsilon^2 / K.$$

Finally, using the fact that we centered the affine functions defining $f$ by subtracting $\mathbf{w}^{(1)} \cdot \mathbf{x} + t_1$ from each of them, we obtain the desired statement. This completes the proof of Proposition 10.2.4.

□

### 10.2.2 Finding a Relevant Direction

This section contains the proof of Proposition 10.2.8, which states that if our current subspace approximation does not lead to a good enough error, then Algorithm 25 would return at least one vector with non-trivial projection onto the hidden subspace.

**Proposition 10.2.8.** *Let $d, k, K \in \mathbb{Z}_+$, $\epsilon, \delta \in (0,1)$ and $\epsilon' \leq c\epsilon^2/(K^2\sqrt{k})$, for some sufficiently small absolute constant $c > 0$. Let $D$ be a distribution supported on $\mathbb{R}^d \times [K]$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $V$ be a $k$-dimensional subspace of $\mathbb{R}^d$. Let $f$ be a multiclass linear classifier with weight vectors $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)} \in \mathbb{R}^d$ and define $\mathrm{OPT} = \mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y]$. Furthermore, let $h_{\mathcal{S}}$ be a piecewise constant approximation of $D$, with respect to an $\epsilon'$-approximating partition $\mathcal{S}$ defined relative to the subspace $V$ (as in Definitions 10.2.2 and 10.2.3).*

*There exists a sample size $N = d(k/\epsilon')^{O(k)}(K/\epsilon)^{O(1)} \log(1/\delta)$ and a universal constant $C > 0$, such that if $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_{\mathcal{S}}(\mathbf{x}) \neq y] > C \cdot \mathrm{OPT} + \epsilon$, then Algorithm 25 given $N$ samples, runs in $\mathrm{poly}(N)$ time and with probability at least $1 - \delta$ returns a set of unit vectors $\mathcal{E}$ of size $|\mathcal{E}| = \mathrm{poly}(1/\epsilon, K)$ with the following property: there exists a $\mathbf{v} \in \mathcal{E}$ and $i, j \in [K]$, $i \neq j$, such that*

$$\frac{(\mathbf{w}^{(i)} - \mathbf{w}^{(j)})^{\perp V}}{\|(\mathbf{w}^{(i)} - \mathbf{w}^{(j)})^{\perp V}\|} \cdot \mathbf{v} \geq \mathrm{poly}(1/K, \epsilon) \, .$$

Before proving Proposition 10.2.8, we develop the necessary tools for working with multiclass linear classifiers restricted to cubes extending along the orthogonal complement of a subspace. First, we define the notion of a projected classifier onto a set of linear equations (Definition 10.2.9).

The intuition behind this definition is that when we restrict our classifier to sufficiently

small cubes on $V$, its behavior mainly depends on its components in the orthogonal complement of $V$. In other words, the classifier is nearly equivalent to one that fixes its component on $V$ to a specific point. We denote by the *projected classifier* of $f$ with respect to a set of linear equations $\mathcal{G}$ the classifier obtained by fixing the coordinates of $\mathbf{x}^V$ according to $\mathcal{G}$.

**Definition 10.2.9** (Projected Classifier)**.** *Let* $f(\mathbf{x}) = \operatorname{argmax}_{i\in[K]}(\mathbf{w}^{(i)}\cdot\mathbf{x}+t_i)$ *with* $\mathbf{w}^{(i)} \in \mathbb{R}^d, t_i \in \mathbb{R}$ *for* $i \in [K]$*, let* $\mathcal{G} = \{\mathbf{v}^{(1)}\cdot\mathbf{x} = z_1, \ldots, \mathbf{v}^{(k)}\cdot\mathbf{x} = z_k\}$ *be a set of* $k$ *linear equations and let* $V$ *be the subspace spanned by orthonormal vectors* $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)} \in \mathbb{R}^d$*. We denote by* $f_{\mathcal{G}}(\mathbf{x})$ *the linear classifier* $f$ *projected onto* $\mathcal{G}$*, i.e.,* $f_{\mathcal{G}}(\mathbf{x}) = \operatorname{argmax}_{i\in[K]}((\mathbf{w}^{(i)})^{\perp_V}\cdot\mathbf{x} + t_i + t'_{i,\mathcal{G}})$*, where* $t'_{i,\mathcal{G}} = \sum_{j=1}^{k} z_j a_{ij}$ *where* $a_{ij} = \mathbf{w}^{(i)}\cdot\mathbf{v}^{(j)}$*.*

Now we prove Lemma 10.2.10, which states that for any partition of a subspace into sufficiently small cubes, the target function is well-approximated by a piecewise function constructed using a projected classifier for each region. This will be our main supporting result for proving Proposition 10.2.8, as it allows us to reduce the problem to instances on the orthogonal complement and use Proposition 10.2.4.

**Lemma 10.2.10** (Projection Lemma)**.** *Let* $d, K, k \in \mathbb{Z}_+, \epsilon, \epsilon' > 0$*,* $f : \mathbb{R}^d \to [K]$ *be a multiclass linear classifier,* $V$ *be a* $k$*-dimensional subspace of* $\mathbb{R}^d, k \geq 1$*. Let* $\mathcal{S}$ *be an* $\epsilon'$*-approximating partition of* $\mathbb{R}^d$ *with respect to* $V$ *as in Definition 10.2.2, with* $\epsilon' \leq \epsilon^2/(K^2\sqrt{k})$*. Then there exists a set* $\mathcal{T} \subseteq \mathcal{S}$ *such that* $\sum_{S\in\mathcal{T}} \mathbf{Pr}_{\mathcal{N}(\mathbf{0},\mathbf{I})}[S] \geq 1-\epsilon$ *and for any* $S \in \mathcal{T}$*, it holds that*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) \neq f_{\mathcal{G}_S}(\mathbf{x}) \mid \mathbf{x} \in S] \leq \epsilon\ ,$$

*where* $f_{\mathcal{G}_S}$ *denotes the projected classifier onto a set of linear equations* $\mathcal{G}_S$ *that is*

*associated with $S$, as defined in Definition 10.2.9.*

*Proof.* Let $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)} \in \mathbb{R}^d$ and $t_1, \ldots, t_K \in \mathbb{R}$ such that $f(\mathbf{x}) = \text{argmax}_{i \in [K]}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i)$. Let $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}$ be an orthonormal basis of $V$. Fix a set of the partition $S \in \mathcal{S}$ and denote by $f_{\mathcal{G}_S} = \text{argmax}((\mathbf{w}^{(i)})^{\perp V} \cdot \mathbf{x} + t_i + t'_i)$ (see Definition 10.2.9 for the definition of $t'_i$) the projected classifier that corresponds to the set of linear equations $\mathcal{G}_S = \{\mathbf{v}^{(1)} \cdot \mathbf{x} = z_1, \ldots, \mathbf{v}^{(k)} \cdot \mathbf{x} = z_k\}$ (see Definition 10.2.2 for the definition of $\mathcal{G}_S$). Let $\mathbf{w}^{(i,j)} = \mathbf{w}^{(i)} - \mathbf{w}^{(j)}$, $t_{i,j} = t_i - t_j$ and $t'_{i,j} = t'_i - t'_j$. By the union bound, we have that

$$\mathbf{Pr}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[f(\mathbf{x}) \neq f_{\mathcal{G}_S}(\mathbf{x}) \mid \mathbf{x} \in S] \leq$$
$$\textstyle\sum_{i \neq j} \mathbf{Pr}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[\text{sign}(\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}) \neq \text{sign}((\mathbf{w}^{(i,j)})^{\perp V} \cdot \mathbf{x} + t_{i,j} + t'_{i,j}) \mid \mathbf{x} \in S] \,.$$

Fix $i, j \in [K]$ with $i \neq j$. Notice that for a sign change to occur, i.e., $\text{sign}(\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}) \neq \text{sign}((\mathbf{w}^{(i,j)})^{\perp V} \cdot \mathbf{x} + t_{i,j} + t'_{i,j})$ it is necessary that

$$\left|\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}\right| \leq \left|\mathbf{w}^{(i,j)} \cdot \mathbf{x}^V - t'_{i,j}\right|.$$

To see that, note that we can add and subtract $\mathbf{w}^{(i,j)} \cdot \mathbf{x}^V$, and we get that if for some $\mathbf{x} \in S$ we have $\text{sign}(\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}) \neq \text{sign}((\mathbf{w}^{(i,j)})^{\perp V} \cdot \mathbf{x} + t_{i,j} + t'_{i,j})$ we also have, $\text{sign}(\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}) \neq \text{sign}((\mathbf{w}^{(i,j)}) \cdot \mathbf{x} + t_{i,j} + t'_{i,j} - \mathbf{w}^{(i,j)} \cdot \mathbf{x}^V)$, hence it should hold that $\left|\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}\right| \leq \left|\mathbf{w}^{(i,j)} \cdot \mathbf{x}^V - t'_{i,j}\right|$. Therefore, it suffices to bound the probability that $\mathbf{Pr}[|\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}| \leq |\mathbf{w}^{(i,j)} \cdot \mathbf{x}^V - t'_{i,j}| \mid \mathbf{x} \in S]$. By the definition of $f_{\mathcal{G}_S}$, if $\mathbf{x} \in S$ then for any $i, j \in [K]$ we have

$$\left|\mathbf{w}^{(i,j)} \cdot \mathbf{x}^V - t'_{i,j}\right| = \left|\sum_{l=1}^{k} (\mathbf{w}^{(i,j)} \cdot \mathbf{v}^{(l)})(\mathbf{x} \cdot \mathbf{v}^{(l)} - z_l)\right|.$$

This equality holds because by definition the projection of $\mathbf{x}$ onto the subspace $V$ with orthonormal basis $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}\}$ is given by $\mathbf{x}^V = \sum_{l=1}^{k} (\mathbf{x} \cdot \mathbf{v}^{(l)}) \mathbf{v}^{(l)}$, and the term $t'_{i,j}$ is defined as $t'_{i,j} = \sum_{l=1}^{k} z_l \, (\mathbf{w}^{(i,j)} \cdot \mathbf{v}^{(l)})$. Next, by applying the triangle inequality we have

$$\left| \sum_{l=1}^{k} (\mathbf{w}^{(i,j)} \cdot \mathbf{v}^{(l)})(\mathbf{x} \cdot \mathbf{v}^{(l)} - z_l) \right| \leq \sum_{l=1}^{k} \left| \mathbf{w}^{(i,j)} \cdot \mathbf{v}^{(l)} \right| \left| \mathbf{x} \cdot \mathbf{v}^{(l)} - z_l \right|.$$

By the definition of the partition $S$, if $\mathbf{x} \in S$ then for every $l \in \{1, \ldots, k\}$ the projection of $\mathbf{x}$ onto the direction $\mathbf{v}^{(l)}$ is $\epsilon'$-close to the corresponding value $z_l$, i.e., $\left| \mathbf{x} \cdot \mathbf{v}^{(l)} - z_l \right| \leq \epsilon'$. Therefore,

$$\sum_{l=1}^{k} \left| \mathbf{w}^{(i,j)} \cdot \mathbf{v}^{(l)} \right| \left| \mathbf{x} \cdot \mathbf{v}^{(l)} - z_l \right| \leq \sum_{l=1}^{k} \left| \mathbf{w}^{(i,j)} \cdot \mathbf{v}^{(l)} \right| \epsilon'.$$

Finally, by applying the Cauchy-Schwarz inequality, we obtain

$$\sum_{l=1}^{k} \left| \mathbf{w}^{(i,j)} \cdot \mathbf{v}^{(l)} \right| \leq \sqrt{k} \sqrt{\sum_{l=1}^{k} \left( \mathbf{w}^{(i,j)} \cdot \mathbf{v}^{(l)} \right)^2} = \sqrt{k} \, \| (\mathbf{w}^{(i,j)})^V \| \, .$$

Combining the above, we have $\left| \mathbf{w}^{(i,j)} \cdot \mathbf{x}^V - t'_{i,j} \right| \leq \| (\mathbf{w}^{(i,j)})^V \| \sqrt{k} \, \epsilon'$. Averaging over $S \in \mathcal{S}$ gives us

$$\begin{aligned}
&\sum_{S \in \mathcal{S}} \mathbf{Pr}[f(\mathbf{x}) \neq f_{\mathcal{G}_S}(\mathbf{x}) \mid \mathbf{x} \in S] \, \mathbf{Pr}[S] \\
&\qquad \leq \sum_{i \neq j \in [K]} \sum_{S \in \mathcal{S}} \mathbf{Pr}[|\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}| \leq \| (\mathbf{w}^{(i,j)})^V \| \sqrt{k} \epsilon' \mid \mathbf{x} \in S] \, \mathbf{Pr}[S] \\
&\qquad \leq \sum_{i \neq j \in [K]} \mathbf{Pr}[|\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}| \leq \| (\mathbf{w}^{(i,j)})^V \| \sqrt{k} \epsilon'] \\
&\qquad \leq \sum_{i \neq j \in [K]} \mathbf{Pr}[|\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}| \leq \| \mathbf{w}^{(i,j)} \| \sqrt{k} \epsilon'] \\
&\qquad \leq \sum_{i \neq j \in [K]} \sqrt{k} \epsilon' \leq K^2 \sqrt{k} \epsilon' \, ,
\end{aligned}$$

where we used the anti-concentration of the Gaussian distribution (see Fact 10.1.7). Finally, since $\epsilon' \leq \epsilon^2/(K^2\sqrt{k})$ it holds that $\sum_{S\in\mathcal{S}} \mathbf{Pr}[f(\mathbf{x}) \neq f_{\mathcal{G}_S}(\mathbf{x}) \mid \mathbf{x} \in S]\,\mathbf{Pr}[S] \leq \epsilon^2$. To conclude the proof, let $\mathcal{S}_{bad}$ be the set of $S \in \mathcal{S}$ such that if $S \in \mathcal{S}_{bad}$ then $\mathbf{Pr}[f(\mathbf{x}) \neq f_{\mathcal{G}_S}(\mathbf{x}) \mid \mathbf{x} \in S] > \epsilon$. Then, we have that

$$\begin{aligned}\epsilon^2 \geq \sum_{S\in\mathcal{S}} \mathbf{Pr}[f(\mathbf{x}) \neq f_{\mathcal{G}_S}(\mathbf{x}) \mid \mathbf{x} \in S]\,\mathbf{Pr}[S] &\geq \sum_{S\in\mathcal{S}_{bad}} \mathbf{Pr}[f(\mathbf{x}) \neq f_{\mathcal{G}_S}(\mathbf{x}) \mid \mathbf{x} \in S]\,\mathbf{Pr}[S] \\ &> \epsilon \sum_{S\in\mathcal{S}_{bad}} \mathbf{Pr}[S]\,.\end{aligned}$$

Hence, $\sum_{S\in\mathcal{S}_{bad}} \mathbf{Pr}[S] \leq \epsilon$. By taking $\mathcal{T} = \mathcal{S}/\mathcal{S}_{bad}$ we conclude the proof. □

We now proceed with the proof of Proposition 10.2.8.

*Proof of Proposition 10.2.8.* Let $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)} \in \mathbb{R}^d$ and $t_1, \ldots, t_K \in \mathbb{R}$ be the parameters defining $f$, i.e., $f(\mathbf{x}) = \mathrm{argmax}_{i\in[K]}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i)$. For brevity, define the difference vectors $\mathbf{w}^{(i,j)} = \mathbf{w}^{(i)} - \mathbf{w}^{(j)}$, and let $(\widehat{\mathbf{w}}^{(i,j)})^{\perp V} = (\mathbf{w}^{(i,j)})^{\perp V} / \left\|(\mathbf{w}^{(i,j)})^{\perp V}\right\|$ denote the normalized projection of $\mathbf{w}^{(i,j)}$ onto the orthogonal complement of a given subspace $V$. Also, define $W = \mathrm{span}(\mathbf{w}_1, \ldots, \mathbf{w}_K)$, and let $C$ be a constant as specified in Proposition 10.2.4. For each cube $S \in \mathcal{S}$ (as defined in Definition 10.2.2), let $\mathcal{G}_S$ be the corresponding set of linear equations, and let $f_{\mathcal{G}_S}$ be the projection of $f$ onto $\mathcal{G}_S$ (see Definition 10.2.9). Furthermore, for every $S \in \mathcal{S}$ and $i \in [K]$, define the population moment $\mathbf{u}^{(S,i)} = \mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{x}\mathbb{1}(y = i) \mid \mathbf{x} \in S]$,and let $\widehat{\mathbf{u}}^{(S,i)} = \mathbf{E}_{(\mathbf{x},y)\sim \widehat{D}}[\mathbf{x}\mathbb{1}(y = i) \mid \mathbf{x} \in S]$ its empirical counterpart (see Line 8). Also, denote by $\widehat{\mathbf{U}}$ the matrix computed at Line 11 of Algorithm 25.

We first show that if the current subspace $V$ does not lead to a good approximation of $f$, meaning that any function defined only on $V$ gets error at least $\Omega(\mathrm{OPT} + \epsilon)$ with

the labels $y$, then when we partition the space into small cubes with respect to $V$, on a nontrivial fraction of them $f$ is far from being constant.

**Claim 10.2.11** (Large Fraction of Good Cubes)**.** *If* $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_{\mathcal{S}}(\mathbf{x}) \neq y] > (C+1)\mathrm{OPT} + 5\epsilon$*, then there exists a subcollection* $\mathcal{T} \subseteq \mathcal{S}$ *satisfying* $\sum_{S\in\mathcal{T}} \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[S] \geq \epsilon$*; and for every* $S \in \mathcal{T}$ *we have* $\min_{i\in[K]} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[f_{\mathcal{G}_S}(\mathbf{x}) \neq i \mid \mathbf{x} \in S] > C\,\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f_{\mathcal{G}_S}(\mathbf{x}) \neq y \mid \mathbf{x} \in S] + \epsilon$.

*Proof of Claim 10.2.11.* Define $\mathcal{T}$ to be the collection of $S \subseteq \mathcal{S}$ such that $\min_{i\in[K]} \mathbf{Pr}[i \neq y \mid \mathbf{x} \in S] > (C+1)\,\mathbf{Pr}[f(\mathbf{x}) \neq y \mid \mathbf{x} \in S] + 2\epsilon$. Assume for the sake of contradiction that $\sum_{S\in\mathcal{T}} \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[S] \leq 2\epsilon$. We have that

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[h_{\mathcal{S}}(\mathbf{x}) \neq y] &\leq \sum_{S\in\mathcal{S}} \min_{\widehat{y}\in[K]} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\widehat{y} \neq y \mid \mathbf{x} \in S] \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[S] + \epsilon \\
&\leq \sum_{S\notin\mathcal{T}} \left( (C+1) \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y \mid \mathbf{x} \in S] + 2\epsilon \right) \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[S] + 3\epsilon \\
&\leq (C+1)\mathrm{OPT} + 5\epsilon\,,
\end{aligned}$$

where in the first line we used the fact (see Fact 10.1.7) that the mass of points outside the discretization is at most $\epsilon$, in the second line we applied our definition of $\mathcal{T}$ and in the last inequality we used that the total error of $f$ is OPT. This contradicts our hypothesis that $\mathbf{Pr}[h_{\mathcal{S}}(\mathbf{x}) \neq y] > (C+1)\mathrm{OPT} + 5\epsilon$. Hence, it must be that $\sum_{S\in\mathcal{T}} \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[S] > 2\epsilon$.

Next, recall that the cube width $\epsilon'$ is chosen to be less than $c\epsilon^2/(K^2\sqrt{k})$ for a sufficiently small constant $c$. Taking $c < 1/(C+1)$, we can apply Lemma 10.2.10 and get that there exists a collection $\mathcal{T}' \subseteq \mathcal{S}$ with $\sum_{S\in\mathcal{T}'} \mathbf{Pr}[S] \geq 1 - \epsilon$, such that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) \neq f_{\mathcal{G}_S}(\mathbf{x}) \mid \mathbf{x} \in S] \leq \epsilon/(C+1)$. Note that for $\mathcal{T}'' = \mathcal{T}' \cap \mathcal{T}$, it must hold that $\sum_{S\in\mathcal{T}''} \mathbf{Pr}[s] \geq \epsilon$. For all $S \in \mathcal{T}''$, it holds that $\min_{i\in[K]} \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[i \neq y \mid \mathbf{x} \in$

$S] > (C+1)\,\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f_{\mathcal{G}_S}(x) \neq y \mid \mathbf{x} \in S] + \epsilon$. Since,

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\Big[i \neq y \mid \mathbf{x} \in S\Big] \leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\Big[f_{\mathcal{G}_S}(\mathbf{x}) \neq i \mid \mathbf{x} \in S\Big] + \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\Big[f_{\mathcal{G}_S}(\mathbf{x}) \neq y \mid \mathbf{x} \in S\Big]\,,$$

we get that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f_{\mathcal{G}_S}(\mathbf{x}) \neq i \mid \mathbf{x} \in S] > C\,\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f_{\mathcal{G}_S}(\mathbf{x}) \neq y \mid \mathbf{x} \in S] + \epsilon$ for all $S \in \mathcal{T}''$, which completes the proof. □

In the next claim, we show that if there exists a cube where the target function is far from being constant, then we can find a vector that correlates with the unknown parameters.

**Claim 10.2.12** (Correlating Moments)**.** *Let $S \in \mathcal{S}$ such that $\min_{i\in[K]} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[f_{\mathcal{G}_S}(\mathbf{x}) \neq i \mid \mathbf{x} \in S] > C\,\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f_{\mathcal{G}_S}(\mathbf{x}) \neq y \mid \mathbf{x} \in S] + \epsilon$. Then there exists $i, j, l \in [K]$ with $i \neq j$ such that $(\widehat{\mathbf{w}}^{(i,j)})^{\perp V} \cdot \mathbf{u}^{(S,l)} \geq \mathrm{poly}(1/K, \epsilon)$.*

*Proof of Claim 10.2.12.* First, notice that the output label of $f_{\mathcal{G}_S}$ is only determined by the projection onto the orthogonal complement of $V$, i.e.,

$$f_{\mathcal{G}_S}(\mathbf{x}) = \mathop{\mathrm{argmax}}_{i\in[K]}((\mathbf{w}^{(i)})^{\perp V} \cdot \mathbf{x} + t_i + t'_{\mathcal{G}_S,i}) = \mathop{\mathrm{argmax}}_{i\in[K]}((\mathbf{w}^{(i)})^{\perp V} \cdot \mathbf{x}^{\perp V} + t_i + t'_{\mathcal{G}_S,i}) = f_{\mathcal{G}_S}(\mathbf{x}^{\perp V})$$

Furthermore, observe that whether $\mathbf{x}$ belongs to $S$ depends only on its projection onto

$V$, hence without loss of generality it can written as $\mathbb{1}(\mathbf{x}^V \in S)$. Thus, we have that

$$\begin{aligned}
&\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f_{\mathcal{G}_S}(\mathbf{x}) \neq y \mid \mathbf{x} \in S] = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f_{\mathcal{G}_S}(\mathbf{x}^{\perp V}) \neq y \mid \mathbf{x} \in S] \\
&= \sum_{y=1}^{K} \int_{\mathbb{R}^d} \mathbb{1}(f_{\mathcal{G}_S}(\mathbf{x}^{\perp V}) \neq y)\mathbb{1}(\mathbf{x}^V \in S) D(\mathbf{x}, y) d\mathbf{x} / \mathbf{Pr}[\mathbf{x}^V \in S] \\
&= \sum_{y=1}^{K} \int_{V^\perp} \mathbb{1}(f_{\mathcal{G}_S}(\mathbf{x}^{\perp V}) \neq y) \left( \frac{1}{\mathbf{Pr}[\mathbf{x}^V \in S]} \int_V D(\mathbf{x}, y)\mathbb{1}(\mathbf{x}^V \in S) d\mathbf{x}^V \right) d\mathbf{x}^{\perp V} \\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[f_{\mathcal{G}_S}(\mathbf{x}) \neq y] \,,
\end{aligned}$$

where $D^S_{V^\perp}$ denotes the marginal distribution on $V^\perp$ obtained by averaging $D$ over $V$ while conditioning on the event $\mathbf{x}^V \in S$. Similarly, we have $\mathbf{Pr}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[f_{\mathcal{G}_S}(\mathbf{x}) \neq i] > C\,\mathbf{Pr}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[f_{\mathcal{G}_S}(\mathbf{x}) \neq y] + \epsilon$ for all $i \in [K]$. Moreover, by the properties of the standard Gaussian distribution, the $\mathbf{x}$-marginal of $D^S_{V^\perp}$ is a standard Gaussian.

As a result, by applying Proposition 10.2.4 for $f_{\mathcal{G}_S}$, we have that there exist $i, j \in [K], i \neq j$ such that $(\widehat{\mathbf{w}}^{(i,j)})^{\perp V} \cdot \mathbf{E}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[\mathbf{x}\mathbb{1}(y = i)] \geq \mathrm{poly}(1/K, \epsilon)$. Furthermore, note that by averaging over $V$ conditioning on $S$ we have that $\mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{x}^{\perp V}\mathbb{1}(y = i) \mid \mathbf{x} \in S] = \mathbf{E}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[\mathbf{x}\mathbb{1}(y = i)]$. Therefore, there exist $i, j \in [K], i \neq j$, such that $(\widehat{\mathbf{w}}^{(i,j)})^{\perp V} \cdot \mathbf{u}^{(S,l)} \geq \mathrm{poly}(1/K, \epsilon)$ for some $l \in [K]$.

□

In the next claim, we show that by taking enough samples, the empirical moments are close to the population ones.

**Claim 10.2.13** (Concentration of Empirical Distribution)**.** *Let $p(\epsilon, 1/K)$ be a sufficiently small polynomial in $\epsilon$ and $1/K$. If $N = d(k/\epsilon')^{Ck}(K/\epsilon)^C \log(1/\delta)$ for a sufficiently large universal constant $C > 0$, then with probability $1 - \delta$ we have that $\left\|(\widehat{\mathbf{u}}^{(S,i)} - \mathbf{u}^{(S,i)})^{\perp V}\right\| \leq p(\epsilon, 1/K)$ and $\mathbf{Pr}_{\widehat{D}}[S] \geq \mathbf{Pr}_D[S]/2$ for all $S \in \mathcal{S}, i \in [K]$.*

*Proof of Claim 10.2.13.* First note that by the definition of an approximating partition (see Definition 10.2.2) and Fact 10.1.7, for all $S \in \mathcal{S}$ we have that $\mathbf{Pr}_D[S] = (\epsilon'/k)^{\Omega(k)}$ and $|\mathcal{S}| = (k/\epsilon')^{O(k)}$. Therefore, by the union bound and Hoeffding's inequality, if $N \geq (k/\epsilon')^{Ck} \log(|\mathcal{S}|/\delta) = (k/\epsilon')^{O(k)} \log(1/\delta)$ for a sufficiently large absolute constant $C > 0$, with probability at least $1 - \delta$ we have that $|\mathbf{Pr}_{\widehat{D}}[S] - \mathbf{Pr}_D[S]| \leq \mathbf{Pr}_D[S]/2$ for all $S \in \mathcal{S}$. Hence, $\mathbf{Pr}_{\widehat{D}}[S] \geq \mathbf{Pr}_D[S]/2$, i.e., the number of samples that fall in each set $S \in \mathcal{S}$ is at least $N \mathbf{Pr}_D[S]/2 = N(\epsilon'/k)^{\Omega(k)}$.

Furthermore, note that by the definitions of $\mathbf{u}^{(S,i)}$ and $\widehat{\mathbf{u}}^{(S,i)}$ it holds that

$$(\widehat{\mathbf{u}}^{(S,i)} - \mathbf{u}^{(S,i)})^{\perp V} = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\widehat{D}}[\mathbf{x}^{\perp V} \mathbb{1}(y = i) \mid \mathbf{x} \in S] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{x}^{\perp V} \mathbb{1}(y = i) \mid \mathbf{x} \in S]$$

Therefore, to prove our concentration result, it suffices to show that when conditioning on $\mathbf{x} \in S$ the random vector $\mathbf{x}^{\perp V} \mathbb{1}(y = i)$ with $(\mathbf{x}, y) \sim D$ is subgaussian in every unit direction with subgaussian norm bounded by a constant. To see this, note that whether $\mathbf{x}$ belongs in $S$ depends only on the projection of $\mathbf{x}$ onto $V$, and $\mathbf{x}^{\perp V}$ is independent of $\mathbf{x}^V$. Now fix a unit vector $\mathbf{v} \in \mathbb{R}^d$, we have that for any $p \geq 1$

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\left[\left(\mathbf{v} \cdot \mathbf{x}^{\perp V} \mathbb{1}(y = i) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbf{v} \cdot \mathbf{x}^{\perp V} \mathbb{1}(y = i) \mid \mathbf{x} \in S]\right)^p \mid S\right]^{1/p} \\
&\lesssim \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\left[\left(\mathbf{v} \cdot \mathbf{x}^{\perp V} \mathbb{1}(y = i)\right)^p \mid S\right]^{1/p} \\
&\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\left[\left(\mathbf{v} \cdot \mathbf{x}^{\perp V}\right)^p\right]^{1/p} \\
&\leq \mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}[(z)^p]^{1/p} \lesssim \sqrt{p}\,,
\end{aligned}$$

where in the first inequality we used the triangle and Jensen's inequalities, while in the second we exploited the independence of $\mathbf{x}^V$ and $\mathbf{x}^{\perp V}$ and the fact that the indicator

function is bounded by 1. Therefore $\mathbf{x}^{\perp V}\mathbb{1}(y=i)$ is subgaussian. Hence, by the use of a standard covering argument it follows that if the number of samples that falls in each discretization set $S$ is at least $d\mathrm{poly}(K,1/\epsilon)\log(K/\delta)$ for a sufficiently large polynomial, then $\left\|(\widehat{\mathbf{u}}^{(S,i)}-\mathbf{u}^{(S,i)})^{\perp V}\right\| \leq p(\epsilon,1/K)$ for all $i\in[K]$.

Combining the above results along with the union bound gives us that there exists an $N = d(k/\epsilon')^{O(k)}(K/\epsilon)^{O(1)}\log(1/\delta)$ that suffices. □

Applying Claims 10.2.11 and 10.2.12, we have that if $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_{\mathcal{S}}(\mathbf{x})\neq y] > (C+1)\mathrm{OPT}+\epsilon$ then there exists a subset $\mathcal{T}\subseteq\mathcal{S}$ of the $\epsilon'$-approximating partition $\mathcal{S}$ with $\sum_{S\in\mathcal{T}}\mathbf{Pr}_D[S]=\Omega(\epsilon)$ such that for all $S\in\mathcal{T}$ we have $(\widehat{\mathbf{w}}^{(i,j)})^{\perp V}\cdot\mathbf{u}^{(S,l)}\geq\mathrm{poly}(1/K,\epsilon)$ for some $i,j,l\in[K], i\neq j$. Moreover, if $N=d(k/\epsilon')^{C'k}(K/\epsilon)^{C'}\log(1/\delta)$ for a sufficiently large universal constant $C'>0$, using the concentration result of Claim 10.2.13, we have that the empirical moments of each set satisfy the same guarantees, i.e., $(\widehat{\mathbf{w}}^{(i,j)})^{\perp V}\cdot\widehat{\mathbf{u}}^{(S,l)}\geq\mathrm{poly}(1/K,\epsilon)$, for some $i,j,l\in[K], i\neq j$, for all $S\in\mathcal{T}$. Furthermore, note that by Claim 10.2.13 we also have that $\sum_{S\in\mathcal{T}}\mathbf{Pr}_{\widehat{D}}[S]=\Omega(\epsilon)$.

Therefore, there exists a subset $\mathcal{T}'\subseteq\mathcal{T}$ and $i,j\in[K]$ with $\sum_{S\in\mathcal{T}'}\mathbf{Pr}_{\widehat{D}}[S]\geq\Omega(\epsilon/K^2)$ such that for all $S\in\mathcal{T}'$ we have that $(\widehat{\mathbf{w}}^{(i,j)})^{\perp V}\cdot\widehat{\mathbf{u}}^{(S,l)}\geq\mathrm{poly}(1/K,\epsilon)$ for some $l\in[K]$. Hence, for some $i\neq j\in[K]$ we have that

$$\left((\widehat{\mathbf{w}}^{(i,j)})^{\perp V}\right)^{\mathsf{T}}\widehat{\mathbf{U}}(\widehat{\mathbf{w}}^{(i,j)})^{\perp V}\geq\frac{\epsilon}{K^2}\mathrm{poly}(\epsilon,1/K)=\mathrm{poly}(1/K,\epsilon)\,,$$

where we used the fact that $\mathbf{x}\mathbf{x}^{\mathsf{T}}$ is PSD for any $\mathbf{x}\in\mathbb{R}^d$. Since the set $S$ depends only on the projection $\mathbf{x}^V$, we have $(\mathbf{u}^{(S,i)})^{\perp V}=\mathbf{E}_{\mathbf{x}^{\perp V}}[\mathbf{x}^{\perp V}\,\mathbf{E}_{\mathbf{x}^V}[\mathbb{1}(y=i)\mid\mathbf{x}\in S]]$. Therefore,

we can bound the norm of $(\mathbf{u}^{(S,i)})^{\perp V}$ as follows:

$$\left\|(\mathbf{u}^{(S,i)})^{\perp V}\right\| \leq \max_{\|\mathbf{u}\|=1} \mathop{\mathbf{E}}_{\mathbf{x}^{\perp V}}[\mathbf{u}\cdot\mathbf{x}^{\perp V} \mathop{\mathbf{E}}_{\mathbf{x}^{V}}[\mathbb{1}(y=i) \mid \mathbf{x}\in S]]$$
$$\leq \max_{\|\mathbf{u}\|=1} (\mathop{\mathbf{E}}_{\mathbf{x}^{\perp V}}[(\mathbf{u}\cdot\mathbf{x}^{\perp V})^2] \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[\mathbb{1}(y=i) \mid \mathbf{x}\in S])^{1/2} \leq 1\,,$$

where we used the Cauchy-Schwarz inequality and the dual characterization of the $\ell_2$ norm. In the last inequality we used that $\mathbf{x}^{\perp V}$ is a standard Gaussian, and henceforth the first expectation is at most 1, and the second expectation is at most 1 as well because $\mathbb{1}(y=i)$ is bounded by 1. By the triangle inequality, it follows that

$$\left\|\widehat{\mathbf{U}}\right\|_F \leq \sum_{S\in\mathcal{S},i\in[K]} \|(\widehat{\mathbf{u}}^{(S,i)})^{\perp V}\| \mathop{\mathbf{Pr}}_{\widehat{D}}[S] \leq \sum_{S\in\mathcal{S},i\in[K]} (1+\mathrm{poly}(\epsilon,1/K)) \mathop{\mathbf{Pr}}_{\widehat{D}}[S] \leq 2K\,.$$

Hence, by Claim 10.5.1, there exists a unit eigenvector $\mathbf{v}$ of $\widehat{\mathbf{U}}$ with eigenvalue at least $\mathrm{poly}(\epsilon/K)$ (for a sufficiently small polynomial), such that $(\widehat{\mathbf{w}}^{(i,j)})^{\perp V}\cdot\mathbf{v} \geq \mathrm{poly}(1/K,\epsilon)$, for some $i\neq j\in[K]$. Moreover, the number of such eigenvectors is at most $2K/\mathrm{poly}(1/K,\epsilon) = \mathrm{poly}(K,1/\epsilon)$. This completes the proof of Proposition 10.2.8. □

Before proving our main result of this section, Theorem 10.2.1, we first show that if we have learned one direction of our target model to sufficiently high accuracy, then we may, without loss of generality, restrict the model to the remaining directions. Specifically, we establish the following:

**Lemma 10.2.14.** *Let $K\geq 2, \epsilon\in(0,1)$ and $f(\mathbf{x}) = \mathrm{argmax}_{i\in[K]}(\mathbf{w}^{(i)}\cdot\mathbf{x}+t_i)$ with $\mathbf{w}^{(i)}\in\mathbb{R}^d, t_i\in\mathbb{R}$ for $i\in[K]$ be a multiclass linear classifier. Let $V$ be a non-empty subspace of $\mathbb{R}^d$ such that for all $i,j\in[K]$ with $i\neq j$ it holds that $\|(\mathbf{w}^{(i)}-\mathbf{w}^{(j)})^V\| \leq$*

$(\epsilon^2/K^4)\|\mathbf{w}^{(i)} - \mathbf{w}^{(j)}\|$. *Then, it holds that*

$$\mathbf{Pr}[f(\mathbf{x}) \neq f^{\perp V}(\mathbf{x})] \lesssim \epsilon \ ,$$

*where* $f^{\perp V}(\mathbf{x}) = \mathrm{argmax}((\mathbf{w}^{(i)})^{\perp V} \cdot \mathbf{x} + t_i)$.

*Proof.* For simplicity, let $\mathbf{w}^{(i,j)} = \mathbf{w}^{(i)} - \mathbf{w}^{(j)}$ and $t_{i,j} = t_i - t_j$. Using the union bound inequality, we have that

$$\mathbf{Pr}[f(\mathbf{x}) \neq f^{\perp V}(\mathbf{x})] \leq \sum_{i,j \in [K]} \mathbf{Pr}[\mathrm{sign}(\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}) \neq \mathrm{sign}((\mathbf{w}^{(i,j)})^{\perp V} \cdot \mathbf{x} + t_{i,j})] \ .$$

We can write $\mathbf{w}^{(i,j)} = (\mathbf{w}^{(i,j)})^V + (\mathbf{w}^{(i,j)})^{\perp V}$. According to the assumptions of Lemma 10.2.14, it holds that $\|(\mathbf{w}^{(i,j)})^V\| \leq (\epsilon^2/K^4)\|\mathbf{w}^{(i,j)}\|$ and $\|(\mathbf{w}^{(i,j)})^{\perp V}\| \geq \sqrt{1 - (\epsilon^4/K^8)}\|\mathbf{w}^{(i,j)}\|$. Therefore, using the independence of any two orthogonal directions of a standard Gaussian, along with Fact 10.1.7, we have that

$$\begin{aligned}
&\mathbf{Pr}[\mathrm{sign}(\mathbf{w}^{(i,j)} \cdot \mathbf{x} + t_{i,j}) \neq \mathrm{sign}((\mathbf{w}^{(i,j)})^{\perp V} \cdot \mathbf{x} + t_{i,j})] \\
&\leq \mathbf{Pr}[|(\mathbf{w}^{(i,j)})^V \cdot \mathbf{x}| \geq |(\mathbf{w}^{(i,j)})^{\perp V} \cdot \mathbf{x} + t_{i,j}|] \\
&\leq \mathop{\mathbf{Pr}}_{z \sim \mathcal{N}(0,1)}[|z| \geq \frac{\|(\mathbf{w}^{(i,j)})^{\perp V}\|(\epsilon/K^2)}{\|(\mathbf{w}^{(i,j)})^V\|}] + \epsilon/K^2 \\
&\leq \mathop{\mathbf{Pr}}_{z \sim \mathcal{N}(0,1)}[|z| \geq \frac{\sqrt{1 - \epsilon^4/K^8}}{\epsilon/K^2}] + \epsilon/K^2 \leq 2\epsilon/K^2 \ .
\end{aligned}$$

Therefore, we have that

$$\mathbf{Pr}[f(\mathbf{x}) \neq f^{\perp V}(\mathbf{x})] \leq 2\epsilon \ .$$

□

### 10.2.3 Proof of Theorem 10.2.1

In this section, given Proposition 10.2.8, we proceed to the proof of Theorem 10.2.1. In particular, we show that Algorithm 24 returns, with probability at least $1-\delta$, a hypothesis $h$ whose 0-1 error is $O(\mathrm{OPT})+\epsilon$.

*Proof of Theorem 10.2.1.* Let $\mathbf{w}^{*(1)},\ldots,\mathbf{w}^{*(K)} \in \mathbb{R}^d$ and $t_1^*,\ldots,t_K^* \in \mathbb{R}$ be the parameters of the affine functions of an optimal classifier $f^*$, i.e., $f^*(\mathbf{x}) = \mathrm{argmax}_{i\in[K]}(\mathbf{w}^{*(i)} \cdot \mathbf{x} + t_i^*)$ and denote by $W^* = \mathrm{span}(\mathbf{w}^{*(1)},\ldots,\mathbf{w}^{*(K)})$. Let $C > 0$ be the constant that satisfies the statement of Proposition 10.2.8. The algorithm maintains a list of vectors $L_t$ (updated at Line 8), and define $V_t = \mathrm{span}(L_t)$, $\dim(V_t) = k_t$. We set $\epsilon' = \mathrm{poly}(\epsilon, 1/K)$, for a sufficiently small polynomial (Line 4) which will be quantified later, and for each $t \in [T]$ let $\mathcal{S}_t$ be an arbitrary $\epsilon'$-approximating partitions with respect to $V_t$ (see Definition 10.2.2). Let $h_t : \mathbb{R}^d \to [K]$ be the piecewise constant classifier defined by $h_t = h_{\mathcal{S}_t}$ as in Definition 10.2.3, for the distribution $D$.

To prove the correctness of Algorithm 24, we need to show that if $h_t$ has significant error, then the algorithm improves its current approximation $V_t$ to $W^*$. To quantify this improvement, we define the potential function

$$\Phi_t = \sum_{i=1}^{K} \left\| (\mathbf{b}^{(i)})^{\perp V_t} \right\|^2 ,$$

where $\mathbf{b}^{(1)},\ldots,\mathbf{b}^{(K)}$ is an orthonormal basis of $W^*$.

Our first step is to show that as long as the algorithm has not terminated, then with high probability we can find a weight vector that correlates non-trivially with the unknown subspace $W^*$.

**Claim 10.2.15** (Large Error Implies Good Correlation)**.** *Suppose that at iteration $t$*

*the classifier $h_t$ satisfies $\mathbf{Pr}_{(x,y)\sim D}[h_t(\mathbf{x}) \neq y] > C \cdot \mathrm{OPT} + 2\epsilon$, and that $\epsilon'$ is bounded as $\epsilon' \leq c\epsilon^2/(K^2\sqrt{k_t})$, for a sufficiently small absolute constant $c > 0$. Then there exists sample size, $N \geq d(k_t/\epsilon')^{O(k_t)}(K/\epsilon)^{O(1)} \log(1/\delta)$, such that with probability at least $1-\delta$ there exists unit vectors $\mathbf{v}^{(t)} \in V_{t+1}$ and $\mathbf{w} \in W^*$ such that $\mathbf{w} \cdot (\mathbf{v}^{(t)})^{\perp V_t} \geq \mathrm{poly}(1/K, \epsilon)$.*

*Proof of Claim 10.2.15.* Let $\mathcal{E}_{t+1} = \{\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k')}\}$ be the set of unit vectors returned by Algorithm 25 at iteration $t+1$. For the sake of contradiction assume that for every $\mathbf{u} \in \mathcal{E}_{t+1}$ and every unit vector $\mathbf{w} \in W^*$, it holds that $|\mathbf{u} \cdot \mathbf{w}| \leq \epsilon^2/(C^2K^4)$.

By Lemma 10.2.14, there exists a classifier $f(\mathbf{x}) = \mathrm{argmax}_{i\in[K]}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i^*)$ such that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) \neq f^*(\mathbf{x})] \leq \epsilon/C$, and for every $i \in [K]$, the vector $\mathbf{w}^{(i)}$ is defined as $\mathbf{w}^{(i)} = (\mathbf{w}^{*(i)})^{\perp U}$, where $U = \mathrm{span}(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k')})$.

Since $f$ approximates $f^*$ up to error $\epsilon/C$, we have $\mathbf{Pr}[f(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon/C$. Thus, the assumption of Claim 10.2.15 that $\mathbf{Pr}_{(x,y)\sim D}[h_t(\mathbf{x}) \neq y] > C \cdot \mathrm{OPT} + 2\epsilon$ implies that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_t(\mathbf{x}) \neq y] > C\,\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] + \epsilon$. Therefore, by applying Proposition 10.2.8 for $f$, we deduce that the returned set $\mathcal{E}_{t+1}$ should contain a vector $\mathbf{u}^{(l)} \in \mathcal{E}_{t+1}$ so that for some indices $i, j \in [K]$, $i \neq j$, it holds $\mathbf{u}^{(l)} \cdot \mathbf{w}^{(i,j)} \geq \mathrm{poly}(1/K, \epsilon) \left\|(\mathbf{w}^{(i,j)})^{\perp V_t}\right\|$, where $\mathbf{w}^{(i,j)} \stackrel{\mathrm{def}}{=} \mathbf{w}^{(i)} - \mathbf{w}^{(j)}$ with $(\mathbf{w}^{(i,j)})^{\perp V_t} \neq \mathbf{0}$. However, since $\mathbf{w}^{(i,j)} \in U^{\perp}$ it should hold that $\mathbf{w}^{(i,j)} \cdot \mathbf{u}^{(l)} = 0$, which is a contradiction. Therefore, there exists $\mathbf{u} \in \mathcal{E}_{t+1}$ and unit vector $\mathbf{w} \in W^*$, such that $|\mathbf{u} \cdot \mathbf{w}| \geq \epsilon^2/(C^2K^4)$, which concludes the proof. □

Our next claim quantifies the decrease of the potential function $\Phi_t$ between the iterations $t$ and $t+1$.

**Claim 10.2.16** (Potential Decrease)**.** *Let $\beta \geq 0$. If there exists unit vector $\mathbf{v}^{(t)} \in V_{t+1}$ and unit vector $\mathbf{w} \in W^*$ such that*

$$\mathbf{w} \cdot (\mathbf{v}^{(t)})^{\perp V_t} \geq \beta \,,$$

*then* $\Phi_{t+1} \leq \Phi_t - \beta^2$.

*Proof of Claim 10.2.16.* First, note that since $\{\mathbf{b}^{(i)}\}_{i=1}^K$ is an orthonormal basis for $W^*$, any vector $\mathbf{w} \in W^*$ can be written as $\sum_{i=1}^K (\mathbf{b}^{(i)} \cdot \mathbf{w})\mathbf{b}^{(i)}$. Thus, we can write $\mathbf{w} \cdot (\mathbf{v}^{(t)})^{\perp V_t} \geq \beta$ as $\sum_{i=1}^K (\mathbf{b}^{(i)} \cdot \mathbf{w})(\mathbf{b}^{(i)} \cdot (\mathbf{v}^{(t)})^{\perp V_t}) \geq \beta$. Recall that our potential function is defined as $\Phi_t = \sum_{i=1}^K \left\|(\mathbf{b}^{(i)})^{\perp V_t}\right\|^2$. Hence, by the Pythagorean theorem, we have that

$$\Phi_{t+1} = \sum_{i=1}^K \|(\mathbf{b}^{(i)})^{\perp V_t}\|^2 - \|(\mathbf{b}^{(i)})^{V_{t+1} \cap \perp V_t}\|^2 = \Phi_t - \sum_{i=1}^K \|(\mathbf{b}^{(i)})^{V_{t+1} \cap \perp V_t}\|^2 \,.$$

Using the dual characterization of the norm, we can further bound

$$\sum_{i=1}^K \|(\mathbf{b}^{(i)})^{V_{t+1} \cap \perp V_t}\|^2 \geq \sum_{i=1}^K \left|\mathbf{b}^{(i)} \cdot (\mathbf{v}^{(t)})^{\perp V_t}\right|^2$$

Note that from Cauchy-Schwarz inequality we have that

$$\sum_{i=1}^K \left|\mathbf{b}^{(i)} \cdot (\mathbf{v}^{(t)})^{\perp V_t}\right|^2 \geq \left|\sum_{i=1}^K (\mathbf{b}^{(i)} \cdot \mathbf{w})(\mathbf{b}^{(i)} \cdot (\mathbf{v}^{(t)})^{\perp V_t})\right|^2 \,,$$

where we used that $\sum_{i=1}^K \left|\mathbf{b}^{(i)} \cdot \mathbf{w}\right|^2 = 1$ from Pythagorean theorem. Therefore, we have

$$\begin{aligned}\Phi_{t+1} &\leq \Phi_t - \left|\sum_{i=1}^K (\mathbf{b}^{(i)} \cdot \mathbf{w})(\mathbf{b}^{(i)} \cdot (\mathbf{v}^{(t)})^{\perp V_t})\right|^2 \\ &\leq \Phi_t - \beta^2 \,.\end{aligned}$$

This completes the proof of Claim 10.2.16. □

Note that by Claims 10.2.15 and 10.2.16, if $h_t$ has large misclassification error, we can decrease the potential $\Phi_t$ by a non-trivial amount. However, since the potential

function is nonnegative, this implies that at some iteration $t$, $h_t$ must have a small error. We aim to show that once this happens, the subsequent hypotheses $h_t$ also maintain small error. In the following claim, we show that by adding new vectors to the subspace $V_t$, we decrease the error of our piecewise constant approximation.

**Claim 10.2.17** (Error Decrease)**.** *For each $t \in [T]$, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_{t+1}(\mathbf{x}) \neq y] \leq \mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_t(\mathbf{x}) \neq y]$.*

*Proof of Claim 10.2.17.* Since Proposition 10.2.8 holds for any $\epsilon'$-partitions $\mathcal{S}_t, t \in [T]$, independent of the choice of basis and threshold points, we can assume that all the classifiers $h_t$ are computed with respect to extensions of a common orthonormal basis and that all threshold points are aligned. Hence, $\mathcal{S}_{t+1}$ is a subdivision of $\mathcal{S}_t$. Thus, each set $S \in \mathcal{S}_t$ is equal to a union of sets $S_1, \ldots, S_l \in \mathcal{S}_{t+1}$, i.e., $S = \cup_{i=1}^{l} S_i$. Therefore, by the definition of $h_t$, we have that

$$
\begin{aligned}
&\mathop{\mathbf{Pr}}_{(x,y)\sim D}[h_{t+1}(\mathbf{x}) \neq y, \mathbf{x} \in S] \\
&= \sum_{i=1}^{l} \mathop{\mathbf{Pr}}_{(x,y)\sim D}[h_{t+1}(\mathbf{x}) \neq y \mid \mathbf{x} \in S_i] \, \mathbf{Pr}[S_i] = \sum_{i=1}^{l} \min_{\widehat{y}\in[K]} \mathop{\mathbf{Pr}}_{(x,y)\sim D}[y \neq \widehat{y} \mid \mathbf{x} \in S_i] \, \mathbf{Pr}[S_i] \\
&\leq \min_{\widehat{y}\in[K]} \sum_{i=1}^{l} \mathop{\mathbf{Pr}}_{(x,y)\sim D}[y \neq \widehat{y} \mid \mathbf{x} \in S_i] \, \mathbf{Pr}[S_i] = \min_{\widehat{y}\in[K]} \mathop{\mathbf{Pr}}_{(x,y)\sim D}[\widehat{y} \neq y, \mathbf{x} \in \cup_{i=1}^{l} S_i] \\
&\leq \mathop{\mathbf{Pr}}_{(x,y)\sim D}[h_t(\mathbf{x}) \neq y, \mathbf{x} \in S] \; .
\end{aligned}
$$

Summing over all $S \in \mathcal{S}$ we conclude the proof. □

In the following claim, we show that if we use a sufficiently large number of samples, then with high probability we obtain an accurate estimate of the population piecewise constant approximation function.

**Claim 10.2.18** (Concentration of the Piecewise Constant Approximation)**.** *Let $\epsilon, \epsilon', \delta \in (0,1)$ and $k, K \in \mathbb{Z}_+$ with $\epsilon' \leq \epsilon/2$. Let $V \subseteq \mathbb{R}^d$ be a $k$-dimensional subspace, and consider a piecewise constant approximation $h : \mathbb{R}^d \to [K]$ of $D$, defined with respect to an $\epsilon'$-approximating partition of $V$ (see Definition 10.2.3). Let $\widehat{D}$ be the empirical distribution obtained from $N$ i.i.d. samples drawn from $D$, and let $\widehat{h}$ be a piecewise constant approximation of $\widehat{D}$ defined with respect to the same partition. Then there exists $N = (k/\epsilon')^{O(k)}(K/\epsilon)^{O(1)} \log(1/\delta)$ such that, with probability at least $1-\delta$, we have $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\widehat{h}(\mathbf{x}) \neq y] \leq \mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] + \epsilon$ .*

*Proof.* Let $\mathcal{S}$ denote the approximating partition used to define both $h$ and $\widehat{h}$. By the definition of an approximating partition (Definition 10.2.2) and using Fact 10.1.7, we have that for any $S \in \mathcal{S}$ it holds $\mathbf{Pr}_D[S] = (\epsilon'/k)^{\Omega(k)}$ which implies that $|\mathcal{S}| = (k/\epsilon')^{O(k)}$. Therefore, by the union bound and Hoeffding's inequality it holds that if $N \geq C \log(|\mathcal{S}|K/\delta)/(\epsilon/|S|)^2 = (k/\epsilon')^{O(k)}(K/\epsilon)^{O(1)} \log(1/\delta)$ for a sufficiently large universal constant $C > 0$, then with probability at least $1-\delta$, it holds that for every $S \in \mathcal{S}$ and every $i \in [K]$, we have that $|\mathbf{Pr}_{\widehat{D}}[\mathbf{x} \in S, y = i] - \mathbf{Pr}_D[\mathbf{x} \in S, y = i]| \leq \epsilon/(4|\mathcal{S}|)$. Hence, with probability $1-\delta$ we have that for all $S \in \mathcal{S}$ it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[\widehat{h}(\mathbf{x}) \neq y] \leq \mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] + \epsilon/(2|\mathcal{S}|)$. The claim follows, as the total error is simply the summation of the error over cubes and the error outside the partition which is at most $\epsilon'$ from Definition 10.2.2. □

Note that from Lines 3, 5 of Algorithm 24 we perform at most $T = \mathrm{poly}(K, 1/\epsilon)$ iterations. Moreover, in each iteration we add at most $\mathrm{poly}(K, 1/\epsilon)$ vectors to our current vector set $L_t$, hence the dimension $k_t$ of $V_t$ satisfies $k_t = \mathrm{poly}(K, 1/\epsilon)$ for all $t = 1, \ldots, T$.

Assume that for every $t = 1, \ldots, T$ the error satisfies $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_t(\mathbf{x}) \neq y] >$

$C \cdot \mathrm{OPT} + \epsilon$. Note that since $k_t = \mathrm{poly}(K, 1/\epsilon)$ for all $t \in [T]$, if we choose $\epsilon' = \mathrm{poly}(\epsilon/K)$ for a sufficiently small polynomial, then for all $t \in [T]$ we have $\epsilon' \leq c\epsilon^2/(K^2\sqrt{k_t})$ for a sufficiently small absolute constant $c > 0$. Hence, using that $N = d2^{\mathrm{poly}(K,1/\epsilon)} \log(1/\delta)$ for a sufficiently large polynomial (Line 4 of Algorithm 24), we can apply Claim 10.2.15 and conclude that with probability $1 - \delta$ there exist unit vectors $\mathbf{v}^{(t)} \in V_{t+1}$ and unit vectors $\mathbf{w}^{(t)} \in W^*$ for $t \in [T]$ such that $\mathbf{w}^{(t)} \cdot (\mathbf{v}^{(t)})^{\perp V_t} \geq \mathrm{poly}(1/K, \epsilon)$. By Claim 10.2.16, it follows that with probability at least $1 - \delta$, the potential function decreases by at least $\mathrm{poly}(\epsilon, 1/K)$ at each iteration, that is $\Phi_t \leq \Phi_{t-1} - \mathrm{poly}(\epsilon, 1/K)$. After $T$ iterations, it obtain that $\Phi_T \leq \Phi_0 - T\mathrm{poly}(\epsilon, 1/K) = 1 - T\mathrm{poly}(\epsilon, 1/K)$. However, since $T$ is chosen to be a sufficiently large polynomial of $\epsilon$ and $K$, we would arrive at a contradiction, since $\Phi_T \geq 0$. Hence, we have that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_t(\mathbf{x}) \neq y] \leq C \cdot \mathrm{OPT} + \epsilon$ for some $t \in [T]$. Since the error of $h_t$ can only decrease over the iterations (see Claim 10.2.17) and $h_t$ is close to its empirical version (Claim 10.2.18), we have that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] < C \cdot \mathrm{OPT} + \epsilon$.

**Sample and Computational Complexity:** From the analysis above, we have that the algorithm terminates in $T = \mathrm{poly}(1/\epsilon, K)$ iterations and at each iteration we draw $d2^{\mathrm{poly}(K,1/\epsilon)} \log(1/\delta)$ samples. Hence, we have that the total number of samples is of the order of $d2^{\mathrm{poly}(K,1/\epsilon)} \log(1/\delta)$. Moreover, the algorithm runs in $\mathrm{poly}(N)$ time as all the computational operations can be implemented in polynomial time. □

### 10.2.4 Proof of Lemma 10.2.7

This section proves Lemma 10.2.7, one of the supporting technical lemmas used in the proof of Proposition 10.2.4. For completeness we restate the lemma below.

**Lemma 10.2.7.** *Let* $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)} \in \mathbb{R}^d$, $t_1, \ldots, t_K \in \mathbb{R}_-$ *and* $s : \mathbb{R}^d \to \mathbb{R}$ *such that*

$s(\mathbf{x}) = \max_{i\in[K]} \max(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i, 0)$. *There exists* $\gamma > 0$ *such that*

$$\Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0]/96 \leq \Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma] \leq 1/48$$

*and*

$$\Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq n\gamma] \lesssim (1/n^2) \Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma]$$

*for all* $n \in \mathbb{Z}_+$.

*Proof.* The proof of the lemma immediately follows from Claim 10.2.19 and Claim 10.2.20. First, we establish Claim 10.2.19, which provides a bound on the tails of a piecewise affine function. Then, using Claim 10.2.20, we show that there is indeed a suitable $\gamma$ to which this tail bound can be applied.

**Claim 10.2.19** (Tail Bounds)**.** *Let* $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)} \in \mathbb{R}^d$, $t_1, \ldots, t_K \in \mathbb{R}$ *and* $s : \mathbb{R}^d \to \mathbb{R}$ *such that* $s(\mathbf{x}) = \max_{i\in[K]} \max(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i, 0)$. *For any* $\gamma > 0$ *such that* $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma] \leq 1/48$, *we have that* $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq n\gamma] \lesssim (1/n^2)\,\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma]$ *for all* $n \in \mathbb{Z}_+$.

*Proof.* By the properties of addition and scalar multiplication of independent Gaussian random variables, we have that for any $n \in \mathbb{Z}_+$ a standard Gaussian random variable $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ can be written as $\mathbf{x} = \frac{1}{n}\sum_{i=1}^{n^2} \mathbf{z}^{(i)}$ where $\mathbf{z}^{(i)}$ are drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Now if for some $i \in [n^2]$ we have that $s(\mathbf{z}^{(i)}) \geq n\gamma$, then there exists $j \in [K]$ such that $\mathbf{w}^{(j)} \cdot \mathbf{z}^{(i)} + t_j \geq n\gamma$ (by definition of $s(\mathbf{x})$) and, furthermore, we have that

$$\mathbf{w}^{(j)} \cdot \mathbf{x} + t_j = \frac{1}{n}\mathbf{w}^{(j)} \cdot \mathbf{z}^{(i)} + t_j + \frac{1}{n} \sum_{k\neq i, k\in[n^2]} \mathbf{w}^{(j)} \cdot \mathbf{z}^{(k)} \geq \gamma + \mathbf{w}^{(j)} \cdot \left(\frac{1}{n} \sum_{k\neq i, k\in[n^2]} \mathbf{z}^{(k)}\right).$$

Note that $\frac{1}{n}\sum_{k\neq i} \mathbf{z}^{(k)}$ is distributed according to $\mathcal{N}(\mathbf{0}, \frac{n^2-1}{n^2}\mathbf{I})$. Hence, by symmetry we

have that $\mathbf{Pr}[\mathbf{w}^{(j)} \cdot (\frac{1}{n}\sum_{k\neq i}\mathbf{z}^{(k)}) \geq 0] = 1/2$. As a result we have that $\mathbf{Pr}[s(\mathbf{x}) \geq \gamma \mid s(\mathbf{z}^{(i)}) \geq n\gamma] \geq 1/2$, for all $i \in [n^2]$.

Denote by $p_n = \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq n\gamma]$. To prove Claim 10.2.19, it suffices to show that $p_n \leq 4p_1/n^2$, for all $n \geq 1$. We will prove it by induction. Note that for $n = 1$ the hypothesis holds trivially. Assume as an inductive hypothesis that $p_{n-1} \leq 4p_1/(n-1)^2$ for some $n \geq 2$. We will prove that $p_n \leq 4p_1/n^2$ for $n \geq 2$.

Fix $i \in [n^2]$ and denote by $A$ the event that $s(\mathbf{x}) \geq \gamma$ and by $B_i^{(n)}$ the event that $s(\mathbf{z}^{(i)}) \geq n\gamma$. Note that

$$\begin{aligned}\mathbf{Pr}\left[A, B_i^{(n)}, \bigcap_{j\neq i} B_j^{(n)}\right] &\geq \mathbf{Pr}[A, B_i^{(n)}] - \sum_{\substack{j=1\\ j\neq i}}^{n^2} \mathbf{Pr}[B_i^{(n)}, B_j^{(n)}] \\ &\geq p_n/2 - (n^2-1)p_n^2 \,,\end{aligned}$$

where in the first inequality we used the union bound, and in the second inequality we used the independence of the $\mathbf{z}^{(i)}$'s and that $\mathbf{Pr}[s(\mathbf{x}) \geq \gamma \mid s(\mathbf{z}^{(i)}) \geq n\gamma] \geq 1/2$, for all $i \in [n^2]$.

Note that it holds $p_n \leq p_{n-1}$, since $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq n\gamma] \leq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq (n-1)\gamma]$. Consequently, by the inductive hypothesis we have that $(n^2-1)p_n^2 \leq 4p_1(n^2-1)/(n-1)^2 p_n$. Hence, since $(n^2-1)/(n-1)^2 \leq 3$ for all $n \geq 2$ and $p_1 \leq 1/48$, we have that $(n^2-1)p_n^2 \leq p_n/4$.

Summing over all $i \in [n^2]$ we have

$$p_1 = \mathbf{Pr}[A] \geq \mathbf{Pr}\left[A, \bigcup_{i\in[n^2]}\left(B_i^{(n)}, \bigcap_{j\neq i} B_j^{(n)}\right)\right] \geq n^2 p_n/4 \,.$$

Therefore, we have that $p_n \leq 4p_1/n^2$, which completes the proof of Claim 10.2.19. □

To apply Claim 10.2.19, we must choose a threshold $\gamma$ such that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(0,I)}[s(\mathbf{x}) \geq \gamma] \leq 1/48$. We now show the following result.

**Claim 10.2.20.** *For any sufficiently small $c \in (0,1)$, there exists $\gamma > 0$ such that*

$$c \geq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma] \geq \frac{c}{2} \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0] \,.$$

*Proof of Claim 10.2.20.* Notice that the quantity $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma]$ is continuous with respect to $\gamma$ for any $\gamma > 0$. Indeed, $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\gamma \leq s(\mathbf{x}) \leq \gamma + \epsilon] \leq \sum_{i=1}^{K} \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[-t_i + \gamma \leq \mathbf{w}^{(i)} \cdot \mathbf{x} \leq -t_i + \gamma + \epsilon]$ where each term such that $\mathbf{w}^{(i)} \neq \mathbf{0}$ tends to 0 as $\epsilon$ tends to 0 and each term such that $\mathbf{w}^{(i)} = \mathbf{0}$ is equal to 0 from the fact that $-t_i + \gamma > 0$. Also note that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma]$ tends to $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0]$ as $\gamma$ approaches 0; hence, there exists $\gamma'$ such that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0] - \epsilon/2 < \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma'] \leq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0]$. Therefore, if $c \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0]$ we have that $\gamma'$ satisfies the claim, because $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0] \geq \epsilon$.

If $c < \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0]$, assume that for all $\gamma$ such that $c \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma]$ we have that $c/2 \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma]$. Let $\gamma_0 = \sup\{\gamma : \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma] \geq c\}$. Notice that this set is non-empty as $\gamma'$ belongs in this set; hence $\gamma_0 \geq \gamma' > 0$. We have that for all $\delta > 0$ sufficiently small $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma_0 - \delta] \geq c$ while $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma_0 + \delta] \leq c/2$, which is a contradiction as $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma]$ is continuous for all $\gamma > 0$. Thus, there exists $\gamma$ such that $c \geq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \geq \gamma] \geq c/2 \geq (c/2)(\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[s(\mathbf{x}) \neq 0])$. □

Combining the the above two statements completes the proof of Lemma 10.2.7. □

## 10.3 Learning Multi-Index Models with Agnostic Noise

In this section, we extend the algorithm presented in Section 10.2 for a general class of MIMs that satisfy specific regularity conditions (Definition 10.1.2), establishing Theorem 10.1.3. We begin by stating the formal versions of the regularity conditions and the theorem. Following this, we provide the pseudocode for the extended algorithm along with the necessary definitions.

As we will discuss later, we need a continuity condition on the target function in order to be able to accurately approximate it over small regions. To this end, we define the Gaussian Surface Area of a multiclass classifier. A bound on this quantity will be crucial for the analysis of our algorithm.

**Definition 10.3.1** (Gaussian Surface Area)**.** *For a Borel set $A \subseteq \mathbb{R}^d$, its Gaussian surface area is defined by $\Gamma(A) := \liminf_{\delta \to 0} \frac{\mathbf{Pr}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[\mathbf{x} \in A_\delta \setminus A]}{\delta}$, where $A_\delta = \{\mathbf{x} : \mathrm{dist}(\mathbf{x}, A) \leq \delta\}$. Let $\mathcal{Y}$ be a set of finite cardinality. For a multiclass function $f : \mathbb{R}^d \mapsto \mathcal{Y}$, we define its Gaussian surface area to be the total surface area of each of the label sets $K_z = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = z\}$, i.e., $\Gamma(f) = \sum_{z \in \mathcal{Y}} \Gamma(K_z)$. For a class of multiclass concepts $\mathcal{C}$, we define $\Gamma(\mathcal{C}) := \sup_{f \in \mathcal{C}} \Gamma(f)$.*

Now we proceed to define the regularity conditions that we assume in order to obtain efficient algorithms.

**Definition 10.3.2** (Well-Behaved Multi-Index Model)**.** *Let $d, K, m \in \mathbb{Z}_+, \alpha, \zeta, \sigma, \tau \in (0, 1)$ and $\mathcal{Y} \subseteq \mathbb{Z}$ of finite cardinality. We say that a class of functions $\mathcal{F}$ from $\mathbb{R}^d \mapsto \mathcal{Y}$ is a family of concept classes of $(m, \zeta, \alpha, K, \tau, \sigma, \Gamma)$-well-behaved Multi-Index Models if, for every $f \in \mathcal{F}$, the following conditions hold:*

1. *There exists a subspace $W \subseteq \mathbb{R}^d$ of dimension at most $K$ such that $f$ depends only on the projection onto $W$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$ it holds $f(\mathbf{x}) = f(\mathbf{x}^W)$.*

2. *The Gaussian surface area of $f$ is bounded by $\Gamma$, i.e., $\Gamma(f) \leq \Gamma$.*

3. *Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For any random variable $y$ supported on $\mathcal{Y}$ with $\mathbf{Pr}[f(\mathbf{x}) \neq y] \leq \zeta$ and any subspace $V \subseteq \mathbb{R}^d$,*

   (a) *either there exists a function $g : V \to \mathcal{Y}$ such that $\mathbf{Pr}_{\mathbf{x},y}[g(\mathbf{x}^V) \neq f(\mathbf{x})] \leq \tau$,*

   (b) *or with probability at least $\alpha$ over $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{x}_0$ is independent of $\mathbf{x}$, there exists a polynomial $p : U \to \mathbb{R}$ of degree at most $m$ and a label $z \in \mathcal{Y}$ such that $\mathbf{E}[p(\mathbf{x}^U)] = 0$ and $\mathbf{E}[p(\mathbf{x}^U)\mathbb{1}(y = z) \mid \mathbf{x}^V = \mathbf{x}_0^V] > \sigma \|p(\mathbf{x}^U)\|_2$, where $U = (V + W) \cap V^\perp$.*

We can now state the formal version of Theorem 10.1.3.

**Theorem 10.3.3** (Approximately Learning Multi-Index Models with Agnostic Noise)**.** *Fix $d, K, m \in \mathbb{Z}_+$, $\zeta, \alpha, \tau, \sigma \in (0, 1)$ and $\Gamma \in \mathbb{R}_+$. Let $D$ be a distribution over $\mathbb{R}^d \times \mathcal{Y}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and let $\epsilon, \delta \in (0, 1)$. Let $\mathcal{F}$ a class of $(m, \zeta, \alpha, K, \tau, \sigma, \Gamma)$-well-behaved Multi-Index Models according to Definition 10.3.2 and let* OPT *such that* $\mathrm{OPT} = \inf_{f \in \mathcal{F}} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y]$. *If $\zeta \geq \mathrm{OPT} + \epsilon$, then, Algorithm 26 draws $N = d^{O(m)} 2^{\mathrm{poly}(mK\Gamma|\mathcal{Y}|/(\epsilon\alpha\sigma))} \log(1/\delta)$ i.i.d. samples from $D$, runs in time $\mathrm{poly}(N)$, and returns a hypothesis $h$ such that, with probability at least $1 - \delta$, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq \tau + \mathrm{OPT} + \epsilon$ .*

Definition 10.3.2 outlines the crucial properties necessary for our algorithm to produce a satisfactory solution. We first assume that our concept class consists of Multi-Index Models with finite Gaussian surface area. This condition bounds the complexity of the

decision boundary of the multi-index models. We leverage this condition in order to show that if we partition our space to sufficiently small cubes then the function should be almost constant within each cube. This is a condition that holds for many classes of interest, such as multiclass linear classifiers and intersections of halfspaces.

In addition, we impose a structural condition on the target class to guarantee progress at each step of the algorithm. Intuitively, if $V$ is the subspace we have currently identified, we require that either it already provides a good approximation (Condition 3(a)) or we can use moments to uncover further directions in the hidden subspace $W$ Condition 3(b). Specifically, we assume that for any perturbation $y$ of a function $f$ from the class that we are attempting to learn, and for any given subspace $V$, one of the following must hold: (i) there exists a good approximation to $y$ using a function defined over $V$ (Condition 3(a)), or (ii) with some probability over orthogonal cylinders to $V$, there exists a low-degree, zero-mean, variance-one polynomial along directions that correlate with the ground-truth subspace $W$, which non-trivially correlates with the level sets of $y$ over that orthogonal cylinder (Condition 3(b). In this manner, extracting these directions and adding them to $V$ enables an improved approximation of the hidden-subspace $W$. Although Condition 3(b) is defined over specific points with a measure of zero, we demonstrate in Lemma 10.3.5 that, by leveraging our assumptions, there exists a $\alpha$ fraction of finite width cylinders that exhibit this property.

Condition 3(a) of Definition 10.3.2 further provides an approximation factor that the optimal function within the subspace $V$ can achieve. This function can be approximated using a non-parametric regression algorithm. In fact, in Lemma 10.3.4, we formally show that a piece-wise constant approximation is enough to approximate the unknown function $g$.

We prove this by leveraging the Gaussian noise sensitivity of $f$ to show that a noisy version of $g$ is nearly constant on sufficiently thin cubes defined over $V$. This way using the noisy version of $g$ we can construct a good enough piecewise constant approximation. Consequently, the function that minimizes the disagreement over these cubes will also exhibit low error.

**Algorithm 26** Learning Multi-Index Models with Agnostic Noise

**Input:** Accuracy $\epsilon \in (0,1)$; failure probability $\delta \in (0,1)$; sample access to distribution $D$ over $\mathbb{R}^d \times \mathcal{Y}$ with $\mathbf{x}$–marginal $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

**Output:** Hypothesis $h$ such that with probability $\geq 1 - \delta$,

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[\, h(\mathbf{x}) \neq y \,] \;\leq\; \tau + \mathrm{OPT} + \epsilon, \quad \mathrm{OPT} \;=\; \inf_{f \in \mathcal{F}} \mathbf{Pr}[f(\mathbf{x}) \neq y].$$

Choose $T, p$ sufficiently large polynomials in $1/\alpha$, $1/\epsilon$, $1/\sigma$, $K$, $\Gamma$, $|\mathcal{Y}|$, $m$, and a universal constant $C$.

Initialize $L_1 \leftarrow \emptyset$, $t \leftarrow 1$, $\epsilon' \leftarrow 1/p$, $N \leftarrow d^{C\,m}\, 2^{T^C} \log \frac{1}{\delta}$.

**while** $t \leq T$ **do**

    Draw $S_t$: $N$ i.i.d. samples from $D$.

    $\mathcal{E}_t \leftarrow$ Alg. 27$(\epsilon, \epsilon', \delta, \mathrm{span}(L_t), S_t)$.

    $L_{t+1} \leftarrow L_t \cup \mathcal{E}_t$.

    $t \leftarrow t + 1$.

**end while**

Construct $\epsilon'$–approximating partition $\mathcal{S}$ w.r.t. $\mathrm{span}(L_{T+1})$ (Def. 10.2.2).

Draw $N$ i.i.d. samples from $D$ and build piecewise-constant classifier $h_{\mathcal{S}}$: for each $S \in \mathcal{S}$, label by majority among samples in $S$.

**Return:** $h_{\mathcal{S}}$

**Lemma 10.3.4** (Piecewise Constant Approximation Suffices)**.** *Let $d, k \in \mathbb{Z}_+$, $\epsilon, \epsilon' \in (0,1)$ and let $\mathcal{Y}$ be a set of finite cardinality. Let $D$ be a distribution supported on $\mathbb{R}^d \times \mathcal{Y}$, whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and let $V \subseteq \mathbb{R}^d$ be a $k$-dimensional subspace, with $k \geq 1$. Suppose that for a function $f \in \mathcal{F}$, it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] \leq \epsilon$. Additionally, assume that there exists a function $g : V \to \mathcal{Y}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[g(\mathbf{x}^V) \neq f(\mathbf{x})] \leq \tau$. Let $\eta = \epsilon'^2/(C\Gamma(f)|\mathcal{Y}|\sqrt{k})$, for $C > 0$ a sufficiently large universal constant, and $\mathcal{S}$ be*

**Algorithm 27** Estimating a Relevant Direction

**Input:** tolerance $\epsilon > 0$, approximation $\epsilon' > 0$, confidence $\delta > 0$, subspace $V \subseteq \mathbb{R}^d$, samples $\{(\mathbf{x}^{(i)}, y_i)\}_{i=1}^N \sim D$ with $\mathbf{x}$-marginal $\mathcal{N}(\mathbf{0}, \mathbf{I})$
**Output:** set of unit vectors $\mathcal{E}$
Choose $t, \eta$ as sufficiently small polynomials in $\epsilon, \alpha, \sigma, 1/K, 1/\Gamma, 1/m, 1/|\mathcal{Y}|$
Form empirical distribution $\widehat{D}$ from the samples
Construct $\mathcal{S}$, an $\epsilon'$-approximating partition w.r.t. $V$ (Def. 10.2.2)
**for all** $S \in \mathcal{S}$ **do**
    **for all** $i \in \mathcal{Y}$ **do**
        Compute polynomial $p_{S,i}(\cdot) \in \mathcal{P}_m$ satisfying

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\Big[(\mathbf{1}_{y=i} - p_{S,i}(\mathbf{x}^{\perp V}))^2 \mid \mathbf{x} \in S\Big] \leq \min_{p' \in \mathcal{P}_m} \mathbf{E}\Big[(\mathbf{1}_{y=i} - p'(\mathbf{x}^{\perp V}))^2 \mid \mathbf{x} \in S\Big] + \eta^2.$$

        Let $U_{S,i}$ be the unit-eigenvectors of $\mathbf{E}_{\mathbf{x}\sim D_{\mathbf{x}}}\Big[\nabla p_{S,i}(\mathbf{x}^{\perp V})\, \nabla p_{S,i}(\mathbf{x}^{\perp V})^\intercal \mid \mathbf{x} \in S\Big]$ with eigenvalues $\geq \sigma^2/(4K)$
    **end for**
    Set $U_S \leftarrow \bigcup_{i \in \mathcal{Y}} U_{S,i}$
**end for**
Compute

$$\widehat{\mathbf{U}} = \sum_{S \in \mathcal{S}} \sum_{\mathbf{u} \in U_S} \Big(\mathbf{u}^{\perp V}\Big)\Big(\mathbf{u}^{\perp V}\Big)^\intercal \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim \widehat{D}}[S].$$

**Return:** all unit-eigenvectors of $\widehat{\mathbf{U}}$ with eigenvalues $\geq t$

*an $\eta$-piecewise approximating partition with respect to $V$ (see Definition 10.2.2). Then, it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_{\mathcal{S}}(\mathbf{x}) \neq y] \leq \tau + \epsilon + \epsilon'$, where $h_{\mathcal{S}}$ is an $\eta$-piecewise constant approximation function (see Definition 10.2.3).*

*Proof.* Fix $i \in \mathcal{Y}$ and $\rho \in (0,1)$ and consider the boolean functions $f_i(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{1}(f(\mathbf{x}) = i)$, $g_i(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{1}(g(\mathbf{x}^V) = i)$. We denote the Gaussian noise operator by $T_\rho$, as defined in Definition 10.1.9. By Fact 10.1.10, we have that $\|f_i - g_i\|_{L_1} \geq \|T_\rho f_i - T_\rho g_i\|_{L_1}$. Hence, from the triangle inequality we also have that $\|f_i - g_i\|_{L_1} + \|T_\rho f_i - f_i\|_{L_1} \geq \|f_i - T_\rho g_i\|_{L_1}$. Furthermore, by the Ledoux-Pisier inequality (Fact 10.1.11), we have

that $\|T_\rho f_i - f_i\|_{L_1} \leq \sqrt{\pi}\Gamma(f_i)\rho$. Therefore, we have that

$$\|f_i - T_\rho g_i\|_{L_1} \leq \|f_i - g_i\|_{L_1} + \sqrt{\pi}\Gamma(f_i)\rho \,.$$

Moreover, note that for any continuously differentiable function $F : \mathbb{R}^d \mapsto \mathbb{R}$, it holds that $F(\mathbf{a}) - F(\mathbf{b}) = \int_0^1 \nabla F(\mathbf{b} + t(\mathbf{a} - \mathbf{b})) \cdot (\mathbf{a} - \mathbf{b}) \mathrm{d}t$. By the Cauchy-Schwarz inequality, we have that $|F(\mathbf{a}) - F(\mathbf{b})| \leq \max_{t\in(0,1)} \|\nabla F(\mathbf{b} + t(\mathbf{a} - \mathbf{b}))\| \|\mathbf{a} - \mathbf{b}\|$. By taking $F = T_\rho g_i$ and using Fact 10.1.10, we have that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it holds

$$|T_\rho g_i(\mathbf{x}) - T_\rho g_i(\mathbf{y})| \leq \frac{\|\mathbf{x} - \mathbf{y}\|}{\rho} \,.$$

For each $S \in \mathcal{S}$, pick a point $\mathbf{x}_S \in S$. Then for any $\mathbf{x} \in S$, the distance between $\mathbf{x}$ and $\mathbf{x}_S$ is at most the diameter of $S$. In our setting, since the partition is constructed along a $k$-dimensional subspace $V$ with edge length roughly $\eta$ (i.e., each coordinate is in an interval of length $\eta$), the distance between any $\mathbf{x} \in S$ and $\mathbf{x}_S$ is at most $\|\mathbf{x} - \mathbf{x}_S\| \leq \sqrt{k}\eta$. Thus, for every $\mathbf{x} \in S$

$$|T_\rho g_i(\mathbf{x}) - T_\rho g_i(\mathbf{x}_S)| \leq \frac{\|\mathbf{x} - \mathbf{x}_S\|}{\rho} \leq \frac{\sqrt{k} \cdot \eta}{\rho}.$$

By averaging over the partition $\mathcal{S}$, we have that

$$\sum_{S\in\mathcal{S}} \|(T_\rho g_i(\mathbf{x}_S) - T_\rho g_i)\mathbb{1}(\mathbf{x} \in S)\|_{L_1} \leq \sum_{S\in\mathcal{S}} \frac{\sqrt{k}\eta}{\rho} \mathbf{Pr}[S] \leq \frac{\sqrt{k}\eta}{\rho} \,.$$

This in turn gives that

$$\sum_{S\in\mathcal{S}} \|(T_\rho g_i(\mathbf{x}_S) - f_i)\mathbb{1}(\mathbf{x} \in S)\|_{L_1} \leq \|f_i - g_i\|_{L_1} + \sqrt{\pi}\Gamma(f_i)\rho + \frac{\sqrt{k}\eta}{\rho} \,.$$

Furthermore, by linearity of expectation and the fact that $\sum_{i\in\mathcal{Y}} g_i(\mathbf{x}) = 1$ we have that $\sum_{i\in\mathcal{Y}} T_\rho g_i(\mathbf{x}) = 1$ for $\mathbf{x} \in \mathbb{R}^d$. Moreover, $T_\rho g_i(\mathbf{x}) \in [0,1]$ for all $\mathbf{x} \in \mathbb{R}^d, i \in \mathcal{Y}$.

Now, we define a multiclass function $G : \mathbb{R}^d \to \mathcal{Y}$ in the following way: for each point $\mathbf{x} \in \mathbb{R}^d$, $G(\mathbf{x})$ is a random label chosen according to the probability distribution induced by $T_\rho g_i(\mathbf{x}_S)$, where $S \in \mathcal{S}$ is the region such that $\mathbf{x} \in S$. Furthermore, for every probability vector $\mathbf{p} \in \mathbb{R}^{|\mathcal{Y}|}$ (where each element of $\mathbf{p}$ represents the probability of the corresponding label) and the unit vector $\mathbf{e}_z$, which represents the one-hot encoding of the label $z \in \mathcal{Y}$. We have that

$$\|\mathbf{p} - \mathbf{e}_z\|_1 = |\mathbf{p}_z - 1| + \sum_{i\neq z, i\in\mathcal{Y}} |\mathbf{p}_i| = 2\,\mathop{\mathbf{Pr}}_{i\sim\mathbf{p}}[i \neq z]\ .$$

Denote by $A = \bigcup_{S\in\mathcal{S}} S$ the set of points that lie inside the regions of $\mathcal{S}$. Considering the disagreement between $G$ and $f$ in $A$, we obtain

$$\begin{aligned}
\mathbf{Pr}[G(\mathbf{x}) \neq f(\mathbf{x}), \mathbf{x} \in A] &\leq \sum_{S\in\mathcal{S}} \mathbf{Pr}[G(\mathbf{x}) \neq f(\mathbf{x}), \mathbf{x} \in S] \\
&= \sum_{i\in\mathcal{Y}} \sum_{S\in\mathcal{S}} \left\|(T_\rho g_i(\mathbf{x}_S) - f_i)\mathbb{1}(\mathbf{x} \in S)\right\|_{L_1} /2 \\
&\leq \sum_{i\in\mathcal{Y}} \|f_i - g_i\|_{L_1}/2 + \sqrt{\pi}\Gamma(f)\rho + |\mathcal{Y}|\sqrt{k}\eta/\rho\ .
\end{aligned}$$

Noting that $\sum_{i\in\mathcal{Y}} \|f_i - g_i\| = 2\tau$, set $\rho = \epsilon'/(C\Gamma(f))$ and $\eta = \epsilon'\rho/(C|\mathcal{Y}|\sqrt{k})$, for a sufficiently large constant $C > 0$. With these choices, we obtain $\mathbf{Pr}[G(\mathbf{x}) \neq f(\mathbf{x}), \mathbf{x} \in A] \leq \tau + \epsilon'/2$. By the triangle inequality we have that $\mathbf{Pr}[G(\mathbf{x}) \neq y, \mathbf{x} \in A]] \leq \tau + \epsilon + \epsilon'/2$. Since $h_{\mathcal{S}}$ chooses the label that minimizes the error at each $S \in \mathcal{S}$, we have that $\mathbf{Pr}[h_{\mathcal{S}}(\mathbf{x}) \neq y, \mathbf{x} \in A] \leq \mathbf{Pr}[G(\mathbf{x}) \neq y, \mathbf{x} \in A]$. Finally, considering that $\mathbf{Pr}[\mathbf{x} \notin A] \leq \eta$ (recall that $\eta \leq \epsilon'^2$) completes the proof of Lemma 10.3.4. □

In the following lemma, we show that, due to the bound on the Gaussian surface area of $f$, if there is a non-negligible probability that the conditional moments of $f$'s level sets along a subspace $V$ are large, the same property holds when $V$ is partitioned into sufficiently small cubes. That is, if we partition $V$ into thin cubes, then with non-trivial probability the conditional moments computed over each cube remain large. The intuition behind the proof is that, since the function is smooth (by the GSA bound), re-randomizing over sufficiently small cubes barely affects its output. However, by re-randomizing over cubes, the moments of the levelsets become independent of which point in the cube we condition on. As a result, if we observe distinguishing moments when conditioning on a point in the cube, then these moments persist when conditioning on the entire cube.

**Lemma 10.3.5** (Existence of Large Moments on Small Cubes)**.** *Let $D$ be a distribution over $\mathbb{R}^d \times \mathcal{Y}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and let $\mathcal{F}$ be a class of $(m, \zeta + O(\epsilon), \alpha, K, \tau, \sigma, \Gamma)$-well-behaved Multi-Index Models according to Definition 10.3.2. Let $V$ be a $k$-dimensional subspace of $\mathbb{R}^d$, with $k \geq 1$, and let $\mathcal{S}$ be an $\eta$-approximating partition of $V$ according to Definition 10.2.2, with $\eta = \epsilon^3/(C|\mathcal{Y}|\Gamma^2 k^2)$, for $C > 0$ a sufficiently large universal constant.*

*Suppose that for $f \in \mathcal{F}$ with $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] \leq \zeta$ there exists no function $g : V \to \mathcal{Y}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[g(\mathbf{x}^V) \neq f(\mathbf{x})] \leq \tau$, then there exists $\mathcal{T} \subseteq \mathcal{S}$ with $\sum_{S\in\mathcal{T}} \mathbf{Pr}[S] \geq \alpha$ such that for each $S \in \mathcal{T}$ there exists a zero-mean polynomial $p : U \to \mathbb{R}$ of degree at most $m$ and $z \in \mathcal{Y}$ such that $\mathbf{E}_{(\mathbf{x},y)\sim D}[p(\mathbf{x}^U)\mathbb{1}(y = z) \mid \mathbf{x} \in S] > \sigma\|p(\mathbf{x}^U)\|_2$, where $U = (V + W) \cap V^\perp$.*

*Proof.* Since $\mathcal{F}$ satisfies Definition 10.3.2, with probability at least $\alpha$ over $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, independent of $\mathbf{x}$, there exists a zero-mean polynomial $p : U \to \mathbb{R}$ of degree at most $m$

such that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[p(\mathbf{x}^U)\mathbb{1}(y=z) \mid \mathbf{x}^V = \mathbf{x}_0^V] > \sigma\|p(\mathbf{x}^U)\|_2 \,,$$

where $U = (V+W)\cap V^\perp$.

In order to prove the statement, we need to show that the above condition can be extended from a single point to encompass an entire cube $S$. To achieve this, we introduce a random variable $y'$, which is $(\zeta + O(\epsilon))$-close to $f$, such that, when conditioned on any $S \in \mathcal{S}$, $y'$ is distributed uniformly within $S$ along the directions defined by $V$, while the other directions remain consistent with those of $y$.

For any $\mathbf{x} \in \mathbb{R}^d$, denote by $\mathcal{S}_\mathbf{x}$ the set $S \in \mathcal{S}$ such $\mathbf{x} \in S$. We define a random vector $\mathbf{x}'$ such that $\mathbf{x}'^{\perp V} = \mathbf{x}^{\perp V}$ and $\mathbf{x}'^V$ is the sampled from the standard Gaussian over $V$ conditioned on $\mathcal{S}_\mathbf{x}$, i.e., $\mathcal{N}(\mathbf{0},\mathbf{I}) \mid \mathcal{S}_\mathbf{x}$. Note that $\mathbf{x}'$ also follows the standard normal distribution over $\mathbb{R}^d$ because its $V$-component is resampled from $\mathcal{N}(\mathbf{0},\mathbf{I})$.

In the following claim we will prove that $f(\mathbf{x})$ and $f(\mathbf{x}')$ are approximately close.

**Claim 10.3.6.** *It holds that* $\mathbf{Pr}[f(\mathbf{x}) \neq f(\mathbf{x}')] = O(\epsilon)$.

*Proof.* For each $i \in \mathcal{Y}$, define the boolean functions $f_i(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{1}(f(\mathbf{x}) = i)$ and let $g^{(i)}_{\mathbf{x}_0}(\mathbf{x}) \stackrel{\text{def}}{=} T_\rho f_i(\mathbf{x}_0^{\perp V} + \mathbf{x}^V)$, where $T_\rho$ is the Gaussian noise operator as given in Definition 10.1.9. Fix a point $\mathbf{x}_0 \in \mathbb{R}^d$ and a cube $S \in \mathcal{S}$. We first show that for any two points $\mathbf{x}, \mathbf{z} \in S$, the difference $|g^{(i)}_{\mathbf{x}_0}(\mathbf{x}) - g^{(i)}_{\mathbf{x}_0}(\mathbf{z})|$ is small.

By using that the gradient of $g^{(i)}_{\mathbf{x}_0}(\mathbf{x})$ is bounded by $1/\rho$ and that the distance between any $\mathbf{x}, \mathbf{z} \in S$ is at most $\|\mathbf{x} - \mathbf{x}_S\| \leq \sqrt{k}\eta$, we get that

$$|g^{(i)}_{\mathbf{x}_0}(\mathbf{x}) - g^{(i)}_{\mathbf{x}_0}(\mathbf{z})| \leq \frac{\eta\sqrt{k}}{\rho} \,,$$

for any $\mathbf{x}, \mathbf{z} \in S$. Consequently by the triangle inequality and Claim 10.1.13, we have

$$\sum_{S \in \mathcal{S}} \mathop{\mathbf{E}}_{\mathbf{x}_0, \mathbf{x}^V} [|(f_i(\mathbf{x}_0^{\perp V} + \mathbf{x}^V) - g^{(i)}_{\mathbf{x}_0}(\mathbf{x}))\mathbb{1}(\mathbf{x} \in S)|]$$
$$\leq \sum_{S \in \mathcal{S}} \mathop{\mathbf{E}}_{\mathbf{x}_0, \mathbf{x}^V} [|(f_i(\mathbf{x}_0^{\perp V} + \mathbf{x}^V) - g^{(i)}_{\mathbf{x}_0}(\mathbf{x}_S))\mathbb{1}(\mathbf{x} \in S)|] + \eta\sqrt{k}/\rho \lesssim \sqrt{\rho}\Gamma(f_i) + \eta\sqrt{k}/\rho \,,$$

for any collection of points $\mathbf{x}_S \in S$. Furthermore, by linearity of expectation and the fact that $\sum_{i \in \mathcal{Y}} f_i(\mathbf{x}) = 1$, we have that $\sum_{i \in \mathcal{Y}} g^{(i)}_{\mathbf{x}_0}(\mathbf{x}) = 1$ and $g^{(i)}_{\mathbf{x}_0}(\mathbf{x}) \in [0, 1]$ for all $\mathbf{x}, \mathbf{x}_0 \in \mathbb{R}^d, i \in \mathcal{Y}$. Therefore, we can consider the multiclass function $G : \mathbb{R}^d \to \mathcal{Y}$ such that $G(\mathbf{x})$ is a random label sampled from the probability distribution induced by $g^{(i)}_{\mathbf{x}}(\mathbf{x}_S)$, where $S \in \mathcal{S}$ such that $\mathbf{x} \in S$. Note that for every probability vector $\mathbf{p} \in \mathbb{R}^{|\mathcal{Y}|}$ and label $z \in \mathcal{Y}$ we have that $\|\mathbf{p} - \mathbf{e}_z\|_1 = 2\,\mathbf{Pr}_{i \sim \mathbf{p}}[i \neq z]$. Therefore, we have that the $\mathbf{Pr}[f(\mathbf{x}) \neq G(\mathbf{x})] \lesssim \sqrt{\rho}\Gamma(f) + \eta\sqrt{k}|\mathcal{Y}|/\rho \lesssim \epsilon$.

Notice that since $\mathbf{x}'$ is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I})$, we have that $\mathbf{Pr}[f(\mathbf{x}') \neq G(\mathbf{x}')] = O(\epsilon)$. However, as $G(\mathbf{x})$ is independent $\mathbf{x}^V$, we have that $\mathbf{Pr}[f(\mathbf{x}') \neq G(\mathbf{x}')] = \mathbf{Pr}[f(\mathbf{x}') \neq G(\mathbf{x})] = O(\epsilon)$. Therefore, by the triangle inequality we have that $\mathbf{Pr}[f(\mathbf{x}) \neq f(\mathbf{x}')] = O(\epsilon)$, which completes the proof of Claim 10.3.6. □

Define $y'$ such that or all $\mathbf{z} \in \mathbb{R}^d$, the distribution of $y'$ given $\mathbf{x}' = \mathbf{z}$ is the same as the distribution of $y$ given $\mathbf{x} = \mathbf{z}$. By the definition of $y'$, it follows that $\mathbf{Pr}_{\mathbf{x}', y'}[f(\mathbf{x}') \neq y'] = \mathbf{Pr}_{\mathbf{x}'}[\mathbf{E}_{y'}[f(\mathbf{x}') \neq y' \mid \mathbf{x}']] \leq \zeta$. Therefore, it holds that $\mathbf{Pr}[f(\mathbf{x}) \neq y'] \leq \zeta + O(\epsilon)$.

Therefore, with probability at least $\alpha$ over $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ independent of $\mathbf{x}$, there exists a zero-mean polynomial $p : U \to \mathbb{R}$ of degree at most $m$ and a label $y \in \mathcal{Y}$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x}, y'}[p(\mathbf{x}^U)\mathbb{1}(y' = z) \mid \mathbf{x}^V = \mathbf{x}_0^V] > \sigma\|p(\mathbf{x}^U)\|_2 \,.$$

However, since the moments of $y'$ are identical when conditioned on each $\mathbf{x}_0^V$ that belongs to the same region in $S$, we deduce that if the moment condition holds for some polynomial at a particular point $\mathbf{x}_0^V$, then it must hold for the entire cube $S$, such that $\mathbf{x}_0^V \in S$. Therefore, there exists a subset $\mathcal{T} \subseteq \mathcal{S}$ of the partition with $\sum_{S\in\mathcal{T}} \mathbf{Pr}[S] \geq \alpha$ satisfying the following: for all $S \in \mathcal{T}$ there exists a zero-mean polynomial $p : U \to \mathbb{R}$ of degree at most $m$ such that $\mathbf{E}_{\mathbf{x},y'}[p(\mathbf{x}^U)\mathbb{1}(y' = z) \mid \mathbf{x} \in S] > \sigma\|p(\mathbf{x}^U)\|_2$ .

Noticing that the moments with respect to $U$ on each $S \in \mathcal{S}$ are the same for $y$ and $y'$, completes the proof of Lemma 10.3.5. □

The remaining part of the section is structured as follows: In Section (10.3.1), we provide the analysis of Algorithm 27 with Proposition 10.3.7. In Section (10.3.2), we analyze Algorithm 26 providing the full proof of Theorem 10.3.3.

### 10.3.1 Finding a Relevant Direction

In this section, we prove Proposition 10.3.7, which states that if the target function class satisfies Definition 10.3.2 and the current approximation of the ground-truth subspace provides a suboptimal piecewise approximation, then Algorithm 27 outputs a small set of vectors, at least one of which has a non-trivial projection onto the ground-truth subspace.

**Proposition 10.3.7.** *Let $d, k \in \mathbb{Z}_+$, $\epsilon, \delta \in (0, 1)$ and $\epsilon' \leq c\epsilon^3/(|\mathcal{Y}|\Gamma^2k^2))$, for $c$ a sufficiently small universal constant. Let $D$ be a distribution supported on $\mathbb{R}^d \times \mathcal{Y}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$, let $V \subseteq \mathbb{R}^d$ be a $k$-dimensional subspace, and let $\mathcal{F}$ be a class of $(m, \zeta + O(\epsilon), \alpha, K, \tau, \sigma, \Gamma)$-well-behaved Multi-Index Models according to Definition 10.3.2. Let $f \in \mathcal{F}$ with $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] = \mathrm{OPT}$ and let $W$ be a subspace such that $f(\mathbf{x}) = f(\mathbf{x}^W)$ for all $\mathbf{x} \in \mathbb{R}^d$, with $K = \dim(W)$. Additionally, let $h_{\mathcal{S}}$ be a*

*piecewise constant approximation of $D$, with respect to an $\epsilon'$-approximating partition $\mathcal{S}$, according to $V$, as defined in Definitions 10.2.2 and 10.2.3.*

*There exists $N = (dm)^{O(m)}(k/\epsilon')^{O(k)} \log(1/\delta)(K\Gamma|\mathcal{Y}|/(\sigma\epsilon\alpha))^{O(1)}$ such that: If $\zeta \geq \mathrm{OPT}$ and $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_{\mathcal{S}}(\mathbf{x}) \neq y] > \tau + \mathrm{OPT} + 3\epsilon$, then Algorithm 27 given $N$ i.i.d. samples from $D$, runs in $\mathrm{poly}(N)$ time, and with probability at least $1-\delta$, returns a list of unit vectors $\mathcal{E}$ of size $|\mathcal{E}| = \mathrm{poly}(mK\Gamma|\mathcal{Y}|/(\epsilon\alpha\sigma))$ such that: For some $\mathbf{v} \in \mathcal{E}$ and unit vector $\mathbf{w} \in W$ it holds that*

$$\mathbf{w}^{\perp V} \cdot \mathbf{v} \geq \mathrm{poly}(\epsilon\alpha\sigma/(mK\Gamma|\mathcal{Y}|)) \; .$$

Before proving Proposition 10.3.7, we show that once a direction of our target model is learned to sufficiently small accuracy, the observed moments depend only on the remaining directions.

The intuition is as follows. Re-randomizing over directions that have negligible projection onto $W$ barely affects the output of the function. Thus, the function is close to a version $y'$ of the noisy label $y$ in which these components have been averaged. Consequently, if the function is far from being a function of the current subspace $V$ (i.e., it still depends significantly on directions outside $V$), then the level sets of $y'$ will exhibit large distinguishing moments. In this way, the influence of the re-randomized (negligible) directions is eliminated, leaving only the significant, distinguishing moments on the remaining directions.

The proof follows a structure similar to that of Lemma 10.3.5, with the main difference being the set of directions over which re-randomization occurs.

**Lemma 10.3.8.** *Let $D$ be a distribution over $\mathbb{R}^d \times \mathcal{Y}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Let*

*$\mathcal{F}$ be a class of $(m, \zeta + O(\epsilon), \alpha, K, \tau, \sigma, \Gamma)$-well-behaved Multi-Index Models as defined in Definition 10.3.2. Let $V$ be a $k$-dimensional subspace of $\mathbb{R}^d$, with $k \geq 1$, and let $\mathcal{S}$ be an $\eta$-approximating partition of $V$ (see Definition 10.2.2), where $\eta = \epsilon^3/(C|\mathcal{Y}|\Gamma^2 k^2)$ and $C > 0$ is a sufficiently large universal constant. Let $f \in \mathcal{F}$ be a function that depends only on a $K$-dimensional subspace $W$, and let $y$ be random variable over $\mathcal{Y}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] \leq \zeta$.*

*Suppose that no function $g : V \to \mathcal{Y}$ exists with $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[g(\mathbf{x}^V) \neq f(\mathbf{x})] \leq \tau$. Then, there exists $\mathcal{T} \subseteq \mathcal{S}$ with $\sum_{S\in\mathcal{T}} \mathbf{Pr}[S] \geq \alpha$ such that for each $S \in \mathcal{T}$ the following hold:*

1. *There exists a label $z \in \mathcal{Y}$ and a zero-mean, variance-one polynomial $p : U \to \mathbb{R}$ of degree at most $m$ such that $\mathbf{E}_{(\mathbf{x},y)\sim D}[p(\mathbf{x}^U)\mathbb{1}(y = z) \mid \mathbf{x} \in S] > \sigma$, where $U{=}(V{+}W) \cap V^\perp$.*

2. *For each such polynomial $p$ satisfying (1), it holds that $\nabla p(\mathbf{x}^U) \cdot \mathbf{v} = 0$ for every $\mathbf{v} \in E$ and all $\mathbf{x} \in \mathbb{R}^d$, where $E \subseteq V^\perp$ is a subspace satisfying $\|\mathbf{v}^W\| \leq \epsilon/\Gamma(f)$ for every unit vector $\mathbf{v} \in E$.*

*Proof.* Let $\mathbf{x}, \mathbf{x}'$ be random variables defined as $\mathbf{x} = \mathbf{x}^{\perp E} + \mathbf{x}^E$ and $\mathbf{x}' = \mathbf{x}^{\perp E} + \mathbf{z}^E$, where $\mathbf{z}$ is an independent standard Gaussian vector. Denote by $\mathcal{S}$ an $\eta$-approximating partition with respect to $V$. Notice that for any unit direction $\mathbf{w} \in W$, both $\mathbf{w} \cdot \mathbf{x}$ and $\mathbf{w} \cdot \mathbf{x}'$ are distributed according to a standard Gaussian. Furthermore, by independence we have that

$$\mathbf{E}[(\mathbf{w} \cdot \mathbf{x})(\mathbf{w} \cdot \mathbf{x}')] = \mathbf{E}[(\mathbf{w} \cdot \mathbf{x}^{\perp E})^2] = \|\mathbf{w}^{\perp E}\|^2 .$$

Given that $E$ is the subspace of small projections, (i.e., for any unit vector $\mathbf{w} \in W$, we have $\|\mathbf{w}^E\| \leq \epsilon/\Gamma(f)$), it follows that $\mathbf{E}[(\mathbf{w} \cdot \mathbf{x}^{\perp E})^2] \geq 1 - (\epsilon/\Gamma(f))^2$ for any unit vector

$\mathbf{w} \in W$. This means that $\mathbf{x}$ and $\mathbf{x}'$ are at least $1 - (\epsilon/\Gamma(f))^2$ correlated in $W$. By the Gaussian noise sensitivity bound (see Fact 10.1.11), we have that

$$\Pr_{\mathbf{x},\mathbf{x}'}[f(\mathbf{x}) \neq f(\mathbf{x}')] \leq O(\epsilon) \ .$$

Next, define a new random variable $y'$ over $\mathcal{Y}$ which is distributed like $y$ but with respect to $\mathbf{x}'$; that is, for every $\mathbf{z} \in \mathbb{R}^d$ and $i \in \mathcal{Y}$, $D(y' = i \mid \mathbf{x}' = \mathbf{z}) = D(y = i \mid \mathbf{x} = \mathbf{z})$. We have that

$$\mathbf{Pr}[f(\mathbf{x}) \neq y'] \leq \mathbf{Pr}[f(\mathbf{x}') \neq y'] + O(\epsilon) \leq \mathbf{Pr}[f(\mathbf{x}) \neq y] + O(\epsilon) \leq \zeta + O(\epsilon) \ .$$

As a result, by Lemma 10.3.5, we have that there exists $\mathcal{T} \subseteq \mathcal{S}$ with $\sum_{S \in \mathcal{T}} \mathbf{Pr}[S] \geq \alpha$ such that for each $S \in \mathcal{T}$ there exists a zero-mean, variance-one polynomial $p : U \to \mathbb{R}$ of degree at most $m$ and a label $z \in \mathcal{Y}$ such that $\mathbf{E}_{\mathbf{x},y'}[p(\mathbf{x}^U)\mathbb{1}(y' = z) \mid \mathbf{x} \in S] > \sigma$, where $U = (V + W) \cap V^\perp$. However we have that for all $S \in \mathcal{T}$

$$\mathop{\mathbf{E}}_{\mathbf{x},y'}[p(\mathbf{x}^U)\mathbb{1}(y' = z) \mid \mathbf{x} \in S] = \mathbf{E}\left[\mathop{\mathbf{E}}_{\mathbf{x}^E}[p(\mathbf{x}^U) \mid \mathbf{x}^{\perp E}] \mathop{\mathbf{E}}_{\mathbf{x}^E}[\mathbb{1}(y = i) \mid \mathbf{x}^{\perp E}] \mid \mathbf{x} \in S\right] \ ,$$

where we used that $y'$ is independent of $\mathbf{x}^E$. Notice that $p'(\mathbf{x}^U) \stackrel{\text{def}}{=} \mathbf{E}_{\mathbf{x}^E}[p(\mathbf{x}^U) \mid \mathbf{x}^{\perp E}]$ is a zero-mean polynomial of degree at most $m$ with variance at most one by Jensen's inequality. Furthermore, as $p'$ is independent of $\mathbf{x}^E$ we have that $\nabla p'(\mathbf{x}^U) \cdot \mathbf{v} = 0$ for any $\mathbf{v} \in E$ and $\mathbf{x} \in \mathbb{R}^d$. □

We now proceed with the proof of Proposition 10.3.7.

*Proof of Proposition 10.3.7.* Denote by $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)}$ an orthonormal basis for the subspace $W$. Let $\widehat{\mathbf{U}}$ be the matrix computed in Line 13 of Algorithm 27 and let $\eta^2$ be

the $L_2^2$ error chosen in the polynomial-regression step (see Line 8 of Algorithm 27).

First, observe that since $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_{\mathcal{S}}(\mathbf{x}) \neq y] > \tau + \mathrm{OPT} + 3\epsilon$, Lemma 10.3.4 for every implies that for every function $g : V \to \mathcal{Y}$ it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[g(\mathbf{x}^V) \neq f(\mathbf{x})] \geq \tau$. Then, by Lemma 10.3.8 there exists a $\mathcal{T} \subseteq \mathcal{S}$ with $\sum_{S\in\mathcal{T}} \mathbf{Pr}[S] \geq \alpha$, such that for each $S \in \mathcal{T}$ there exists a label $z \in \mathcal{Y}$ and a zero-mean variance-one polynomial $q_S : U \to \mathbb{R}$ of degree at most $m$ satisfying $\mathbf{E}_{(\mathbf{x},y)\sim D}[q_S(\mathbf{x}^U)\mathbb{1}(y = z) \mid \mathbf{x} \in S] > \sigma$, where $U = (V + W) \cap (V^\perp)$. Furthermore, for each $S \in \mathcal{S}$, define $E_S \stackrel{\text{def}}{=} \mathrm{span}(\{\mathbf{v} \in U_S : \left\|\mathbf{v}^W\right\| \leq \epsilon/(\sqrt{|U_S|}\Gamma)\})$. Then, for every unit vector $\mathbf{v} \in E_S$, we have $\nabla q_S(\mathbf{x}^U) \cdot \mathbf{v} = 0$ since $\left\|\mathbf{v}^W\right\| \leq \epsilon/\Gamma$ holds for every unit vector $\mathbf{v} \in E_S$.

In the following claim, we leverage the existence of non-trivial moments over $U$ for regions $S \in \mathcal{T}$. We show that if the sample size is sufficiently large then for each region $S \in \mathcal{T}$ the associated set $U_S$ (see Line 9 of Algorithm 27) contains a vector that correlates with $W$.

**Claim 10.3.9** (Existence of Correlating Vectors)**.** *Let $C > 0$ be a sufficiently large universal constant. Fix $S \in \mathcal{T}$. If the number of samples that fall in $S$ is $N_S \geq (dm)^{Cm} \log(|\mathcal{Y}|/\delta)/\eta^C$, then with probability at least $1 - \delta$ there exists a unit vector $\mathbf{v} \in U_S$ and $i \in [K]$ such that*

$$(\mathbf{w}^{(i)})^{\perp V} \cdot \mathbf{v} \geq \mathrm{poly}(\epsilon\sigma/(mK|\mathcal{Y}|\Gamma) \, .$$

*Moreover, it holds that $|U_S| = \mathrm{poly}(mK|\mathcal{Y}|/\sigma)$.*

*Proof of Claim 10.3.9.* Denote by $p_{S,i}$ (for $S \in \mathcal{S}$ and $i \in \mathcal{Y}$) the degree at most $m$ polynomial computed at Line 8 of Algorithm 27. Notice that membership of a point $\mathbf{x}$ in a cube $S$ depends solely on its projection onto $V$, i.e., on $\mathbf{x}^V$. Therefore, the

error guarantee obtained in the regression step $\mathbf{E}_{(\mathbf{x},y)\sim D}[(\mathbb{1}(y = i) - p_{S,i}(\mathbf{x}^{\perp V}))^2 \mid \mathbf{x} \in S] \leq \min_{p'\in\mathcal{P}_m} \mathbf{E}_{(\mathbf{x},y)\sim D}[(\mathbb{1}(y = i) - p'(\mathbf{x}^{\perp V}))^2 \mid \mathbf{x} \in S] + \eta^2$ is equivalent to $\mathbf{E}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[(\mathbb{1}(y = i) - p_{S,i}(\mathbf{x}))^2] \leq \min_{p'\in\mathcal{P}_m} \mathbf{E}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[(\mathbb{1}(y = i) - p'(\mathbf{x}))^2] + \eta^2$, where $D^S_{V^\perp}$ is the marginal obtained by averaging $D$ over $V$ conditioned on $S$, i.e., $D^S_{V^\perp}(\mathbf{x}^{\perp V}, y) = \mathbf{E}_{\mathbf{x}^V}[D(\mathbf{x}, y) \mid \mathbf{x} \in S]$.

Moreover, by the properties of the Gaussian distribution, we have that the $\mathbf{x}$-marginal of $D_{V^\perp}$ is a standard Gaussian. Therefore, if $N_S \geq (dm)^{Cm}\log(|\mathcal{Y}|/\delta)/\eta^C$ for a sufficiently large universal constant $C > 0$, then by applying the union bound and Fact 10.5.2, we have that with probability at least $1 - \delta$:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[(\mathbb{1}(y = i) - p_{S,i}(\mathbf{x}))^2] \leq \min_{p'\in\mathcal{P}_m} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[(\mathbb{1}(y = i) - p'(\mathbf{x}))^2] + \eta^2 \ .$$

Furthermore, by the orthogonality of Hermite polynomials, we have that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[(\mathbb{1}(y = i) - p_{S,i}(\mathbf{x}))^2] = \sum_{\beta\subseteq\mathbb{N}^d} \Big(\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[\mathrm{He}_\beta(\mathbf{x})\mathbb{1}(y = i)] - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[\mathrm{He}_\beta(\mathbf{x})p_{S,i}(\mathbf{x})]\Big)^2$$

In particular, if we decompose the error into its Hermite polynomial components of degree $t$, we have that $\sum_{t=1}^m \mathbf{E}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[(\mathbb{1}(y = i)^{[t]} - p^{[t]}_{S,i}(\mathbf{x}))^2] \leq \eta^2$.

Next, by the rotational invariance of the Gaussian distribution, we may, without loss of generality, choose an orthonormal basis $\mathbf{e}_1, \ldots, \mathbf{e}_{k'}$ for the subspace $U$ projected onto $E_S^\perp$. Note that $k' \geq 1$, otherwise, the polynomial $q_S$ (which is defined over this space) could not exist. Furthermore, since $U$ is a subset of $V^\perp$, we similarly have that $\mathbf{E}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[q_S(\mathbf{x}^U)\mathbb{1}(y = z)] > \sigma$ for some $z \in \mathcal{Y}$. Decomposing $q_S(\mathbf{x}^U)$ in the basis of Hermite polynomials we have that $q_S(\mathbf{x}^U) = \sum_{\beta\in\mathbb{N}^{d'}, 1\leq\|\beta\|_1\leq m} \widehat{q}_S(\beta)\mathrm{He}_\beta(\mathbf{x}^U)$, where $d' = \dim(U)$. Moreover, note that since $q_S$ has no component outside $\mathbf{e}_1, \ldots, \mathbf{e}_{k'}$, we

have that $q_S(\mathbf{x}^U) = \sum_{\beta \in J} \widehat{q}_S(\beta) \mathrm{He}_\beta(\mathbf{x}^U)$, where $J$ denotes the set of $\beta \in \mathbb{N}^{d'}$ such that $\beta_i \geq 1$ for some $i \in [k']$. Considering the correlation of $q_S$ with the region of the label $z$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[q_S(\mathbf{x}^U)\mathbb{1}(y=z)] &= \sum_{\beta\in J}\widehat{q}_S(\beta)\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[\mathrm{He}_\beta(\mathbf{x}^U)\mathbb{1}(y=z)] \\ &\leq \left(\sum_{\beta\in J}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[\mathrm{He}_\beta(\mathbf{x}^U)\mathbb{1}(y=z)]^2\right)^{1/2},\end{aligned}$$

where we used the fact that $\|q_S\|_2 = 1$ and the Cauchy-Schwarz inequality. Hence, we deduce that the sum of the squares of the Hermite coefficients corresponding to degrees up to $m$ in the expansion of the random variable $\mathbb{1}(y=z)$ is at least $\sigma^2$.

Now evaluating the quadratic form of $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim D_\mathbf{x}}[\nabla p_{S,z}(\mathbf{x}^{\perp V})\nabla p_{S,z}(\mathbf{x}^{\perp V})^\intercal \mid \mathbf{x} \in S]$ (Line 9 of Algorithm 27) using Fact 10.1.8 gives us

$$\begin{aligned}\sum_{i=1}^{k'}\mathbf{e}_i^\intercal\mathbf{M}\mathbf{e}_i = \sum_{i=1}^{k'}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[(\nabla p_{S,z}(\mathbf{x})\cdot\mathbf{e}_i)^2] &\geq \sum_{\beta\in\mathbb{N}^d}\left(\sum_{i=1}^{k'}\beta_i\right)(\widehat{p}_{S,z}(\beta))^2 \\ &\geq \sum_{\beta\in J}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D^S_{V^\perp}}[\mathrm{He}_\beta(\mathbf{x}^U)\mathbb{1}(y=z)]^2 - 2\eta \\ &\geq \sigma^2 - 2\eta \geq \sigma^2/2\end{aligned}$$

for $\eta \leq \sigma/2$. Thus, since $\dim(U) \leq K$ we have that $\mathbf{e}_j^\intercal\mathbf{M}\mathbf{e}_j \geq \sigma^2/(2K)$ for some $j \in [k']$.

Since $\mathbf{M}$ is symmetric and therefore diagonalizable, we can denote by $\mathbf{u}^{(i)}$ the unit eigenvectors of $\mathbf{M}$ and by $\lambda_i$ the corresponding eigenvalues. Additionally, let $U_{S,z}$ denote the set of eigenvectors of $\mathbf{M}$ with eigenvalues greater than $\sigma^2/4K$ (see Line 9 of

Algorithm 27). Decomposing $\mathbf{e}_j$ in the basis of eigenvectors gives us

$$\mathbf{e}_j^\intercal \mathbf{M}\, \mathbf{e}_j = \sum_{i=1}^{d-k} \lambda_i \, (\mathbf{e}_j \cdot \mathbf{u}^{(i)})^2.$$

Using that $U_{S,z}$ contains eigenvectors of $\mathbf{M}$ with eigenvalues greater than $\sigma^2/4K$, we get that

$$\frac{\sigma^2}{4K} + \sum_{\mathbf{u}^{(i)} \in U_{S,z}} \lambda_i (\mathbf{e}_j \cdot \mathbf{u}^{(i)})^2 \geq \mathbf{e}_j^\intercal \mathbf{M}\, \mathbf{e}_j \geq \frac{\sigma^2}{2K} \, .$$

Hence, there exists an eigenvector $\mathbf{u}^{(i)} \in U_{S,z}$ such that $\lambda_i (\mathbf{e}_j \cdot \mathbf{u}^{(i)})^2 \geq \sigma^2/(4K|U_{S,z}|)$. This implies that $(\mathbf{e}_j \cdot \mathbf{u}^{(i)})^2 \geq \sigma^2/(4|U_{S,z}|K\mathrm{tr}(\mathbf{M}))$ as $\mathbf{M}$ is positive definite. Therefore, by Fact 10.5.3, we have that $|U_{S,z}| = \mathrm{poly}(mK/\sigma), \mathrm{tr}(\mathbf{M}) = O(m)$ which implies that $|\mathbf{e}_j \cdot \mathbf{u}^{(i)}| \geq \mathrm{poly}(\sigma/(mK))$. Finally, note that since $\mathbf{e}_j \cdot \mathbf{u}^{(i)} \neq 0$, we have that $\mathbf{u}^{(i)}$ cannot belong in the space $E_S$. Thus, there exists a unit vector $\mathbf{w} \in W$ such that $|\mathbf{u}^{(i)} \cdot \mathbf{w}| \geq \epsilon/(\sqrt{|U_S|}\Gamma) \geq \mathrm{poly}(\epsilon\sigma/(mK|\mathcal{Y}|\Gamma)$, which concludes the proof of Claim 10.3.9.

□

By the definition of an approximating partition (see Definition 10.2.2) and Fact 10.1.7, for all $S \in \mathcal{S}$ we have that $\mathbf{Pr}_D[S] = (\epsilon'/k)^{\Omega(k)}$ and $|\mathcal{S}| = (k/\epsilon')^{O(k)}$. Therefore, by the union bound and Hoeffding's inequality, it holds that if the sample size $N \geq (k/\epsilon')^{Ck} \log(|\mathcal{S}|/\delta) = (k/\epsilon')^{O(k)} \log(1/\delta)$ for a sufficiently large constant $C > 0$, then with probability at least $1 - \delta$ we have that $|\mathbf{Pr}_{\widehat{D}}[S] - \mathbf{Pr}_D[S]| \leq \mathbf{Pr}_D[S]/2$ for all $S \in \mathcal{S}$. Hence, $\mathbf{Pr}_{\widehat{D}}[S] \geq \mathbf{Pr}_D[S]/2$, i.e., the number of samples that fall in each set $S \in \mathcal{S}$ is at least $N\, \mathbf{Pr}_D[S]/2 = N(\epsilon'/k)^{\Omega(k)}$.

If we choose $N$ so that $N \geq (dm)^{Cm}(k/\epsilon')^{Ck} \log(1/\delta)(K\Gamma|\mathcal{Y}|/(\sigma\epsilon\alpha))^C$ for a sufficiently large universal constant $C > 0$, by Claim 10.3.9, we have that for all $S \in \mathcal{T}$ it holds that $\mathbf{v} \cdot (\mathbf{w}^{(i)})^{\perp V} \geq \mathrm{poly}(\epsilon\sigma/(mK|\mathcal{Y}|\Gamma)$ for some $\mathbf{v} \in U_S, i \in [K]$. Hence, we have

that there exists a subset $\mathcal{T}' \subseteq \mathcal{T}$ and $i \in [K]$ with $\sum_{S \in \mathcal{T}'} \mathbf{Pr}_{\widehat{D}}[S] \geq \sum_{S \in \mathcal{T}'} \mathbf{Pr}_{D}[S]/2 = \Omega(\alpha/K)$ such that for all $S \in \mathcal{T}'$ we have that $\mathbf{v} \cdot (\mathbf{w}^{(i)})^{\perp V} \geq \mathrm{poly}(\epsilon\sigma/(mK|\mathcal{Y}|\Gamma)$ for some $\mathbf{v} \in U_S$.

Denote by $\widehat{\mathbf{U}}$ the matrix computed in Line 13 of Algorithm 27. For some $i \in [K]$, we have that

$$\left((\mathbf{w}^{(i)})^{\perp V}\right)^{\mathsf{T}} \widehat{\mathbf{U}} (\mathbf{w}^{(i)})^{\perp V} \gtrsim \frac{\alpha}{K} \mathrm{poly}(\epsilon\sigma/(mK|\mathcal{Y}|\Gamma) \geq \mathrm{poly}(\epsilon\alpha\sigma/(mK|\mathcal{Y}|\Gamma) \ ,$$

where we used the fact that $\widehat{\mathbf{U}}$ is a PSD matrix.

Moreover, as $\|\mathbf{v}\| = 1$ for all $\mathbf{v} \in U_S, S \in \mathcal{S}$, by the triangle inequality we can see that

$$\left\|\widehat{\mathbf{U}}\right\|_F \leq \sum_{S \in \mathcal{S}, \mathbf{v} \in U_S} \|\mathbf{v}\| \mathop{\mathbf{Pr}}_{\widehat{D}}[S] \leq \sum_{S \in \mathcal{S}, \mathbf{v} \in U_S} \mathop{\mathbf{Pr}}_{\widehat{D}}[S] \leq \mathrm{poly}(mK|\mathcal{Y}|/\sigma) \ ,$$

where we used the fact that $|U_S| = \mathrm{poly}(mK|\mathcal{Y}|/\sigma)$ by Claim 10.3.9.

Hence, by Claim 10.5.1, we have that there exists a unit eigenvector $\mathbf{v}$ of $\widehat{\mathbf{U}}$ with eigenvalue at least $\mathrm{poly}(\epsilon\alpha\sigma/(mK|\mathcal{Y}|\Gamma)$ for a sufficiently small polynomial, such that $(\mathbf{w}^{(i)})^{\perp V} \cdot \mathbf{v} \geq \mathrm{poly}(\epsilon\alpha\sigma/(mK\Gamma|\mathcal{Y}|))$ for some $i \in [K]$. Moreover, the number of such eigenvectors is at most $\mathrm{poly}(mK|\mathcal{Y}|\Gamma/(\epsilon\alpha\sigma))$, which completes the proof of Proposition 10.3.7. □

### 10.3.2 Proof of Theorem 10.3.3

In this section, given Proposition 10.3.7, we prove Theorem 10.3.3. Recall that, in Proposition 10.3.7, we have shown that if our current approximation to the hidden subspace is not accurate enough to produce a classifier that has sufficiently small error, then Algorithm 27 efficiently finds a direction that has non-trivial correlation to the

hidden subspace. In the proof that follows, we iteratively apply this argument to show that, after a moderate number of iterations, Algorithm 26 outputs a classifier with sufficiently small error.

*Proof of Theorem 10.3.3.* Denote by $f^* \in \mathcal{F}$ a classifier such that $\mathbf{Pr}[f^*(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon/4$. We show that Algorithm 26, with high probability, returns a hypothesis $h$ with 0-1 error at most $\tau + \mathrm{OPT} + \epsilon$. Let $W^*$ be a subspace of dimension $K$ such that $f^*(\mathbf{x}) = f^*(\mathbf{x}^{W^*})$ and denote by $\mathbf{w}^{*(1)}, \ldots, \mathbf{w}^{*(K)} \in \mathbb{R}^d$ an orthonormal basis of $W^*$. Let $L_t$ be the list of vectors maintained by the algorithm (Line 8 of Algorithm 26) and $V_t = \mathrm{span}(L_t)$, $\dim(V_t) = k_t$. Let $\epsilon'$ be the partition width parameter set at Line 4 of Algorithm 26. For $t \in [T]$, let $\mathcal{S}_t$ be arbitrary $\epsilon'$-approximating partitions with respect to $V_t$ (see Definition 10.2.2). Let $h_t : \mathbb{R}^d \to [K]$ be piecewise constant classifiers, defined as $h_t = h_{\mathcal{S}_t}$ according to Definition 10.2.3 for the distribution $D$.

To prove the correctness of Algorithm 26, we need to show that if $h_t$ has significant error, then the algorithm improves its approximation $V_t$ to $W^*$. For quantifying the improvement at every step, we consider the following potential function

$$\Phi_t = \sum_{i=1}^{K} \left\| (\mathbf{w}^{*(i)})^{\perp V_t} \right\|^2 .$$

Note that from Lines 3 and 5 of Algorithm 26, we perform at most $\mathrm{poly}(mK\Gamma|\mathcal{Y}|/(\epsilon\alpha\sigma))$ iterations. Furthermore, in each iteration, we update the vector set with at most $\mathrm{poly}(mK\Gamma|\mathcal{Y}|/(\epsilon\alpha\sigma))$ vectors (see Proposition 10.2.8). Hence, it follows that $k_t \leq \mathrm{poly}(mK\Gamma|\mathcal{Y}|/(\epsilon\alpha\sigma))$, for all $t = 1, \ldots, T$.

Assume that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_t(\mathbf{x}) \neq y] > \tau + \mathrm{OPT} + \epsilon/2$ for all $t = 1, \ldots, T$. Using the fact that $N = d^{Cm} 2^{(mK\Gamma|\mathcal{Y}|/(\epsilon\alpha\sigma))^C} \log(1/\delta)$ for a sufficiently large universal constant

$C > 0$ (Line 4 of Algorithm 26), we can apply Proposition 10.3.7 and conclude that, with probability $1 - \delta$, there exist unit vectors $\mathbf{v}^{(t)} \in V_{t+1}$ for $t = [T]$ such that $\mathbf{w} \cdot (\mathbf{v}^{(t)})^{\perp V_t} \geq \mathrm{poly}(\epsilon\alpha\sigma/(mK\Gamma|\mathcal{Y}|))$. Thus, by Claim 10.2.16, we have that with probability $1 - \delta$, for all $t \in [T]$,

$$\Phi_t \leq \Phi_{t-1} - \mathrm{poly}(\epsilon\alpha\sigma/(mK\Gamma|\mathcal{Y}|)) \; .$$

After $T$ iterations, $\Phi_T \leq \Phi_0 - T\mathrm{poly}(\epsilon\alpha\sigma/(mK\Gamma|\mathcal{Y}|)) = K - T\mathrm{poly}(\epsilon\alpha\sigma/(mK\Gamma|\mathcal{Y}|))$. However, since $T$ is set to be a sufficiently large polynomial of $m, \Gamma, K, |\mathcal{Y}|, 1/\epsilon, 1/\alpha$ and $1/\sigma$ we would arrive at a contradiction, since $\Phi_T \geq 0$. Hence, we have that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h_t(\mathbf{x}) \neq y] \leq \tau + \mathrm{OPT} + \epsilon/2$, for some $t \in \{1, \ldots, T\}$. Since the error of $h_t$ can only be decreasing by Claim 10.2.17 and $h_t$ is close to its sample variant by Claim 10.2.18, we have that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq \tau + \mathrm{OPT} + \epsilon$.

**Sample and Computational Complexity:** From the analysis above we have that the algorithm terminates in $\mathrm{poly}(mK\Gamma|\mathcal{Y}|/(\epsilon\alpha\sigma))$ iterations and at each iteration we draw of order $d^{O(m)}2^{\mathrm{poly}(mK\Gamma|\mathcal{Y}|/(\epsilon\alpha\sigma))}\log(1/\delta)$ samples. Hence, we have that the total number of samples is $d^{O(m)}2^{\mathrm{poly}(mK\Gamma|\mathcal{Y}|/(\epsilon\alpha\sigma))}\log(1/\delta)$. Moreover, we use at most $\mathrm{poly}(N)$ time as all operations can be implemented in sample polynomial time. □

## 10.4 Algorithmic Applications of General MIM Learner

In this section, we highlight selected applications of our main result. First, in Section (10.4.1) we present an application of our algorithmic approach to the class of intersections of halfspaces, establishing Theorem 10.1.6.

Additionally, we note that our result on learning Multiclass Linear Classifiers, as

presented in Theorem 10.2.1, can also be derived as a corollary of Theorem 10.3.3, using similar arguments.

### 10.4.1 Learning Intersections of Halfspaces with Adversarial Label Noise

In this section, we present Theorem 10.4.1, on learning intersections of halfspaces under agnostic noise, which is derived as a corollary of Theorem 10.3.3. The proof of this result follows from Proposition 10.4.2, which relies on a new structural result about the class of intersections of $K$-halfspaces, Proposition 10.4.4. First, we state the formal version of the theorem.

**Theorem 10.4.1** (Learning Intersections of Halfspaces with Adversarial Label Noise)**.** *Fix the class $\mathcal{F}_{K,d}$ of intersections of $K$ halfspaces in $\mathbb{R}^d$. Let $D$ be a distribution over $\mathbb{R}^d \times \{0,1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and let $\epsilon, \delta \in (0,1), K, d \in \mathbb{Z}_+$. Then, Algorithm 26 draws $N = d^2 2^{\mathrm{poly}(K/\epsilon)} \log(1/\delta)$ i.i.d. samples from $D$, runs in time $\mathrm{poly}(N)$, and returns a hypothesis $h$ such that, with probability at least $1-\delta$, it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq K\, O(\mathrm{OPT} \log(1/\mathrm{OPT})) + \epsilon$, where $\mathrm{OPT} = \inf_{f \in \mathcal{F}_{K,d}} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y]$.*

*Proof.* First, note that without loss of generality we can assume that $\mathrm{OPT} \geq \epsilon^2$ by randomly flipping an $\epsilon^2$ fraction of the labels with probability $1/2$. Therefore, by the mean value theorem we have that

$$\begin{aligned}(\mathrm{OPT} + \epsilon^2) \log(1/(\mathrm{OPT} + \epsilon^2) + 1) &\leq (\mathrm{OPT}) \log(1/(\mathrm{OPT}) + 1) + \epsilon^2 \log(1/\mathrm{OPT} + 1) \\ &\leq (\mathrm{OPT}) \log(1/(\mathrm{OPT}) + 1) + \epsilon\,.\end{aligned}$$

Hence, by applying Proposition 10.4.2 we have that the class of intersections of $K$-

halfspaces is a class of $(2, \mathrm{OPT} + \epsilon, \epsilon, K, KO(\mathrm{OPT}\log(1/\mathrm{OPT}+1)+\epsilon), \mathrm{poly}(\epsilon/K), O(\sqrt{\log K}))$-well-behaved Multi-Index Models.

Thus, by Theorem 10.3.3 we have that Algorithm 26 draws $N = d^2 2^{\mathrm{poly}(K/\epsilon)}\log(1/\delta)$ and in $\mathrm{poly}(N)$ time returns a hypothesis $h$ with error $KO(\mathrm{OPT}\log(1/\mathrm{OPT}+1))+O(\epsilon)$. Finally, note that without loss of generality $\mathrm{OPT} \leq 1/2$, since one of the constant classifiers that predict 0 and 1 can achieve error at most $1/2$. Consequently, we have that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq K\,O(\mathrm{OPT}\log(1/\mathrm{OPT})) + O(\epsilon)$, since $x\log(1/x+1) = \Theta(x\log(1/x))$ for all $x \in (0, 1/2]$. □

For the class of intersections of $K$ halfspaces, we adapt the general algorithm Algorithm 27 to perform polynomial regression with polynomials of degree at most 2 in Line 8 of Algorithm 27. This degree suffices for learning intersections of halfspaces, as demonstrated in Proposition 10.4.2. Moreover, the regression task can be efficiently reduced to covariance estimation with respect to the Frobenius norm because the entries of the covariance matrix correspond directly to the Hermite coefficients. Consequently, the regression can be executed in polynomial time using $O(d^2)$ samples, so that the overall sample complexity of the algorithm scales as $d^2$ rather than $d^{O(1)}$.

The following proposition shows that the class of intersections of halfspaces satisfies the regularity conditions in Definition 10.3.2.

**Proposition 10.4.2.** *Let $d, K \in \mathbb{Z}_+, \zeta, \epsilon \in (0,1)$. Let $\mathcal{F}_{K,d}$ be the class of $(2, \zeta, \epsilon, K, KO(\zeta\log(1/\zeta+1)+\epsilon), \mathrm{poly}(\epsilon/K), O(\sqrt{\log K}))$-well-behaved Multi-Index Models according to Definition 10.3.2.*

*Proof.* Let $f \in \mathcal{F}_{K,d}$ and denote by $W$ the subspace spanned by the normal vectors of the halfspaces that define $f$, which has dimension at most $K$. From [KOS08, CCK17]

we have that the Gaussian surface area of $\mathcal{F}_{K,d}$ is bounded by $O(\sqrt{\log(K)})$. It remains to show that Condition 3 of Definition 10.3.2 is satisfied for the class $\mathcal{F}_{K,d}$.

Let $D$ be a distribution over $\mathbb{R}^d \times \{0,1\}$ ,where the $\mathbf{x}$-marginals of $D$ are standard normal, such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] \leq \zeta$. Let $V$ be a subspace of $\mathbb{R}^d$, and define $\tau(x) = x\log(1/x+1)$. Assume that for every function $g : V \to \{0,1\}$ it holds that $\mathbf{Pr}[f(\mathbf{x}) \neq g(\mathbf{x}^V)] > CK\tau(\zeta) + 2\epsilon$, for a sufficiently large constant $C > 0$. In particular, consider the function $h(\mathbf{x}) = \mathrm{argmin}_{i\in\{0,1\}} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq i \mid \mathbf{x}^V]$ which depends only on $\mathbf{x}^V$ (i.e., $h(\mathbf{x}) = h(\mathbf{x}^V)$). Then the same lower bound applies to $h$, hence $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] > (C-4)K\tau(\zeta) + 2\epsilon$, since $\tau(x) \geq x/4$ for all $x \in (0,1]$.

We first show that if the subspace $V$ is far from optimal, then when we partition the space into cylinders determined by points in $V$, the target function is far from constant on a nontrivial fraction of these cylinders. Specifically, we show the following

**Claim 10.4.3** (Fraction of Good Points)**.** *With probability at least $\epsilon$ over a random draw of $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ (independent of $\mathbf{x}$), the following holds: for every $i \in \{0,1\}$,*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y \neq i \mid \mathbf{x}^V = \mathbf{x}_0^V] > (C-4)K\,\tau\left(\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y \mid \mathbf{x}^V = \mathbf{x}_0^V]\right) + \epsilon\,,$$

*and $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y \mid \mathbf{x}^V = \mathbf{x}_0^V] \leq 4/(C-4)$.*

*Proof of Claim 10.4.3.* Assume for contradiction that the set $U$ that contains $\mathbf{x}_0 \in \mathbb{R}^d$ such that for all $i \in \{0,1\}$ it holds that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[y \neq i \mid \mathbf{x}^V = \mathbf{x}_0^V] > (C-4)K\tau\left(\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y \mid \mathbf{x}^V = \mathbf{x}_0^V]\right) + \epsilon$$

has probability mass less than $\epsilon$, i.e., $\mathbf{Pr}[x \in U] < \epsilon$. Then, by the definition of the

piecewise constant classifier we can write

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] = \mathop{\mathbf{E}}_{\mathbf{x}_0^V}\left[\min_{i\in\{0,1\}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}\left[i \neq y \mid \mathbf{x}^V = \mathbf{x}_0^V\right]\right].$$

Note that for every $\mathbf{x}_0^V$ not in $U$, it holds that

$$\min_{i\in\{0,1\}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}\left[i \neq y \mid \mathbf{x}^V = \mathbf{x}_0^V\right] \leq (C-4)K\,\tau\left(\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}\left[f(\mathbf{x}) \neq y \mid \mathbf{x}^V = \mathbf{x}_0^V\right]\right) + \epsilon.$$

Therefore, by taking the expectation over $\mathbf{x}_0^V$ and applying Jensen's inequality (noting that $\tau(x) = x\log(1/x+1)$ is concave for all $x > 0$), we obtain

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] &\leq (C-4)K \mathop{\mathbf{E}}_{\mathbf{x}_0^V}\left[\tau\left(\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}\left[f(\mathbf{x}) \neq y \mid \mathbf{x}^V = \mathbf{x}_0^V\right]\right)\right] + \epsilon + \mathbf{Pr}[\mathbf{x}^V \in U] \\
&\leq (C-4)K\,\tau\left(\mathop{\mathbf{E}}_{\mathbf{x}_0^V}\left[\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}\left[f(\mathbf{x}) \neq y \mid \mathbf{x}^V = \mathbf{x}_0^V\right]\right]\right) + 2\epsilon \\
&= (C-4)K\,\tau\left(\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y]\right) + 2\epsilon\,,
\end{aligned}$$

which contradicts the assumed error lower bound for $h$. Therefore, it holds that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbf{x} \in U] \geq \epsilon$. Finally, note that for every $\mathbf{x}_0$ in $U$ we have, by the definition of $U$,

$$(C-4)K\,\tau\left(\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}\left[f(\mathbf{x}) \neq y \mid \mathbf{x}^V = \mathbf{x}_0^V\right]\right) + \epsilon \leq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D}\left[y \neq i \mid \mathbf{x}^V = \mathbf{x}_0^V\right] \leq 1.$$

By using that $\tau(x) \geq x/4$ for all $x \in (0,1]$ we deduce the desired claim. □

Fix a point $\mathbf{x}_0 \in \mathbb{R}^d$ that satisfies the statement of Claim 10.4.3. Let $D_{V^\perp}$ be a distribution supported on $V^\perp \times \{0,1\}$ defined as $D_{V^\perp}(\mathbf{x}^{\perp V}, y) = D(\mathbf{x}, y \mid \mathbf{x}^V = \mathbf{x}_0^V)$. By the properties of the Gaussian distribution, we have that the $\mathbf{x}$-marginal of $D_{V^\perp}$ is a

standard Gaussian distribution. Moreover, by the definition of $D_{V^\perp}$, we have that for $f'(\mathbf{x}) = f(\mathbf{x}^{\perp V} + \mathbf{x}_0^V)$ and all $i \in \{0,1\}$ it holds that

$$\begin{aligned}\Pr_{(\mathbf{x},y)\sim D_{V^\perp}}[y \neq i] &> (C-4)K\tau(\Pr_{(\mathbf{x},y)\sim D_{V^\perp}}[f'(\mathbf{x}) \neq y]) + \epsilon \\ &\geq ((C-4)K/4)\Pr_{(\mathbf{x},y)\sim D_{V^\perp}}[f'(\mathbf{x}) \neq y]\log(1/\Pr_{(\mathbf{x},y)\sim D_{V^\perp}}[f'(\mathbf{x}) \neq y]) + \epsilon\end{aligned}$$

and $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f'(\mathbf{x}) \neq y] \leq 4/(C-4)$. Finally, applying Proposition 10.4.4 for every intersection of halfspaces $f'(\mathbf{x}) = f(\mathbf{x}^{\perp V} + \mathbf{x}_0^V)$ for $\mathbf{x}_0^V$ that satisfies the statement of Claim 10.4.3 completes the proof of Proposition 10.4.2. □

Below, we state and prove a structural result for intersections of halfspaces. In particular, we show that if an intersection of halfspaces is far from being a constant function, then either its first moment or its second moment exhibits a non-trivial correlation with the weight vectors that define it. This structural property is crucial for our algorithm, as it guarantees that at every iteration we can identify a direction along which to improve our approximation.

**Proposition 10.4.4** (Structural Result for Intersections of Halfspaces)**.** *Let $d \in \mathbb{Z}_+$, $\epsilon, \mathrm{OPT} \in (0,1)$, and let $D$ be a distribution over $\mathbb{R}^d \times \{0,1\}$ whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Let $f \in \mathcal{F}_{K,d}$ be a classifier with error $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y] = \mathrm{OPT}$, and assume that $f$ can be expressed in the form $f(\mathbf{x}) = \prod_{i\in[K]} \mathbb{1}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i \geq 0)$ with $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)} \in \mathbb{R}^d, t_1, \ldots, t_K \in \mathbb{R}$. Let $C > 0$ be a sufficiently large universal constant. If it holds that $\min(\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq 0], \mathbf{Pr}_{(\mathbf{x},y)\sim D}[y \neq 1]) \geq CK\mathrm{OPT}\log(1/\mathrm{OPT}) + \epsilon$ and $\mathrm{OPT} \leq 1/C$, then there exists some $i \in [K]$ such that*

$$\max\left(\left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\left[\mathbf{w}^{(i)} \cdot \mathbf{x}y\right]\right|, \left|\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\left[\left((\mathbf{w}^{(i)} \cdot \mathbf{x})^2 - 1\right)y\right]\right|\right) \geq \mathrm{poly}(1/K, \epsilon)\,.$$

*Proof.* Let $W = \text{span}(\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(K)})$ and without loss of generality assume that $\left\|\mathbf{w}^{(i)}\right\| = 1$ for all $i \in [K]$, since normalizing the weight vectors does not change the output of $f$. Assume that for all $i \in [K]$ the first moments are small, i.e., $|\mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{w}^{(i)} \cdot \mathbf{x}y]| \leq \epsilon p(\epsilon/K)$, where $p$ is a sufficiently small universal polynomial (to be quantified later). We show that there exists $i \in [K]$ such that the second moment is large, i.e., $|\mathbf{E}_{(\mathbf{x},y)\sim D}[((\mathbf{w}^{(i)} \cdot \mathbf{x})^2 - 1)y]| \geq \text{poly}(\epsilon/K)$ for a sufficiently small universal polynomial. We can decompose the second-moment matrix as follows:

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(\mathbf{x}\mathbf{x}^\intercal - \mathbf{I})y] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(\mathbf{x}\mathbf{x}^\intercal - \mathbf{I})f(\mathbf{x})] + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(\mathbf{x}\mathbf{x}^\intercal - \mathbf{I})(\mathbb{1}(f(\mathbf{x}) \neq y, y = 1) - \mathbb{1}(y \neq f(\mathbf{x}), f(\mathbf{x}) = 1))] \\
&= \underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(\mathbf{x}\mathbf{x}^\intercal - \mathbf{I})f(\mathbf{x})]}_{\mathbf{I}_1} + \underbrace{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(\mathbf{x}\mathbf{x}^\intercal - \mathbf{I})(T(\mathbf{x},y) - S(\mathbf{x},y))]}_{\mathbf{I}_2},
\end{aligned}
$$

where $S(\mathbf{x},y) \stackrel{\text{def}}{=} \mathbb{1}(y \neq f(\mathbf{x}), f(\mathbf{x}) = 1), T(\mathbf{x},y) \stackrel{\text{def}}{=} \mathbb{1}(f(\mathbf{x}) \neq y, y = 1)$. Note that $\mathbf{I}_1$ is the second moment of $f(\mathbf{x})$, i.e., the noiseless classifier, and $\mathbf{I}_2$ is the contribution corresponding to the noise.

First, we show the following tail bounds, which will help us bound the contribution corresponding to the noise.

**Claim 10.4.5.** *Let $\epsilon \in (0,1)$ less than a sufficiently small universal constant and let $R : \mathbb{R}^d \to [0,1]$ be a function such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[R(\mathbf{x})] = \epsilon$. Then, for every unit vector $\mathbf{v} \in \mathbb{R}^d$, the following bounds hold:*

*i)* $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[R(\mathbf{x})(\mathbf{v} \cdot \mathbf{x})^2] \lesssim \epsilon \log(1/\epsilon)$.

*ii)* $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[R(\mathbf{x})(\mathbf{v} \cdot \mathbf{x})] \lesssim \epsilon\sqrt{\log(1/\epsilon)}$.

*Proof.* First, we will prove (i) and then we will apply it to prove (ii). Note that the expectation $\mathbf{E}[(\mathbf{v}\cdot\mathbf{x})^2 R(\mathbf{x})]$ is minimized if $R$ puts all of its mass at the tails of the function $(\mathbf{v}\cdot\mathbf{x})^2$, i.e., $R(\mathbf{x}) = \mathbb{1}((\mathbf{v}\cdot\mathbf{x})^2 \geq y_\epsilon)$, where $y_\epsilon$ is the smallest number such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbb{1}((\mathbf{v}\cdot\mathbf{x})^2 \geq y_\epsilon)] \leq \epsilon$. Hence, we have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(\mathbf{v}\cdot\mathbf{x})^2 R(\mathbf{x})] &\leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(\mathbf{v}\cdot\mathbf{x})^2 \mathbb{1}((\mathbf{v}\cdot\mathbf{x})^2 \geq y_\epsilon)] \\
&= \int_{\mathbb{R}^d} (\mathbf{v}\cdot\mathbf{x})^2 \mathbb{1}((\mathbf{v}\cdot\mathbf{x})^2 \geq y_\epsilon)\phi(\mathbf{x})d\mathbf{x} \\
&= \frac{1}{\sqrt{2\pi}} \int_{\sqrt{y_\epsilon}}^{\infty} x^2 e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} y_\epsilon e^{-y_\epsilon/2} + \epsilon \qquad (10.4) \\
&\leq \epsilon \log(1/\epsilon) + \epsilon \lesssim \epsilon \log(1/\epsilon)\ , \qquad (10.5)
\end{aligned}$$

where in (10.4) we simply used integration by parts. In (10.5) we used the fact that the function $xe^{-x/2}$ is decreasing for $x \geq 2$. Note that, by Fact 10.1.7, we have $y_\epsilon \gtrsim \log(1/\epsilon) \geq 2$, and since $\epsilon$ is assumed to be sufficiently small.

Finally, we can prove (ii) by simply using Cauchy-Schwarz

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbf{v}\cdot\mathbf{x}R(\mathbf{x})]^2 \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(\mathbf{v}\cdot\mathbf{x})^2 R(\mathbf{x})] \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[R(\mathbf{x})] \lesssim \epsilon^2 \log(1/\epsilon)\ ,$$

which concludes the proof of Claim 10.4.5. □

We next show that there exists a direction within $W$ along which the second moment of $f(\mathbf{x})$ is small. More precisely, we have the following:

**Claim 10.4.6.** *There exists an $i \in [K]$ such that*

$$(\mathbf{w}^{(i)})^\intercal \mathbf{I}_1 \mathbf{w}^{(i)} \lesssim -\frac{\mathbf{Pr}[f(\mathbf{x}) = 1]\,\mathbf{Pr}[f(\mathbf{x}) \neq 1]}{K} + p\left(\frac{\epsilon}{K}\right) + \mathrm{OPT}\log(1/\mathrm{OPT})\ .$$

*Proof.* Denote by $\mu \stackrel{\text{def}}{=} \mathbf{E}[\mathbf{x} f(\mathbf{x})]/\,\mathbf{Pr}[f(\mathbf{x}) = 1]$ and define $\bar{\mathbf{x}} \stackrel{\text{def}}{=} \mathbf{x} - \mu$. Let $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\bar{\mathbf{x}}\bar{\mathbf{x}}^\intercal f(x)]$. For any defining vector $\mathbf{w}^{(i)}$, by Fact 10.5.4 we have

$$(\mathbf{w}^{(i)})^\intercal \mathbf{M} \mathbf{w}^{(i)} \leq \mathbf{Pr}[f(\mathbf{x}) = 1]\Big(1 - ce^{-(\max(0,t_i))^2/2}\Big)\ ,$$

for a sufficiently small universal constant $c > 0$. Next, we claim that there exists some $i \in [K]$ for which the bias $t_i$ is at most $2\sqrt{\ln(K/\mathbf{Pr}[f(\mathbf{x}) \neq 1])}$. To see this, assume for contradiction that for every $i \in [K]$ we have $t_i > 2\sqrt{\ln(\frac{K}{\mathbf{Pr}[f(\mathbf{x})\neq 1]})}$ and note that if $t_i > 2\sqrt{\ln(\frac{K}{\mathbf{Pr}[f(\mathbf{x})\neq 1]})}$ then $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i \leq 0] < \mathbf{Pr}[f(\mathbf{x}) \neq 1]/K$. By applying the union bound, we obtain

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) \neq 1] \leq \sum_{i=1}^{K} \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbf{w}^{(i)} \cdot \mathbf{x} + t_i \leq 0] < \mathbf{Pr}[f(\mathbf{x}) \neq 1]\ ,$$

which is a contradiction. Hence, there must exist some index $i \in [K]$ with $(\mathbf{w}^{(i)})^\intercal \mathbf{M} \mathbf{w}^{(i)} \leq \mathbf{Pr}[f(\mathbf{x}) = 1](1 - c\,\mathbf{Pr}[f(\mathbf{x}) \neq 1]/K)$. Fix such an index $i \in [K]$. Then, by definition,

$$(\mathbf{w}^{(i)})^\intercal \mathbf{I}_1\, \mathbf{w}^{(i)} = \ (\mathbf{w}^{(i)})^\intercal \mathbf{M}\, \mathbf{w}^{(i)} - \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) = 1] + \frac{\left(\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\mathbf{w}^{(i)} \cdot \mathbf{x}\, f(\mathbf{x})\right]\right)^2}{\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) = 1]}.$$

Substituting the bound on $(\mathbf{w}^{(i)})^\intercal \mathbf{M}\, \mathbf{w}^{(i)}$, we obtain

$$(\mathbf{w}^{(i)})^\intercal \mathbf{I}_1\, \mathbf{w}^{(i)} \leq -c\,\frac{\mathbf{Pr}[f(\mathbf{x}) = 1]\ \mathbf{Pr}[f(\mathbf{x}) \neq 1]}{K} + \frac{\left(\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\mathbf{w}^{(i)} \cdot \mathbf{x}\, f(\mathbf{x})\right]\right)^2}{\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) = 1]}.$$

Note that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbf{w}^{(i)}\cdot\mathbf{x}f(\mathbf{x})] = \mathbf{E}[\mathbf{w}^{(i)}\cdot\mathbf{x}y] + \mathbf{E}[\mathbf{w}^{(i)}\cdot\mathbf{x}(S(\mathbf{x},y)-T(\mathbf{x},y))]$. Recall that by assumption we have that $|\mathbf{E}_{(\mathbf{x},y)\sim D}[\mathbf{w}^{(i)}\cdot\mathbf{x}y]| \leq \epsilon p(\epsilon/K)$ for all $i \in [K]$. Moreover, using the triangle inequality $\mathbf{Pr}[f(\mathbf{x})=1] \geq \mathbf{Pr}[y=1] - \mathbf{Pr}[f(\mathbf{x})\neq y]$ together with the assumption in Proposition 10.4.4 that $\mathbf{Pr}[y=1] \geq CK\mathrm{OPT}\log(1/\mathrm{OPT}) + \epsilon$, it follows that $\mathbf{Pr}[f(\mathbf{x})=1] \geq \mathrm{OPT} + \epsilon$. Thus,

$$\mathbf{E}[\mathbf{w}^{(i)}\cdot\mathbf{x}y]^2/\,\mathbf{Pr}[f(\mathbf{x})=1] \leq p(\epsilon/K)\ .$$

Next, observe that $|\mathbf{E}[\mathbf{w}^{(i)}\cdot\mathbf{x}S(\mathbf{x},y)]| = |\mathbf{E}[\mathbf{w}^{(i)}\cdot\mathbf{x}R(\mathbf{x})]|$, where we define $R(\mathbf{x}) = \mathbf{E}_{(\mathbf{x},y)\sim D}[S(\mathbf{x},y)\mid\mathbf{x}]$. Since, for all $\mathbf{x}\in\mathbb{R}^d$ it holds that $R(\mathbf{x})\in[0,1]$, and $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[R(\mathbf{x})] \leq \mathrm{OPT}$, by Claim 10.4.5 we have that $|\mathbf{E}[(\mathbf{w}^{(i)}\cdot\mathbf{x})S(\mathbf{x},y)]| \lesssim \mathrm{OPT}\sqrt{\log(1/\mathrm{OPT})}$. Consequently, since $\mathbf{Pr}[f(\mathbf{x})=1] \geq \mathrm{OPT}$, it follows that $\mathbf{E}[\mathbf{w}^{(i)}\cdot\mathbf{x}S(\mathbf{x},y)]^2/\,\mathbf{Pr}[f(\mathbf{x})=1] \lesssim \mathrm{OPT}\log(1/\mathrm{OPT})$. A similar argument shows that $\mathbf{E}[\mathbf{w}^{(i)}\cdot\mathbf{x}T(\mathbf{x},y)]^2/\,\mathbf{Pr}[f(\mathbf{x})=1] \lesssim \mathrm{OPT}\log(1/\mathrm{OPT})$. Combining these bounds, we obtain

$$(\mathbf{w}^{(i)})^{\intercal}\mathbf{I}_1\mathbf{w}^{(i)} \lesssim -\frac{\mathbf{Pr}[f(\mathbf{x})=1]\,\mathbf{Pr}[f(\mathbf{x})\neq 1]}{K} + p\left(\frac{\epsilon}{K}\right) + \mathrm{OPT}\log(1/\mathrm{OPT})\ ,$$

which concludes the proof of Claim 10.4.6. □

Now we bound the contribution corresponding to the noise, using Claim 10.4.5.

**Claim 10.4.7.** *For every unit vector $\mathbf{v}\in\mathbb{R}^d$, it holds that $|\mathbf{v}^{\intercal}\mathbf{I}_2\mathbf{v}| = O(\mathrm{OPT}\log(1/\mathrm{OPT}))$.*

*Proof.* Let $\mathbf{v}\in\mathbb{R}^d$ be a unit vector, and denote by $R(\mathbf{x}) = \mathbf{E}_{(\mathbf{x},y)\sim D}[S(\mathbf{x},y)\mid\mathbf{x}]$. Note that $R(\mathbf{x})\in[0,1]$ for all $\mathbf{x}\in\mathbb{R}^d$, and moreover $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[R(\mathbf{x})] \leq \mathrm{OPT}$. Hence, by Claim 10.4.5 we have that $|\mathbf{E}_{(\mathbf{x},y)\sim D}[(\mathbf{v}\cdot\mathbf{x})^2S(\mathbf{x},y)]| = O(\mathrm{OPT}\log(1/\mathrm{OPT}))$. By using similar arguments we have that $|\mathbf{E}_{(\mathbf{x},y)\sim D}[(\mathbf{v}\cdot\mathbf{x})^2T(\mathbf{x},y)]| = O(\log(1/\mathrm{OPT})\mathrm{OPT})$.

Therefore, by triangle inequality, we have that

$$|\mathbf{v}^\intercal \mathbf{I}_2 \mathbf{v}| \le |\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(\mathbf{v}\cdot\mathbf{x})^2 S(\mathbf{x},y)]| + |\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(\mathbf{v}\cdot\mathbf{x})^2 T(\mathbf{x},y)]| + \mathbf{E}[S(\mathbf{x},y)+T(\mathbf{x},y)]$$
$$= O(\mathrm{OPT}\log(1/\mathrm{OPT}))\ ,$$

which concludes the proof of Claim 10.4.7. □

Denote by $t$ the quantity $t \coloneqq C'\mathrm{OPT}\log(1/\mathrm{OPT})$, for sufficiently large constant $C'$. Combining Claims 10.4.6 and 10.4.7, we have that there exists $i \in [K]$ such that

$$\begin{aligned}(\mathbf{w}^{(i)})^\intercal \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}[(\mathbf{x}\mathbf{x}^\intercal - \mathbf{I})y](\mathbf{w}^{(i)}) &\lesssim -\mathbf{Pr}[f(\mathbf{x})=1]\,\mathbf{Pr}[f(\mathbf{x})\neq 1]/K - t + p(\epsilon/K)\\ &\le -\left(\min_{i\in\{0,1\}}(\mathbf{Pr}[y=i]) - \mathrm{OPT}\right)/(2K) - t + p(\epsilon/K)\\ &\le -\epsilon/(2K) + p(\epsilon/K)\ ,\end{aligned}$$

where we used the fact that one of $\mathbf{Pr}[f(\mathbf{x})=1]$ or $\mathbf{Pr}[f(\mathbf{x})\neq 1]$ must be at least $1/2$, and the assumption of Proposition 10.4.4 that for every label $i \in \{0,1\}$ it holds that $\mathbf{Pr}[y=i] \ge CK\mathrm{OPT}\log(1/\mathrm{OPT}) + \epsilon$, for a sufficiently large constant $C$. Setting $p(\epsilon/K) \le c\epsilon/K$ for a sufficiently small universal constant $c > 0$, we ensure that the error term $p(\epsilon/K)$ is negligible compared to the main term. This concludes the proof of Proposition 10.4.4. □

## 10.5 Omitted Content from Sections 10.2 to 10.4

The following claim shows that if a matrix has a small Frobenius norm but exhibits a large quadratic form in some direction, then by selecting those eigenvectors whose eigenvalues

exceed a given threshold, one obtains a short list of vectors among which at least one correlates well with that direction. This claim is used for proving Propositions 10.2.8 and 10.3.7.

**Claim 10.5.1.** *Let $\mathbf{M} \in \mathbb{R}^{d\times d}$ a symmetric positive semi-definite (PSD) matrix and let $\mathbf{v} \in \mathbb{R}^d$ with $\|\mathbf{v}\| \leq 1$ such that $\mathbf{v}^\intercal \mathbf{M}\mathbf{v} \geq \alpha$. Then there exists a unit eigenvector $\mathbf{u}$ of $\mathbf{M}$ with eigenvalue at least $\alpha/2$ such that $|\mathbf{u}\cdot\mathbf{v}| \gtrsim (\alpha/\|\mathbf{M}\|_F)^{3/2}$ . Moreover, the number of eigenvectors of $\mathbf{M}$ with eigenvalue greater than $\alpha/2$ is at most $4\|\mathbf{M}\|_F/\alpha^2$ .*

*Proof.* From the spectral theorem, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ be the eigenvalues of $\mathbf{M}$ with their corresponding orthonormal eigenvectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(d)} \in \mathbb{R}^d$. Since $\mathbf{M}$ is symmetric and psd, it can be decomposed as $\mathbf{M} = \sum_{i=1}^d \lambda_i \mathbf{x}^{(i)}(\mathbf{x}^{(i)})^\intercal$. Moreover, we can write $\mathbf{v}$ in the eigenbasis of $\mathbf{M}$, i.e., $\mathbf{v} = \sum_{i=1}^d (\mathbf{v}\cdot\mathbf{x}^{(i)})\mathbf{x}^{(i)}$. Therefore, we have that $\mathbf{v}^\intercal\mathbf{M}\mathbf{v} = \sum_{i=1}^d \lambda_i(\mathbf{v}\cdot\mathbf{x}^{(i)})^2$, and using that $\mathbf{v}^\intercal\mathbf{M}\mathbf{v} \geq \alpha$, we get that $\sum_{i=1}^n \lambda_i(\mathbf{v}\cdot\mathbf{x}^{(i)})^2 \geq \alpha$.

Let $S = \{i : \lambda_i \geq \frac{\alpha}{2}\}$. Using the fact that $\sum_{i=1}^d (\mathbf{v}\cdot\mathbf{x}^{(i)})^2 \leq 1$, we have that

$$\sum_{i=1}^d \lambda_i(\mathbf{v}\cdot\mathbf{x}^{(i)})^2 = \sum_{i\notin S} \lambda_i(\mathbf{v}\cdot\mathbf{x}^{(i)})^2 + \sum_{i\in S} \lambda_i(\mathbf{v}\cdot\mathbf{x}^{(i)})^2 \leq \frac{\alpha}{2} + \sum_{i\in S} \lambda_i(\mathbf{v}\cdot\mathbf{x}^{(i)})^2 \,.$$

Hence, we have that $\sum_{i\in S} \lambda_i(\mathbf{v}\cdot\mathbf{x}^{(i)})^2 \geq \frac{a}{2}$. Furthermore, note that there must be at least one index $j \in S$ such that $\lambda_j(\mathbf{v}\cdot\mathbf{x}^{(j)})^2 \geq \frac{a}{2|S|}$. Also, from the fact that $\lambda_j \leq \|\mathbf{M}\|_F$, we have that $(\mathbf{v}\cdot\mathbf{x}^{(j)})^2 \geq \frac{a}{2|S|\|\mathbf{M}\|_F}$. Taking the sum of the squares of the eigenvalues in $S$, we get that

$$|S|\cdot\left(\frac{\alpha}{2}\right)^2 \leq \sum_{i\in S}\lambda_i^2 \leq \|\mathbf{M}\|_F^2 \Rightarrow |S| \leq \frac{4\|\mathbf{M}\|_F^2}{\alpha^2} \,.$$

Combining the above, we have that $\mathbf{v}\cdot\mathbf{x}^{(j)} \gtrsim (\alpha/\|\mathbf{M}\|_F)^{3/2}$, for some $j \in S$, and $|S| \leq 4\|\mathbf{M}\|_F/\alpha^2$, which completes the proof of Claim 10.5.1. □

The following two facts are important properties of polynomial regression which are be used in Proposition 10.3.7 for the analysis of our general algorithm. Fact 10.5.2 shows that by using a sufficiently small number of samples in polynomial time, one can efficiently obtain a low-degree polynomial whose $L_2$ approximation error is within an additive $\epsilon$ of the best achievable by any polynomial of the same degree. Fact 10.5.3 shows that the subspace spanned by the eigenvectors of a low-degree regression polynomial contains only a few eigenvectors corresponding to large eigenvalues.

**Fact 10.5.2** (see, e.g., Lemma 3.3 in [DKK+21a])**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ whose $x$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Let $k \in \mathbb{Z}_+$ and $\epsilon, \delta > 0$. There is an algorithm that draws $N = (dk)^{O(k)} \log(1/\delta)/\epsilon^2$ samples from $\mathcal{D}$, runs in time poly$(N, d)$, and outputs a polynomial $P(\mathbf{x})$ of degree at most $k$ such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + \epsilon,$$

*with probability $1 - \delta$.*

**Fact 10.5.3** (see, e.g., Lemma 3.3 in [DKK+21a])**.** *Fix $\epsilon \in (0, 1)$ and let $P(\mathbf{x})$ be a degree-$k$ polynomial, such that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P(\mathbf{x}))^2] \leq \min_{P' \in \mathcal{P}_k} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - P'(\mathbf{x}))^2] + O(\epsilon).$$

*Let $\mathbf{M} = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\nabla P(\mathbf{x}) \nabla P(\mathbf{x})^\intercal]$ and $V$ be the subspace spanned by the eigenvectors of $\mathbf{M}$ with eigenvalues larger than $\eta$. Then the dimension of the subspace $V$ is $\dim(V) = O(k/\eta)$ and moreover $\mathrm{tr}(\mathbf{M}) = O(k)$.*

Fact 10.5.4 shows that for an intersection of halfspaces, the variance in any defining direction is reduced from one by an amount that depends on how compressed the

function is along that direction. This fact is the key ingredient in our structural result on intersections of halfspaces (see Proposition 10.4.4).

**Fact 10.5.4** (Gaussian Monotonicity, [Vem10a]). *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be an intersection of halfspaces and let $\mathcal{N}_d|_{\mathcal{K}}$ be the truncation of the standard Gaussian distribution in $d$ dimensions $\mathcal{N}_d$ to $\mathcal{K}$. For any $u \in \mathbb{S}^{d-1}$, we have $\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}(\mathbf{u}\cdot\mathbf{x}) \leq 1$. Moreover, if for some $T \in \mathbb{R}$ the halfspace $\{\mathbf{x} : \mathbf{u}\cdot\mathbf{x} + T \geq 0\}$ is one of the defining halfspaces of the intersection then, we have variance reduction along $u$, i.e., $\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}_d|_{\mathcal{K}}}(\mathbf{u}\cdot\mathbf{x}) \leq 1 - \frac{1}{2}e^{-\frac{1}{2}\max\{0,T\}^2}$ for a sufficiently large universal constant $C > 0$.*

### 10.5.1 Related Complexity Measures

Several complexity measures have been proposed for learning real-valued concepts in the realizable setting. One measure closely related to our work is the *generative exponent* $k^*$, introduced in [DPVLB24]. The generative exponent $k^*$ is defined as the smallest positive integer $k$ for which $\mathbf{E}_{y\sim D_y}[\zeta_k^2(y)] \neq 0$, where $\zeta_k(y) \coloneqq \mathbf{E}_{x\sim\mathcal{N}(0,1)}[\mathrm{He}_k(x) \mid f(x) = y]$ and $f : \mathbb{R} \to \mathbb{R}$ is a single-index link function. This measure is used for the task of approximating the hidden direction $\mathbf{w}$ of a single-index model $f(\mathbf{w}\cdot\mathbf{x})$ from samples $(\mathbf{x}, f(\mathbf{w}\cdot\mathbf{x}))$ with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Using the measure $k^*$, [DPVLB24] derived nearly tight SQ bounds, with complexity of order $d^{k^*/2}$. (Technically speaking, their upper bound requires knowledge of $\zeta_{k^*}(y)$ or the existence of an orthonormal polynomial basis with respect to the marginal distribution over the labels $D_y$, where $\zeta_{k^*}$ has a large low-degree component.)

Our definition of well-behaved MIMs (Definition 10.1.2) implicitly introduces a generalization of the generative exponent for PAC learning discrete-valued MIMs in the presence of noise. As we demonstrate in Section (10.5.2), for the special case

of realizable learning of Single-Index models, our complexity measure is essentially equivalent to the generative exponent. We establish this under two additional regularity assumptions: (i) a bound on the GSA of the ground truth concept, and (b) the existence of non-trivial distinguishing moments with respect to an already discovered subspace. (These regularity assumptions hold for several well-studied function classes such as multiclass linear classifiers and intersections of halfspaces.)

Note that our learning algorithm does not require prior knowledge of the moments conditioned on each level-set for two reasons: (i) when the number of labels is finite, we can leverage the moments of all level-sets efficiently; and (ii) the bounded GSA assumption allows us to partition an already discovered subspace into finite-width regions to uncover additional directions using moments. Moreover, by considering the existence of distinguishing moments only when our approximation to the target ground truth subspace is insufficient, our notion allows for relaxed approximation guarantees that yield more efficient algorithms.

Other relevant complexity measures that have gained attention recently are the *information exponent* [BAGJ21] and the *leap complexity* [AAB+21, AAM22, AAM23]. The information exponent is defined as the smallest integer $k$ for which a single-index link function $f : \mathbb{R} \to \mathbb{R}$ has a non-zero Hermite coefficient, i.e., the minimum $k$ such that $\mathbf{E}_{x\sim\mathcal{N}(0,1)}[f(x)\mathrm{He}_k(x)] \neq 0$. The leap complexity generalizes this notion to real-valued MIMs: we say that a function $f : \mathbb{R}^K \to \mathbb{R}$ has leap complexity $\mathrm{Leap}(f) = k$ if every Hermite decomposition of $f$ can be ordered such that each consecutive term introduces new directions via a polynomial of degree at most $k$. Since these measures rely solely on the non-negativity of the Hermite coefficients and do not incorporate label-based information (e.g., via level-sets), they are closely related to the CSQ complexity, which

can be arbitrarily larger than the SQ complexity.

### 10.5.2 Near-equivalence to Generative Exponent for Single-Index Case

In this section, we compare our definition of well-behaved MIMs (Definition 10.1.2) to the generative exponent (Definition 10.5.5) introduced in [DPVLB24]. We first give a formal definition of the generative exponent:

**Definition 10.5.5** (Generative exponent)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a single index link function. Let $D$ be the joint distribution of $(x, f(x))$, where $x \sim \mathcal{N}(0,1)$. The generative exponent $k^*$ is defined as the smallest positive integer $k$ for which $\mathbf{E}_{y\sim D_y}[\zeta_k^2(y)] \neq 0$, where $\zeta_k(y) \coloneqq \mathbf{E}_{x\sim\mathcal{N}(0,1)}[\mathrm{He}_k(x) \mid f(x) = y]$.*

As we have mentioned in the previous subsection, the generative exponent focuses on subspace recovery (in the realizable setting) rather than accurate label prediction. Hence, in order for the two notions to be comparable, we first slightly modify our definition of Definition 10.1.2 to this setting.

**Definition 10.5.6** (Well-Behaved MIMs)**.** *Fix $m \in \mathbb{Z}_+, \sigma \in (0,1)$ and $\mathcal{Y} \subseteq \mathbb{Z}$ of finite cardinality. Let $f : \mathbb{R}^d \to \mathcal{Y}$ be a $K$-MIM function with hidden subspace $W$. We say that a $K$-MIM function $f : \mathbb{R}^d \to \mathcal{Y}$ is $m$-well-behaved MIM if for any subspace $V \subseteq \mathbb{R}^d$ the following conditions holds:*

*a) $V$ contains $W$, i.e., $W \subseteq V$, or*

*b) there exists a point $\mathbf{x}_0 \in \mathbb{R}^d$, a zero mean, unit-variance polynomial $p : U \to \mathbb{R}$ of*

*degree at most $m$ and $z \in \mathcal{Y}$ such that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[p(\mathbf{x}^U)\mathbb{1}(f(\mathbf{x}) = z) \mid \mathbf{x}^V = \mathbf{x}_0^V] \neq 0 \,,$$

*where $U = (V + W) \cap V^\perp$.*

Compared to Definition 10.1.2, our modified definition of well-behaved MIMs in Definition 10.5.6 introduces the following changes:

(i) We omitted the technical assumption of bounded GSA to direct our attention on the existence of distinguishing moments.

(ii) Instead of requiring that the function be approximately equal to one that depends solely on the projection onto $V$ (as specified in condition (a) of Definition 10.1.2), we now require that $V$ contain the hidden subspace $W$. This change emphasizes subspace recovery in the realizable setting and aligns our definition with that of the generative exponent.

(iii) Since we are focusing on the realizable setting, we consider only the moments of $f$ rather than noisy version of $f$.

In the following lemma, we show that for SIMs, the smallest degree $m$ for which Definition 10.5.6 is satisfied is equal to the generative exponent.

**Lemma 10.5.7.** *Let $\mathcal{Y}$ be a finite cardinality set. Let $f : \mathbb{R}^d \to \mathcal{Y}$ be a SIM, that is, there exists a unit vector $\mathbf{w}$ such that $f(\mathbf{x}) = f(\mathbf{x}^\mathbf{w})$ for all $\mathbf{x} \in \mathbb{R}^d$. The smallest positive integer $m$ for which $f$ is an $m$-well-behaved MIM (as per Definition 10.5.6) is equal to the generative exponent $k^*$ of $f$ (see Definition 10.5.5).*

*Proof.* We first show that $f$ is not an $m$-well-behaved MIM for all positive integers $m < k^*$. Note that since $\mathcal{Y}$ is discrete for all $k \in \mathbb{Z}_+$ we have that

$$\mathop{\mathbf{E}}_{y\sim D_y}[\zeta_k^2(y)] = \sum_{y\in\mathcal{Y}} \zeta_k^2(y) \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(0,1)}[f(x\mathbf{w}) = y] \;,$$

where $\zeta_k(y) \stackrel{\text{def}}{=} \mathbf{E}_{x\sim\mathcal{N}(0,1)}[\mathrm{He}_k(x) \mid f(x\mathbf{w}) = y]$ (see Definition 10.5.5). Hence for each positive integer $m < k^*$ by the definition of the generative exponent for all $y \in \mathcal{Y}$ it holds that

$$\begin{aligned}&\mathop{\mathbf{E}}_{x\sim\mathcal{N}(0,1)}[\mathrm{He}_m(x) \mid f(x\mathbf{w}) = y]^2 \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(0,1)}[f(x\mathbf{w}) = y] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{He}_m(\mathbf{w}\cdot\mathbf{x}) \mid f(\mathbf{x}) = y]^2 \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) = y] = 0 \;.\end{aligned}$$

Otherwise, the generative exponent would be less than $k^*$. Thus, for every $m \in [k^* - 1]$ and $y \in \mathcal{Y}$ either $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{He}_m(\mathbf{w}\cdot\mathbf{x}) \mid f(\mathbf{x}) = y]$ or $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[f(\mathbf{x}) = y] = 0$. In either case, this implies that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{He}_m(\mathbf{w}\cdot\mathbf{x})\mathbb{1}(f(\mathbf{x}) = y)] = 0$, for all $m \in [k^* - 1]$ and $y \in \mathcal{Y}$.

Since any zero-mean polynomial of degree less $k^*$ can be decomposed to a linear combination of non-constant Hermite polynomials of degree less than $k^*$, it follows that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[p(\mathbf{w}\cdot\mathbf{x})\mathbb{1}(f(\mathbf{x}) = y)] = 0$, for any zero-mean polynomial $p$ of degree less than $k^*$. Furthermore, because $f$ is a SIM, for any polynomial $p : \mathbb{R}^d \to \mathbb{R}$ we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[p(\mathbf{x})\mathbb{1}(f(\mathbf{x}) = y)] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[p((\mathbf{w}\cdot\mathbf{x})\mathbf{w})\mathbb{1}(f(\mathbf{x}) = y)] \;.$$

Therefore, for any zero-mean polynomial $p$ of degree less $k^*$ it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[p(\mathbf{x})\mathbb{1}(f(\mathbf{x}) = y)] = 0$, so $f$ does not satisfy the condition (b) of Definition 10.5.6 even for $V = \{0\}$.

This concludes the first part of the proof.

Now to conclude the proof it suffices to show that $f$ is a $k^*$-well-behaved MIM. Note that since $\mathbf{E}_{y\sim D_y}[\zeta_{k^*}^2(y)] \neq 0$ and $\mathcal{Y}$ is discrete, we have that there exists a label $y \in \mathcal{Y}$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{He}_{k^*}(\mathbf{w}\cdot\mathbf{x})\mathbb{1}(f(\mathbf{x}) = y)] \neq 0$. Consider a subspace $V$ of $\mathbb{R}^d$ that does not contain $\mathbf{w}$ and denote by $\mathbf{u}$ the projection of $\mathbf{w}$ onto $V^\perp$. Note that since $V$ does not contain $\mathbf{w}$ it holds that $\|\mathbf{u}\| = \left\|\mathbf{w}^{\perp V}\right\| = \mathbf{w}\cdot\mathbf{u} \neq 0$.

Since $f$ is a SIM ,we have that for any $\lambda \in \mathbb{R}$

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{He}_{k^*}(\lambda\mathbf{u}\cdot\mathbf{x})\mathbb{1}(f(\mathbf{x}) = y)] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{He}_{k^*}(\lambda(\mathbf{w}\cdot\mathbf{u})(\mathbf{w}\cdot\mathbf{x}))\mathbb{1}(f(\mathbf{x}) = y)] \,.$$

Thus, setting $\lambda = 1/(\mathbf{w}\cdot\mathbf{u})$ gives us that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\mathrm{He}_{k^*}\left(\frac{\mathbf{u}\cdot\mathbf{x}}{\mathbf{w}\cdot\mathbf{u}}\right)\mathbb{1}(f(\mathbf{x}) = y)\right] \neq 0 \,.$$

Note that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathrm{He}_{k^*}\left(\mathbf{u}\cdot\mathbf{x}/(\mathbf{w}\cdot\mathbf{u})\right)] = 0$ by a change of variables, since $\mathbf{u}/(\mathbf{w}\cdot\mathbf{u})$ is a unit vector. Thus, normalizing the above polynomial implies the existence of a mean-zero, variance-one polynomial $p : \mathbb{R} \to \mathbb{R}$ of degree $k^*$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[p(\mathbf{u}\cdot\mathbf{x})\mathbb{1}(f(\mathbf{x}) = y)] \neq 0 \,.$$

Therefore, there has to exist an $\mathbf{x}_0^V$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[p(\mathbf{u}\cdot\mathbf{x})\mathbb{1}(f(\mathbf{x}) = y) \mid \mathbf{x}^V = \mathbf{x}_0^V] \neq 0$ , otherwise averaging over $\mathbf{x}_0^V$ would result to the above moment being zero. Consequently, we have that $f$ is a $k^*$-well-behaved MIM, which completes the proof. $\square$

## 10.6 Lower Bound For Single-Shot Optimization

In this section, we present an example in which a single-shot approach fails to learn the target function to sufficiently small error even if it is given access to exactly the first $O(1)$-moments. On the other hand, we show that if we take an iterative approach then we can learn the function in polynomial time using just $O_\epsilon(d^2)$ complexity. Our results build on moment-matching constructions and properties of Legendre polynomials.

**Theorem 10.6.1.** *Let $d \in \mathbb{Z}_+$ and $C_1, C_2 > 0$ be sufficiently large universal constants. There exists a family of functions*

$$\mathcal{F} = \{f(\mathbf{U}\mathbf{x}) : \mathbf{U} \in \mathbb{R}^{2\times d},\ \mathbf{U}\mathbf{U}^\intercal = \mathbf{I}_2\},$$

*such that:*

- *No algorithm given the first $C_1$ exact moments of $f(\mathbf{U}\mathbf{x}) \in \mathcal{F}$ can output another function $f' \in \mathcal{F}$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(f(\mathbf{U}\mathbf{x}) - f'(\mathbf{x}))^2] \le 1/C_2$.*
- *There exists an SQ algorithm that for any $\epsilon > 0$, given SQ oracle access to a distribution $D$ over $\mathbb{R}^d \times \{0,1\}$ (with $\mathbf{x} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and $y = f(\mathbf{U}\mathbf{x})$ for some $f \in \mathcal{F}$), outputs an estimate $\widehat{\mathbf{U}}$ satisfying $\|\widehat{\mathbf{U}} - \mathbf{U}\|_2 \le \epsilon$ using $O(d^2)$ queries to $\mathrm{VSTAT}(C_2 d^2/\epsilon^2)$.*

*Proof.* The proof is based on a careful construction of a two-dimensional distribution and a corresponding function family. Assume that the algorithm is given $C_1 = O(1)$ moments. Let $k = 2C_1$. Define the set $A$ (we use $A$ as the indicator function of this set) on $\mathbb{R}^2$ as in Proposition 10.6.2 and let

$$B(x,y) = \mathbb{1}(|x+y| \ge 2b)\ ,$$

where $b = \Theta(k)$ and is the smallest value such that $A(x, y) = 0$ whenever $|x| \geq b$ or $|y| \geq b$. We then define the boolean function

$$f(x, y) = A(x, y) + B(x, y) \ .$$

Let $\mathcal{Q} \subseteq \mathbb{R}^{d \times 2}$ be a set of matrices that satisfy Lemma 11.2.5, i.e.,

$$|\mathcal{Q}| = 2^{d^{\Omega(1)}}, \quad \mathbf{Q}\mathbf{Q}^\intercal = \mathbf{I}_2, \quad \text{and} \quad \|\mathbf{Q}\mathbf{P}^\intercal\|_F \leq O\Big(d^{-\Omega(1)}\Big) \ .$$

For each $[\mathbf{v}, \mathbf{u}] = \mathbf{Q} \in \mathcal{Q}$, we define $f_{\mathbf{Q}}(\mathbf{x})$ as follows: Let $\mathbf{w}^{(1)} = \mathbf{v}/\sqrt{2} + \mathbf{u}/\sqrt{2}$ and $\mathbf{w}^{(2)} = \mathbf{v}/\sqrt{2} - \mathbf{u}/\sqrt{2}$ and we let $f_{\mathbf{Q}}(\mathbf{x}) = f(\mathbf{w}^{(1)} \cdot \mathbf{x}, \mathbf{w}^{(2)} \cdot \mathbf{x})$. Let $\mathcal{F}_{\mathcal{Q}} = \{f_{\mathbf{Q}}(\mathbf{x}) : \mathbf{Q} \in \mathcal{Q}\} \subseteq \mathcal{F}$. For each $\mathbf{Q} \in \mathcal{Q}$, let $D_{\mathbf{Q}}$ denote the joint distribution of $(\mathbf{x}, f_{\mathbf{Q}}(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Define the class of distributions $\mathcal{D} = \{D_{\mathbf{Q}} : \mathbf{Q} \in \mathcal{Q}\}$.

Now, suppose we have the first $k/2$-moments of a distribution $D$ drawn from $\mathcal{D}$. Note that because $\mathbf{E}[p(\mathbf{x})A(\mathbf{w}^{(1)} \cdot \mathbf{x}, \mathbf{w}^{(2)} \cdot \mathbf{x})] = 0$ for any zero mean mean polynomial $p$ of degree at most $k$, we have that $\mathbf{E}[p(\mathbf{x})f(\mathbf{x})] = \mathbf{E}[p(\mathbf{x})B(\mathbf{x})] = \mathbf{E}[p(\mathbf{x})B(\mathbf{w}^{(1)} \cdot \mathbf{x} + \mathbf{w}^{(2)} \cdot \mathbf{x})] = \mathbf{E}[p(\mathbf{x})B(\mathbf{v} \cdot \mathbf{x}\sqrt{2})]$. Therefore, we can exactly extract from the moments only $\mathbf{w}^{(1)} + \mathbf{w}^{(2)} = \mathbf{v} \cdot \mathbf{x}\sqrt{2}$. Therefore, with access to $O(1)$-moments, we can only extract the direction $\mathbf{v}$. We show that for any vector $\mathbf{w}$ so that $\mathbf{w} \cdot \mathbf{u} \leq c_1$ where $c_1 > 0$ is a sufficiently small absolute constant, then the hypothesis $A(\mathbf{w}^{(3)} \cdot \mathbf{x}, \mathbf{w}^{(4)} \cdot \mathbf{x})$ achieves large $L_2^2$ error where $\mathbf{w}^{(3)} = \mathbf{v}/\sqrt{2} + \mathbf{w}/\sqrt{2}$ and $\mathbf{w}^{(4)} = \mathbf{v}/\sqrt{2} - \mathbf{w}/\sqrt{2}$. First, we choose

$\mathbf{w}$ so that $\mathbf{w} \cdot \mathbf{u} = 0$. Note that from Jensen's inequality it holds that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(A(\mathbf{w}^{(1)} \cdot \mathbf{x}, \mathbf{w}^{(2)} \cdot \mathbf{x}) - A(\mathbf{w}^{(3)} \cdot \mathbf{x}, \mathbf{w}^{(4)} \cdot \mathbf{x}))^2] \geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\mathop{\mathbf{E}}_{\mathbf{v}\cdot\mathbf{x}\sim\mathcal{N}(0,1)}[A(\mathbf{w}^{(1)} \cdot \mathbf{x}, \mathbf{w}^{(2)} \cdot \mathbf{x}) - A(\mathbf{w}^{(3)} \cdot \mathbf{x}, \mathbf{w}^{(4)} \cdot \mathbf{x})]\right)^2\right] , \quad (10.6)$$

where the last term corresponds to the variance over the direction $\mathbf{u}$. Note that $A$ is a 2-dimensional instance that depends only on matching the first $C_1$ moments where $C_1$ is an absolute constant therefore Equation (10.6) (as by construction is non zero) is bounded from below by an absolute constant. Therefore, there exists a sufficiently small absolute constant $c_2 > 0$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\mathop{\mathbf{E}}_{\mathbf{v}\cdot\mathbf{x}\sim\mathcal{N}(0,1)}[A(\mathbf{w}^{(1)} \cdot \mathbf{x}, \mathbf{w}^{(2)} \cdot \mathbf{x}) - A(\mathbf{w}^{(3)} \cdot \mathbf{x}, \mathbf{w}^{(4)} \cdot \mathbf{x})]\right)^2\right] \geq c_2 .$$

Furthermore, note that the $L_2^2$ of $A(\mathbf{w}^{(1)} \cdot \mathbf{x}, \mathbf{w}^{(2)} \cdot \mathbf{x})$ and $A(\mathbf{w}^{(3)} \cdot \mathbf{x}, \mathbf{w}^{(4)} \cdot \mathbf{x})$ distance is a decreasing function with respect to $\theta$, where $\theta = \theta(\mathbf{w}, \mathbf{u})$. Using the fact that the set $\theta \in [0, \pi/2]$ is compact, there exists a sufficiently small absolute constant $c_1 > 0$ such that if $\theta(\mathbf{u}, \mathbf{w}) \geq \pi/2 - c_1$, then the $L_2^2$ of $A(\mathbf{w}^{(1)} \cdot \mathbf{x}, \mathbf{w}^{(2)} \cdot \mathbf{x})$ and $A(\mathbf{w}^{(3)} \cdot \mathbf{x}, \mathbf{w}^{(4)} \cdot \mathbf{x})$ is at least $c_2/2$. Using the fact that $d$ is sufficiently large the probability of finding $\mathbf{w}$ so that $\theta(\mathbf{u}, \mathbf{w}) \leq \pi/2 - c_1$ is at least $\exp(-\Omega(d))$.

On the other hand, there exists an SQ Algorithm with only $O(d^2)$ queries can learn the two directions of $f$. With $O(d)$ queries (with accuracy $O(2^{-k}\epsilon) = O(\epsilon)$) recovers an $\epsilon$ approximation to the direction $\mathbf{w} = (\mathbf{v} + \mathbf{u})/\sqrt{2}$. Next if we conditioning on the event $\{\mathbf{w} \cdot \mathbf{x} \geq \delta\}$ where $\delta > 0$ is chosen so that $A(\mathbf{w}(1) \cdot \mathbf{x}, \mathbf{w}^{(2)} \cdot \mathbf{x})$ does not match moments with respect $\mathbf{u}$, which such exists and it is an absolute constant as it reweighs

the probability mass of the intervals, therefore

$$\mathbf{E}\left[y\,\mathbf{x}\mathbf{x}^{\mathsf{T}} \mid \mathbf{w}\cdot\mathbf{x} \geq \delta\right]$$

reveals an $\epsilon$-approximation to the direction $\mathbf{u}$. This completes the proof. □

### 10.6.1 Moment-Matching Construction

We now describe the constructions used in the proof. These results build piecewise constant functions and moment-matching distributions that are central to our lower bound.

**Proposition 10.6.2.** *Let $c > 1$ be a universal constant and $k \in \mathbb{Z}_+$. There exists a distribution $A$ on $\mathbb{R}^2$, for which the following are true:*

1. *$A$ matches its first $k$ moments with the 2-dimensional standard Gaussian.*
2. *for all $\mathbf{x} \in \mathrm{supp}(A)$ it holds that $|\mathbf{x}_1|, |\mathbf{x}_2| \leq ck$ and $A$ is a union of $k+3$ intervals (in each orthogonal direction).*

*Proof.* Let $x, y$ be two orthogonal directions. Define $A(x,y) = U(x)U(y)/\,\mathbf{Pr}[x \in U]^2$ where $U$ is the set from Proposition 10.6.3. Then, we have that $A$ satisfies Items 1 and 2. □

**Proposition 10.6.3.** *For any $k \in \mathbb{Z}_+$ there exists a set $U \subseteq \mathbb{R}$ with $\mathbf{Pr}_{x\sim\mathcal{N}(0,1)}[x \in U] = 1/2$ such that $U$ is a union of $k$ intervals with $|\max(U)| = O(k)$ and*

$$\mathop{\mathbf{E}}_{y\sim\mathcal{N}(0,1)}\left[y^t \mid y \in U\right] = \mathop{\mathbf{E}}_{y\sim\mathcal{N}(0,1)}[y^t]$$

*for all integers $t = 0, \ldots, k$.*

*Proof.* We first show that there exists a piecewise constant function $f : \mathbb{R} \to \{0, 1\}$ that matches the moments of the standard normal distribution and vanishes for $|x| \geq O(k)$. This is formalized in the following lemma.

**Lemma 10.6.4.** *Fix $k \in Z_+$ and $\eta \in (0, 1)$. There exists a $O(k)$-piecewise-constant function $f : \mathbb{R} \to \{0, 2\}$ such that $f(x) = 0$ for $|x| \geq \Theta(k)$ and for all $t \leq k$ it holds that*

$$\left| \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t f(y)] - \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t] \right| = 0 \, .$$

Note that because $f(x)$ takes values in $\{0, 2\}$, it holds that $\mathbf{Pr}[f(x) = 2] = 1/2$ (which holds from Lemma 10.6.4 for $t = 0$). It also holds that

$$\left| \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t \mid f(y) = 2] - \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t] \right| = \left| \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t f(y)] \frac{1}{2 \, \mathbf{Pr}[f(y) = 2]} - \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t] \right| = 0 \, .$$

Taking $U = \mathbb{1}(f(x) = 2)$, we complete the proof of Proposition 10.6.3. □

### 10.6.2 Proof of Lemma 10.6.4

To prove Lemma 10.6.4, we first show that there is a function that satisfies our properties up to an error and has exponential pieces.

**Lemma 10.6.5.** *Fix $k \in Z_+$ and $\delta \gtrsim k$ and $\eta \in (0, 1)$. There exists a $(1/\eta + 1)^{\mathrm{poly}(k)}$-piecewise-constant function $f_\eta : \mathbb{R} \to \{0, 2\}$ such that $f(x) = 0$ for $|x| \geq \delta$ and for all $t \leq k$ it holds that*

$$\left| \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t f(y)] - \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t] \right| \leq \eta \, .$$

*Proof.* We first show that there exists a continue function that satisfies are properties.

**Claim 10.6.6.** *Fix $k \in Z_+$ and $\delta \gtrsim k$. There exists an $f : \mathbb{R} \to [0,2]$ such that $f(x) = 0$ for $|x| \geq \delta$ and for all $t \leq k$ it holds that*

$$\mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)}[y^t f(y)] = \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)}[y^t] \,.$$

We fix the following parameters throughout the proof (where $C$ denotes a sufficiently large absolute constant) $s = (\eta/(\delta k))^{\mathrm{poly}(k)}$ and we choose $s$ so that there exists $i' \in \mathbb{Z}$ so that $-\delta + i's = 0$. We partition the subset $[-\delta, \delta]$ in pieces $[is, (i+1)s)$ for $i \in \mathbb{Z}$. We define $\tilde{f}$ to be the following random piecewise-constant function: For each $i \in \{-1/\delta, \ldots, 1/\delta\}$ we let $\tilde{f}(x)$ be constant in the interval $x \in [is, (i+1)s)$, taking the following value:

$$\tilde{f}(x) = \begin{cases} 2, & \text{with probability } p_i := \int_{is}^{(i+1)s} \phi(x) f(x) \mathrm{d}x / (2 \int_{is}^{(i+1)s} \phi(x) \mathrm{d}x) \\ 0, & \text{with probability } 1 - p_i \end{cases} \tag{10.7}$$

and we define $\tilde{f}(x) = 0$ with probability 1 in the entire $(-\infty, -\delta] \cup [\delta, +\infty)$. Our goal is to show that $\left| \mathbf{E}_{y \sim \mathcal{N}(0,1)}[y^t \tilde{f}(y)] - \mathbf{E}_{y \sim \mathcal{N}(0,1)}[y^t] \right| \leq \eta$.

First note that for any $i \in [-\delta, -\delta + 1, \ldots, \delta]$ and $t \leq k$, we have on expectation that

$$\mathop{\mathbf{E}}_{\tilde{f}} \left[ \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)}[y^t \tilde{f}(y) \mathbb{1}(y \in (is, (i+1)s))] \right] = \int_{is}^{(i+1)s} \phi(y) f(y) \mathrm{d}y \frac{\int_{is}^{(i+1)s} y^t \phi(y) \mathrm{d}y}{\int_{is}^{(i+1)s} \phi(y) \mathrm{d}y} \,. \tag{10.8}$$

Because we have choose $s$ so that no interval $(is, (i+1)s)$ contains 0, that means that $y^t$ has constant sign inside the integration. Therefore, from mean valued theorem for integrals, we have that that there exists $\xi_1, \xi_2 \in (is, (i+1)s)$ so that the following are

true:

$$\int_{is}^{(i+1)s} y^t \phi(y) \mathrm{d}y = \xi_1^t \int_{is}^{(i+1)s} \phi(y) \mathrm{d}y \quad \text{and} \quad \int_{is}^{(i+1)s} \phi(y) f(y) y^t \mathrm{d}y = \xi_2^t \int_{is}^{(i+1)s} \phi(y) f(y) \mathrm{d}y \ .$$

Combining above with Equation (10.8), we have that

$$\begin{aligned} &\left| \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t f(y) \{ y \in (is, (i+1)s) \}] - \mathop{\mathbf{E}}_{\tilde{f}} \left[ \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t \tilde{f}(y) \{ y \in (is, (i+1)s) \}] \right] \right| \\ &= |\xi_1^t - \xi_2^t| \int_{is}^{(i+1)s} \phi(y) f(y) \mathrm{d}y \leq 2s |\xi_1^t - \xi_2^t| \leq 2 s^2 t \delta^{t-1} \ , \end{aligned}$$

where in the first inequality we used that $|f(y)\phi(y)| \leq 2$ and in the second we use that $|\xi_1^t - \xi_2^t| = \int_{\xi_1}^{\xi_2} t x^{t-1} \mathrm{d}x$ and $|\xi_1|, |\xi_2| \leq \delta$. We start using a probabilistic argument, by taking the expectation and using Fubini's theorem, we have that

$$\begin{aligned} &\left| \mathop{\mathbf{E}}_{\tilde{f}} \left[ \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t \tilde{f}(y) - y^t] \right] \right| \\ &= \left| \mathop{\mathbf{E}}_{\tilde{f}} \left[ \sum_{i=-\delta}^{\delta-1} \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [(y^t \tilde{f}(y) - y^t) \{ y \in (is, (i+1)s) \}] \right] - \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t \{ y \notin (-\delta, \delta) \}] \right| \\ &\leq \left| \mathop{\mathbf{E}}_{y \sim \mathcal{N}(0,1)} [y^t f(y) - y^t] \right| + \sum_{i=-\delta}^{\delta-1} 2 s^2 t \delta^{t-1} = 4 s^2 t \delta^t = O(\eta) \ , \end{aligned}$$

where we used that by construction $\mathbf{E}_{y \sim \mathcal{N}(0,1)} [y^t f(y) - y^t] = 0$.

□

To complete the proof note that from Lemma 10.6.8, we have that there exists $f : \mathbb{R} \mapsto \{0, 2\}$ that satisfies the properties of *Lemma* 10.6.4.

### 10.6.3 Proof of Claim 10.6.6

*Proof of Claim 10.6.6.* Let

$$f(x) = \left(1 + \frac{p(x)\mathbb{1}(x \in [-1,1])}{\phi(x)}\right)\mathbb{1}(|x| \le \delta) \ ,$$

for some appropriate chosen $k$-degree polynomial $p$. We have that

$$\mathop{\mathbf{E}}_{y\sim\mathcal{N}(0,1)}[y^t f(y)] = \mathop{\mathbf{E}}_{y\sim\mathcal{N}(0,1)}[y^t\mathbb{1}(|y| \le \delta)] + \int_{-1}^{1} p(y)y^t \mathrm{d}y \ .$$

We want to chose $p$ so that $\mathbf{E}_{y\sim\mathcal{N}(0,1)}[y^t f(y)] = \mathbf{E}_{y\sim\mathcal{N}(0,1)}[y^t]$, therefore, we want

$$\int_{-1}^{1} p(y)y^t \mathrm{d}y = \mathop{\mathbf{E}}_{y\sim\mathcal{N}(0,1)}[y^t\mathbb{1}(|y| \ge \delta)] \ .$$

Equivalently because Legendre polynomials form a complete basis, it suffices to find a polynomial $p$ so that

$$\int_{-1}^{1} p(y)P_t(y) \mathrm{d}y = \mathop{\mathbf{E}}_{y\sim\mathcal{N}(0,1)}[P_t(y)\mathbb{1}(|y| \ge \delta)] \ ,$$

for all $t \le k$. Therefore, we can write $p(y) = \sum_{i=0}^{k} a_i P_i(y)$ and $a_i = \frac{2i+1}{2}\int_{-1}^{1} P_i(y)p(y)\mathrm{d}y$. Hence, we have that

$$a_i = \frac{2}{2i+1}\mathop{\mathbf{E}}_{y\sim\mathcal{N}(0,1)}[P_t(y)\mathbb{1}(|y| \ge \delta)] \ .$$

Because for the Legendre polynomials it holds that $|P_t(y)| \le 1$ for all $|y| \le 1$ in order to have $|p(x)| \le \phi(1)$ it suffices to show that $\sum_{i=0}^{k} |a_i| \le \phi(1)$. Using the property that

$P_k(x) = 2^{-k}\sum_{i=1}^{\lceil k/2\rceil}\binom{k}{i}\binom{2k-2i}{k}x^{k-2i}$, we have that

$$a_t = \frac{2}{2t+1}\mathop{\mathbf{E}}_{y\sim\mathcal{N}(0,1)}[P_t(y)\mathbb{1}(|y|\geq\delta)] = \frac{2}{2t+1}2^{-t}\sum_{i=1}^{\lceil t/2\rceil}\binom{t}{i}\binom{2t-2i}{k}\mathop{\mathbf{E}}_{y\sim\mathcal{N}(0,1)}[y^{t-2i}\mathbb{1}(|y|\geq\delta)]$$
$$\lesssim \frac{\exp(-\delta^2/2)}{2t+1}2^{-t}\sum_{i=1}^{\lceil t/2\rceil}\binom{t}{i}\binom{2t-2i}{k}\delta^{t-2i} \lesssim \frac{\exp(-\delta^2/2)}{2t+1}P_t(\delta)\ .$$

Furthermore, using the Corollary 5.4 in [DKS17], we have that $|P_t(\delta)| \leq O(\delta^t)$ . Hence, we have that $|a_i| \lesssim \frac{\exp(-\delta^2/2)\delta^k}{2i+1}$. Choosing $\delta \gtrsim k$, we have that $\sum_{i=0}^{k}|a_i| \leq \phi(1)$ and that completes the proof.

□

### 10.6.4 Reducing the Number of Intervals

The proof is very similar to the proof of [DKZ20] but with the extra condition that we bound the maximum breakpoint of the piecewise constant function. For completeness, we state and add the extra details of the statements.

**Proposition 10.6.7.** *Let $k, \ell$ be positive integers with $\ell \geq k+3$ and $a, b \in \mathbb{R}$ with $b > a$. Let $D$ be a continuous distribution over $\mathbb{R}$ and let $\nu_0, \ldots, \nu_{k-1} \in \mathbb{R}$. Fix $b > 0$. If for any $\eta > 0$ there exists an at most $\ell$-piecewise constant function $g_\eta : \mathbb{R} \to \{a, b\}$ with $\{b_1^\eta, b_2^\eta, \ldots, b_{\ell-1}^\eta\}$ breakpoints such that $|\mathbf{E}_{z\sim D}[g_\eta(z)z^t] - \nu_t| \leq \eta$ for every non-negative integer $t < k$ and $|b_1^\eta|, |b_{\ell-1}^\eta| \leq b$, then there exists an at most $(k+3)$-piecewise constant function $f : \mathbb{R} \to \{a, b\}$ with $\{b_1, b_2, \ldots, b_{k+2}\}$ breakpoints such that $\mathbf{E}_{z\sim D}[f(z)z^t] = \nu_t$, for every non-negative integer $t < k$ and $|b_1|, |b_{k+2}| \leq b$.*

*Proof.* Note that, we can always transform the function $g_\eta : \mathbb{R} \mapsto \{a, b\}$ to a $g'_\eta : \mathbb{R} \mapsto \{\pm 1\}$ that satisfies similar properties. We define $g'_\eta(z) \stackrel{\text{def}}{=} (2g_\eta(z) - a - b)/(b - a)$

and let $\nu'_t = 2\nu_t/(b-a) + (a+b)/(b-a)\,\mathbf{E}_{z\sim D}[z^t]$ and $\eta' = \eta(2/(b-a))$. Hence, we have that for any $\eta' > 0$, there exists an at most $\ell$-piecewise constant function $g'_{\eta'} : \mathbb{R} \to \{\pm 1\}$ such that $|\,\mathbf{E}_{z\sim D}[g'_\eta(z)z^t] - \nu'_t| \leq \eta'$ for every non-negative integer $t < k$. By applying Lemma 10.6.9 and Lemma 10.6.8, we obtain that there exists an at most $(k+1)$-piecewise constant function $f' : \mathbb{R} \to \{\pm 1\}$ such that $\mathbf{E}_{z\sim D}[f'(z)z^t] = \nu'_t$ which keeps the first and last breakpoints the same as the original function, for every non-negative integer $t < k$. By setting $f(z) = (f'(z)(b-a) + a + b)/2$, we complete the proof of Proposition 10.6.7. □

**Lemma 10.6.8.** *Let $k$ be a positive integer and $b > 0$. Let $D$ be a continuous distribution over $\mathbb{R}$ and let $\nu_0, \ldots, \nu_{k-1} \in \mathbb{R}$. If for any $\eta > 0$ there exists an at most $(k+3)$-piecewise constant function $g_\eta : \mathbb{R} \to \{\pm 1\}$ with $\{b_1^\eta, b_2^\eta, \ldots, b_{k+2}^\eta\}$ breakpoints such that $|\,\mathbf{E}_{z\sim D}[g_\eta(z)z^t] - \nu_t| \leq \eta$ and $|b_1^\eta|, |b_{k+2}^\eta| \leq b$, for every non-negative integer $t < k$, then there exists an at most $(k+3)$-piecewise constant function $f : \mathbb{R} \to \{\pm 1\}$ with $\{b_1, b_2, \ldots, b_{k+2}\}$ breakpoints such that $\mathbf{E}_{z\sim D}[f(z)z^t] = \nu_t$, for every non-negative integer $t < k$ and $|b_1|, |b_{k+2}| \leq b$.*

Lemma 10.6.8 follows from the above using a compactness argument.

*Proof.* Let $p(z)$ be the pdf of $D$. For every $\eta > 0$, we have that there exists a function $g_\eta$ such that $|\,\mathbf{E}_{z\sim D}[f_\eta(z)z^t] - \nu_t| \leq \eta$, for every non-negative integer $t < k$ and the function $g_\eta$ is at most $(k+1)$-piecewise constant. Let $\mathbf{M} : \overline{\mathbb{R}}^k \mapsto \mathbb{R}^k$, where $M_i(\mathbf{b}) = \sum_{n=0}^{k}(-1)^{n+1}\int_{b_n}^{b_{n+1}} z^i p(z)\mathrm{d}z$ and $b_1 \leq b_2 \leq \ldots \leq b_k$, $b_0 = -\infty$ and $b_{k+1} = \infty$. Here we assume without loss of generality that before the first breakpoint the function is negative because we can always set the first breakpoint to be $-\infty$. It is clear that the function $\mathbf{M}$ is a continuous map and $\overline{\mathbb{R}}^{k+1}$ is a compact set, thus $\mathbf{M}\left(\overline{\mathbb{R}}^{k+1}\right)$ is a compact set. We also have that for every $\eta > 0$ there is a point $\mathbf{b} \in \overline{\mathbb{R}}^{k+1}$ such that

$|\mathbf{M}(\mathbf{b}) \cdot \mathbf{e}_i - \nu_i| \leq \eta$ for all $i < k$. Thus, from compactness, we have that there exists a point $\mathbf{b}^* \in \overline{\mathbb{R}}^{k+1}$ such that $\mathbf{M}(\mathbf{b}^*) = \mathbf{0}$. This completes the proof. □

The following lemma is analogous to Lemma 3.8 in [DKZ20]. We include its proof for completeness, as our work considers more general distributions and requires specifying exact moment values.

**Lemma 10.6.9.** *Let $m$ and $k$ be positive integers such that $m > k + 3$ and $\eta > 0$. Let $D$ be a continuous distribution over $\mathbb{R}$ and let $\nu_0, \ldots, \nu_{k-1} \in \mathbb{R}$. If there exists an $m$-piecewise constant $f : \mathbb{R} \mapsto \{\pm 1\}$ such that $|\mathbf{E}_{z\sim D}[f(z)z^t] - \nu_t| < \eta$ for all non-negative integers $t < k$, then there exists an at most $(m-1)$-piecewise constant $g : \mathbb{R} \mapsto \{\pm 1\}$ such that $|\mathbf{E}_{z\sim D}[g(z)z^t] - \nu_t| < \eta$ for all non-negative integers $t < k$ and if $\{b_1, b_2, \ldots, b_{m-1}\}$ are the breakpoints of $f$ and $\{b'_1, b'_2, \ldots, b'_{m-2}\}$ are the breakpoints of $g$, it holds that $b'_1 = b_1$ and $b'_{m-2} = b_{m-1}$.*

*Proof.* Let $p(z)$ be the pdf of $D$. Let $\{b_1, b_2, \ldots, b_{m-1}\}$ be the breakpoints of $f$, i.e., the points where the function $f$ changes value. Then let $F(z_1, z_2, \ldots, z_{m-1}, z) : \overline{\mathbb{R}}^m \mapsto \mathbb{R}$ be an $m$-piecewise constant function with breakpoints on $z_1, \ldots, z_{m-1}$, where $z_1 < z_2 < \ldots < z_{m-1}$ and $F(b_1, b_2, \ldots, b_{m-1}, z) = f(z)$. For simplicity, let $\mathbf{z} = (z_1, \ldots, z_{m-1})$ and define $M_i(\mathbf{z}) = \mathbf{E}_{z\sim D}[F(\mathbf{z}, z)z^i]$ and let $\mathbf{M}(\mathbf{z}) = [M_0(\mathbf{z}), M_1(\mathbf{z}), \ldots M_{k-1}(\mathbf{z})]^T$. It is clear from the definition that $M_i(\mathbf{z}) = \sum_{n=0}^{m-1} \int_{z_n}^{z_{n+1}} F(\mathbf{z}, z)z^i p(z)\mathrm{d}z = \sum_{n=0}^{m-1} a_n \int_{z_n}^{z_{n+1}} z^i p(z)\mathrm{d}z$, where $z_0 = -\infty$ and $z_m = \infty$ and $a_n$ is the sign of $F(\mathbf{z}, z)$ in the interval $(z_n, z_{n+1})$. Note that $a_n = -a_{n+1}$ for every $0 \leq n < m$. By taking the derivative of $M_i$ in $z_j$, for $0 < j < m$, we get that

$$\frac{\partial}{\partial z_j} M_i(\mathbf{z}) = 2a_{j-1} z_j^i p(z_j) \quad \text{and} \quad \frac{\partial}{\partial z_j} \mathbf{M}(\mathbf{z}) = 2a_{j-1} p(z_j)[1, z_j^1, \ldots, z_j^{k-1}]^T \,.$$

We now argue that for any $\mathbf{z}$ with distinct coordinates that there exists a vector $\mathbf{u} \in \mathbb{R}^{m-1}$ such that $\mathbf{u} = (0, \mathbf{u}_2, \ldots, \mathbf{u}_{k+1}, 0, 0, \ldots, 1, 0)$ and the directional derivative of $\mathbf{M}$ in the $\mathbf{u}$ direction is zero. To prove this, we construct a system of linear equations such that $\nabla_{\mathbf{u}} M_i(\mathbf{z}) = 0$, for all $0 < i < k$. Indeed, we have $\sum_{j=2}^{k} \frac{\partial}{\partial z_j} M_i(\mathbf{z}) \mathbf{u}_j = -\frac{\partial}{\partial z_{m-2}} M_i(\mathbf{z})$ or $\sum_{j=2}^{k} a_{j-1} z_j^i p(z_j) \mathbf{u}_j = -a_{m-3} z_{m-2}^i p(z_{m-2})$, which is linear in the variables $\mathbf{u}_j$. Let $\widehat{\mathbf{u}}$ be the vector with the first $k$ variables and let $\mathbf{w}$ be the vector of the right hand side of the system, i.e., $\mathbf{w}_i = -a_{m-3} z_{m-2}^i p(z_{m-2})$. Then this system can be written in matrix form as $\mathbf{VD}\widehat{\mathbf{u}} = \mathbf{w}$, where $\mathbf{V}$ is the Vandermonde matrix, i.e., the matrix that is $\mathbf{V}_{i,j} = \alpha_i^{j-1}$, for some values $\alpha_i$ and $\mathbf{D}$ is a diagonal matrix. In our case, $\mathbf{V}_{i,j} = z_i^{j-1}$ and $\mathbf{D}_{j,j} = 2a_{j-1} p(z_j)$. It is known that the Vandermonde matrix has full rank iff for all $i \neq j$ we have $\alpha_i \neq \alpha_j$, which holds in our setting. Thus, the matrix $\mathbf{VD}$ is nonsingular and there exists a solution to the equation. Thus, there exists a vector $\mathbf{u}$ with our desired properties and, moreover, any vector in this direction is a solution of this system of linear equations. Note that the vector $\mathbf{u}$ depends on the value of $\mathbf{z}$, thus we consider $\mathbf{u}(\mathbf{z})$ be the (continuous) function that returns a vector $\mathbf{u}$ given $\mathbf{z}$.

We define a differential equation for the function $\mathbf{v} : \overline{\mathbb{R}} \mapsto \overline{\mathbb{R}}^{m-1}$, as follows: $\mathbf{v}(0) = \mathbf{b}$, where $\mathbf{b} = (b_1, \ldots, b_{m-1})$, and $\mathbf{v}'(T) = \mathbf{u}(\mathbf{v}(T))$ for all $T \in \overline{\mathbb{R}}$. If $\mathbf{v}$ is a solution to this differential equation, then we have:

$$\frac{\mathrm{d}}{\mathrm{d}T}\mathbf{M}(\mathbf{v}(T)) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{v}(T)}\mathbf{M}(\mathbf{v}(T))\frac{\mathrm{d}}{\mathrm{d}T}\mathbf{v}(T) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{v}(T)}\mathbf{M}(\mathbf{v}(T))\mathbf{u}(\mathbf{v}(T)) = \mathbf{0}\ ,$$

where we used the chain rule and that the directional derivative in $\mathbf{u}(\mathbf{v}(T))$ direction is zero. This means that the function $\mathbf{M}(\mathbf{v}(t))$ is constant, and for all $0 \leq j < k$, we have $|M_j - \nu_j| < \eta$, because we have that $|\,\mathbf{E}_{z \sim D}[F(z_1, \ldots, z_{m-1}, z) z^t] - \nu_t| < \eta$. Furthermore, since $\mathbf{u}(\mathbf{v}(T))$ is continuous in $\mathbf{v}(T)$, this differential equation will be well founded and

have a solution up until the point where either two of the $z_i$ approach each other or one of the $z_i$ approaches plus or minus infinity (the solution cannot oscillate, since $\mathbf{v}'_{m-2}(T) = 1$ for all $T$).

Running the differential equation until we reach such a limit, we find a limiting value $\mathbf{v}^*$ of $\mathbf{v}(T)$ so that: There is an $i$ such that $\mathbf{v}^*_i = \mathbf{v}^*_{i+1}$, which gives us a function that is at most $(m-2)$-piecewise constant, i.e., taking $F(\mathbf{v}^*, z)$. Which is guaranteed to exist since $\mathbf{v}'_{m-2}(T) = 1$ for all $T$, thus there exists $T'$ such that $\mathbf{v}^*_{m-2} = \mathbf{v}^*_{m-1}$. Thus, we have a function with at most $m-1$ breakpoints and the same moments. Furthermore, because $\mathbf{u}_0 = \mathbf{u}_m = 0$ the differential equation keeps the first and the last breakpoint constant we have that the first and last breakpoints of the new function are the same as in the original function. This completes the proof. □

# Part IV

# Statistical-Computational Tradeoffs via SQ Complexity

# Chapter 11

# SQ-Lower Bounds using Linear Programming

## 11.1 Introduction

This chapter investigates the fundamental computational limits of agnostically learning both Boolean and real-valued functions under the Gaussian distribution. We formally answer the long-standing question of whether polynomial regression algorithms are optimal for this task by developing a general framework to derive strong lower bounds in the Statistical Query (SQ) model. Our central contribution is a novel methodology based on *linear programming duality* in function spaces, which provides a direct and powerful connection between the analytical difficulty of approximating a concept class with polynomials and its inherent computational hardness.

**Contributions for Boolean-valued Concepts** Our main general result for Boolean-valued concepts is the following:

**Theorem 11.1.1** (Generic SQ Lower Bound, Boolean Case)**.** *Let $n, m \in \mathbb{Z}_+$ with $m \leq n^a$ for any constant $0 < a < 1/2$ and $\epsilon \geq n^{-c}$ for some sufficiently small constant $c > 0$. Fix a function $f : \mathbb{R}^m \to \{\pm 1\}$. Let $d$ be the smallest integer such that there exists a degree at most $d$ polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}_m}[|p(\mathbf{x}) - f(\mathbf{x})|] < 2\epsilon$. Let $\mathcal{C}$ be a class of Boolean-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any $\mathbf{P} \in \mathbb{R}^{m \times n}$ such that $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Any SQ algorithm that agnostically learns $\mathcal{C}$ under $\mathcal{N}_n$ to error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

The $L^1$-polynomial regression algorithm and Theorem 11.1.1 characterize the complexity of agnostic learning under the Gaussian distribution – within the class of SQ algorithms – for a range of concept classes. If $d$ is the (minimum) degree for which any function in $\mathcal{C}$ can be $\epsilon$-approximated by a degree-$d$ polynomial in $L^1$-norm, the complexity of agnostically learning $\mathcal{C}$ is, roughly, $n^{\Theta(d)}$.

***Applications of Theorem 11.1.1.*** Note that the above result does not tell us what the optimal degree $d$ is for any given concept class $\mathcal{C}$. Using analytic techniques, we establish explicit lower bounds on the $L^1$ polynomial approximation degree for three fundamental concept classes: Linear Threshold Functions (LTFs), Polynomial Threshold Functions, and Intersections of Halfspaces. As a corollary, we obtain explicit SQ lower bounds for these classes. Our applications are summarized in Table 11.1.

For the class of LTFs, using a known degree lower bound for the sign function [Gan02], we immediately obtain an SQ lower bound of $n^{\Omega(1/\epsilon^2)}$. This bound is optimal (within polynomial factors), improving on the previous SQ lower bound of $n^{\Omega(1/\epsilon)}$ [GGK20, DKZ20]. Our approach is simpler and more general compared to these prior works, immediately extending to other families. For the broader class of degree-$k$ PTFs, we

| **Concept Class** | Lower Bound | Upper Bound |
|---|---|---|
| LTFs | $\Omega\left(1/\epsilon^2\right)$ [Gan02] | $O\left(1/\epsilon^2\right)$ [Gan02, DKN10] |
| Degree-$k$ PTFs | $\Omega\left(k^2/\epsilon^2\right)$ (Thm 11.3.1) | $O\left(k^2/\epsilon^4\right)$ [Kan11] |
| Intersections of $k$ Halfspaces | $\tilde{\Omega}\left(\sqrt{\log k}/\epsilon\right)$ (Thm 11.3.5) | $O\left(\log k/\epsilon^4\right)$ [KOS08] |

Table 11.1: Bounds on the degree $d$ of $\epsilon$-approximating polynomials in $L^1$-error under the Gaussian measure. For each concept class, we obtain an SQ lower bound of $n^{\Omega(d)}$.

establish a degree lower bound of $\Omega(k^2/\epsilon^2)$ (Proposition 11.3.2), which yields an SQ lower bound of $n^{\Omega(k^2/\epsilon^2)}$ for the agnostic learning problem.

Our third explicit degree lower bound is for intersections of $k$ halfspaces. For this concept class, we prove a degree lower bound of $d = \tilde{\Omega}(\sqrt{\log k}/\epsilon)$ which implies a corresponding SQ lower bound of $n^{\tilde{\Omega}(\sqrt{\log k}/\epsilon)}$. In the process, we establish a new structural result translating lower bounds on the Gaussian Noise Sensitivity (GNS) of *any* Boolean function to the $L^1$-polynomial approximation degree of the same function.

Recall that the Gaussian Noise Sensitivity (GNS) of a function $f : \mathbb{R}^n \to \{\pm 1\}$ is defined as $\mathrm{GNS}_\rho(f) \stackrel{\text{def}}{=} \mathbf{Pr}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^\rho}[f(\mathbf{x}) \neq f(\mathbf{y})]$, where $\mathcal{N}_n^\rho$ is the distribution of a $(1-\rho)$-correlated Gaussian pair (i.e., $\mathbf{x}$ and $\mathbf{y}$ are standard Gaussians with correlation $(1-\rho)$). We show the following:

**Theorem 11.1.2** (Structural Result)**.** *Let $f : \mathbb{R}^n \to \{\pm 1\}$ and $p : \mathbb{R}^n \to \mathbb{R}$ be a degree at most $d$ polynomial. Then, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x})-p(\mathbf{x})|] \geq \Omega(1/\log(d))\mathrm{GNS}_{(\log(d)/d)^2}(f)$. Furthermore, for any $\epsilon > 0$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - p(\mathbf{x})|] \geq \mathrm{GNS}_\epsilon(f)/4 - O(d\sqrt{\epsilon})$.*

**Contributions for Real-valued Concepts** For agnostically learning real-valued concepts, we provide two generic lower bound results, analogous to Theorem 11.1.1, for Correlational SQ (CSQ) algorithms and general SQ algorithms respectively. A conceptual message of our results is that $L^2$ regression is essentially optimal against

CSQ algorithms, but not necessarily optimal against general SQ algorithms.

Recall that Correlational SQ (CSQ) algorithms are a subclass of SQ algorithms, where the algorithm is allowed to choose any bounded query function on the examples and obtain estimates of its correlation with the labels. (See Section (1.4.5) for a detailed description.) This class of algorithms is fairly broad, capturing many learning algorithms used in practice (including gradient-descent). For CSQ algorithms, we prove.

**Theorem 11.1.3** (Generic CSQ Lower Bound, Real-valued Case)**.** *Let $n, m \in \mathbb{Z}_+$ with $m \leq n^a$ for any constant $0 < a < 1/2$ and $\epsilon \geq n^{-c}$ for some sufficiently small constant $c > 0$. Let $f : \mathbb{R}^m \to \mathbb{R}$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f^2(\mathbf{x})] = 1$ and $d$ be the smallest integer such that there exists a degree at most $d$ polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\|f - p\|_2 < \epsilon$. Let $\mathcal{C}$ be a class of real-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any matrix $\mathbf{P} \in \mathbb{R}^{m\times n}$ satisfying $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Then, any CSQ algorithm that agnostically learns $\mathcal{C}$ over $\mathcal{N}_n$ to $L^2$-error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

Our lower bound for the general SQ model is presented below. The difference between the two is that the latter uses the $L^1$-norm to measure the approximation of $f$ by polynomials.

**Theorem 11.1.4** (Generic SQ Lower Bound, Real-valued Case)**.** *Let $n, m \in \mathbb{Z}_+$ with $m \leq n^a$ for any constant $0 < a < 1/2$ and $\epsilon \geq n^{-c}$ for some sufficiently small constant $c > 0$. Let $f : \mathbb{R}^m \to \mathbb{R}$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f^2(\mathbf{x})] = 1$ and $d$ be the smallest integer such that there exists a degree at most $d$ polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\|f - p\|_1 < \epsilon$. Let $\mathcal{C}$ be a class of real-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any matrix $\mathbf{P} \in \mathbb{R}^{m\times n}$ satisfying $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Then, any SQ algorithm*

*that agnostically learns $\mathcal{C}$ over $\mathcal{N}_n$ to $L^2$-error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

***Applications of Theorems 11.1.3 and 11.1.4.*** As in the Boolean-valued setting, obtaining explicit (C)SQ lower bounds for agnostically learning real-valued concepts requires analytic tools to establish lower bounds on the degree of polynomial approximations. In this work, we give such lower bounds for two fundamental concept classes: ReLUs and sigmoids. Establishing degree lower bounds for other non-linear activations is left as a question for future work.

| | $p = 1$ | $p = 2$ |
|---|---|---|
| **Concept Class** | **Lower / Upper Bounds** | **Lower / Upper Bounds** |
| ReLUs | $\Omega(1/\epsilon)$ (Cor. 11.5.3)<br>$O(1/\epsilon)$ | $\Omega\left(1/\epsilon^{4/3}\right)$ (Cor. 11.5.3)<br>$O\left(1/\epsilon^{4/3}\right)$ |
| Sigmoids | $\Omega(\log(1/\epsilon))$ (Thm. 11.5.8)<br>$O(\log^2(1/\epsilon))$ | $\Omega\left(\log^2(1/\epsilon)\right)$ (Cor. 11.5.5)<br>$O\left(\log^2(1/\epsilon)\right)$ |

Table 11.2: Bounds on the degree $d$ of $\epsilon$-approximating polynomials in $L^1$ and $L^2$-error under the Gaussian measure. For each concept class, we obtain a CSQ (resp. SQ) lower bound of $n^{\Omega(d)}$, where $d$ is the $L^2$ degree (resp. $L^1$ degree).

Our degree lower bounds applications for both $L^1$ and $L^2$ polynomial approximations are summarized in Table 11.2. Combining these degree lower bounds Theorems 11.1.3 and 11.1.4 implies explicit SQ lower bounds for ReLUs and sigmoids.

Concretely, for agnostically learning ReLUs, we establish a CSQ lower bound of $n^{\Omega(1/\epsilon^{4/3})}$ (matching the $n^{O(1/\epsilon^{4/3})}$ upper bound obtained via $L^2$-regression); and an SQ lower bound of $n^{\Omega(1/\epsilon)}$, improving on the previous best bound of $n^{\Omega((1/\epsilon)^{1/36})}$ [GGK20, DKZ20].

### 11.1.1 Comparison to Prior Work

At the level of results, the most relevant prior works are the two independent works [DKZ20, GGK20], which established the previously best SQ lower bounds for LTFs, ReLUs, and sigmoids under the Gaussian distribution. We have already provided a technical comparison to [DKZ20] in the previous subsection. The work [GGK20] relies on a boosting procedure that translates recent SQ lower bounds for (non-agnostic) learning one-hidden-layer neural networks [DKKZ20] to agnostically learning simple concept classes.

A useful point of technical comparison is the work [DFT+15], which gave an analogue of our results on agnostically learning Boolean functions on the Boolean hypercube. The basic statement is the same — that the complexity of agnostic learning Boolean functions under a discrete product distribution is characterized by the $L^1$-approximation degree — and the duality-based proof techniques are similar. In particular, [DFT+15] sets up a *finite* LP to find a function $f$ that has vanishing Fourier coefficients but is close in $L^1$-norm to the target function. Due to the discrete nature of the setting they consider, [DFT+15] avoids the functional analysis based arguments required to establish duality in our setting.

A more significant difference with our framework is that the hard family of [DFT+15] embeds a copy of $f$ as a junta on a random subset of coordinates, while ours embeds it in a random low-dimensional subspace. This is a critical distinction and is necessary in the Gaussian setting to obtain our tight characterization and the associated applications to LTFs/PTFs and intersections of halfspaces. Finally, we remark that the appendix of [DFT+15] sketches a generalization of their results to arbitrary product distributions (including the Gaussian distribution). We emphasize, however, that the lower bound

obtained from their construction does *not* match the guarantee of the $L^1$-regression algorithm [KKMS08] for the following reason: The exponent for their lower bounds for the continuous setting have to do with the degree necessary to $\epsilon$-approximate the hard function as *a linear combination of d-juntas*. On the other hand, the upper bound of [KKMS08] is related to the approximation by degree-$d$ polynomials. Note that degree-$d$ polynomials are always linear combinations of $d$-juntas, and thus the approximation degree by linear combinations of juntas is lower than the approximation degree by polynomials. In summary, while the lower bound of [DFT+15] is tight for discrete product distributions, this is not true in general.

## Technical Preliminaries

**Notation** We will use $\mathcal{U}(S)$ for the uniform distribution on the set $S$. Let $\mathcal{N}$ denote the standard univariate Gaussian distribution and $\mathcal{N}_n$ denote the standard $n$-dimensional Gaussian distribution. We use $\phi_n$ to denote the pdf of $\mathcal{N}_n$. Sometimes we may use the same symbol for a distribution and its pdf, i.e., denote by $D(\mathbf{x})$ the density that the distribution $D$ gives to the point $\mathbf{x}$.

We use $\langle \mathbf{u}, \mathbf{v} \rangle$ for the inner product of vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. For a matrix $\mathbf{P} \in \mathbb{R}^{m \times n}$, let $\|\mathbf{P}\|_2$ denote its spectral norm and $\|\mathbf{P}\|_F$ denote its Frobenius norm. We use $\mathbf{I}_n$ to denote the $n \times n$ identity matrix. We denote by $\mathcal{P}_d^n$ the class of all polynomials from $\mathbb{R}^n$ to $\mathbb{R}$ with degree at most $d$.

## 11.2 SQ Lower Bound for Boolean-Valued Concepts: Proof of Theorem 11.1.1

The idea of our construction is to find a function $g : \mathbb{R}^m \to [-1, 1]$ whose low-degree moments vanish and is non-trivially close to $f$. Our hard distribution will then embed $g$ in a random $m$-dimensional subspace. Given this construction, we can apply Lemma 1.4.9 to prove Theorem 11.1.1. The following result establishes the existence of such a function $g$.

**Proposition 11.2.1.** *Let $f : \mathbb{R}^m \to \{\pm 1\}$ be such that for any polynomial $p : \mathbb{R}^m \to \mathbb{R}$ of degree at most $d-1$, it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[|p(\mathbf{x})-f(\mathbf{x})|] \geq 2\epsilon$. There exists a function $g\colon \mathbb{R}^m \to [-1, 1]$ such that:*

1. *For any degree at most $d-1$ polynomial $P : \mathbb{R}^m \to \mathbb{R}$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0$, i.e., $g$ has zero low-degree moments, and,*
2. *$\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[|g(\mathbf{x}) - f(\mathbf{x})|] \leq 1 - 2\epsilon$, i.e., $g$ is non-trivially close to $f$.*

*Proof.* Note that such a function $g$ would be a solution to the following infinite linear program (LP):

$$(*) \begin{cases} \mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[|g(\mathbf{x}) - f(\mathbf{x})|] \leq 1 - 2\epsilon & \\ \mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0 & \forall P \in \mathcal{P}^m_{d-1} \\ |g(\mathbf{x})| \leq 1 & \forall \mathbf{x} \in \mathbb{R}^m \end{cases}$$

We claim that the LP $(*)$ is equivalent to the following LP:

$$(**) \begin{cases} -\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})f(\mathbf{x})] + 2\epsilon \leq 0 & \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0 & \forall P \in \mathcal{P}^m_{d-1} \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})h(\mathbf{x})] - \|h\|_1 \leq 0 & \forall h \in L^1(\mathbb{R}^m) \end{cases}$$

We now show the equivalence between the two formulations. We claim that the third constraint of $(*)$ is equivalent with the third constraint of $(**)$. This follows by introducing the "dual variable" $h : \mathbb{R}^m \to \mathbb{R}$. The forward direction follows from Hölder's inequality and the inverse follows from the definition of dual norms as suprema. Finally, for the first constraints, note that since $f$ is Boolean-valued and $\|g\|_\infty \leq 1$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[|g(\mathbf{x}) - f(\mathbf{x})|] = 1 - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})f(\mathbf{x})]$.

At this point, we would like to use "LP duality" to argue that $(**)$ is feasible if and only if its "dual LP" is infeasible. While such a statement turns out to be true, it requires some care to prove since we are dealing with infinite LPs (both in number of variables and constraints). The proof requires a version of the geometric Hahn-Banach theorem from functional analysis.

**Lemma 11.2.2** (Informal)**.** *The LP defined by* $(**)$ *is feasible if only if there is no conical combination of the inequalities of* $(**)$ *that yields the contradictory inequality* $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x}) \cdot 0] + 1 \leq 0$.

A proof of this lemma can be found on Section 11.6. Using Lemma 11.2.2, the LP

defined by $(**)$ is feasible if and only if the following "dual" LP is infeasible:

$$(**') \begin{cases} \|h\|_1 - 2\lambda\,\epsilon < 0 & \\ h(\mathbf{x}) + P(\mathbf{x}) - \lambda\, f(\mathbf{x}) = 0 & \forall \mathbf{x} \in \mathbb{R}^m \\ \lambda \geq 0, h \in L^1(\mathbb{R}^m), P \in \mathcal{P}^m_{d-1} & \end{cases}$$

Suppose that such a solution $(\lambda, h, P)$ exists. We can assume that $\lambda > 0$, since otherwise the first inequality is violated. Moreover, by scaling the solution, we can further assume $\lambda = 1$. Then, the second constraint becomes $h = f - P$ and the first becomes $\|f - P\|_1 < 2\epsilon$. However, this cannot happen by the definition of the degree $d$ (since, by assumption, there is no polynomial of degree less than $d$ such that $\|f - P\|_1 < 2\epsilon$). Therefore, the LP $(**)$ is feasible, which completes our proof. □

Our construction will use rotated versions of the function $g$ from Proposition 11.2.1 to create a family of distributions that is hard to distinguish from a fixed reference distribution. To bound the SQ dimension of this hypothesis testing problem, we will need a generalization of Lemma 16 in [DKKZ20], which bounds the correlation of two rotated versions of $g$. To formally state our lemma, we will need one additional piece of terminology. If $g(\mathbf{x}) = \sum_{J \in \mathbb{N}^m} \hat{g}(J) H_J(\mathbf{x})$ is the Hermite expansion of $g$, the degree-$t$ Hermite part of $g$ is the sum of the terms corresponding to the Hermite polynomials of degree exactly $t$. (For background in multilinear algebra and Hermite analysis, see Section (1.4.4) and Section (1.4.3).) Our main correlation lemma is the following.

**Lemma 11.2.3** (Correlation Lemma)**.** *Let $g : \mathbb{R}^m \to \mathbb{R}$ and $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m \times n}$ be linear maps such that $\mathbf{U}\mathbf{U}^\intercal = \mathbf{V}\mathbf{V}^\intercal = \mathbf{I}_m$. Then, we have that*

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}_n} [g(\mathbf{U}\mathbf{x}) g(\mathbf{V}\mathbf{x})] \leq \sum_{t=0}^{\infty} \|\mathbf{U}\mathbf{V}^\intercal\|_2^t \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}_m} [(g^{[t]}(\mathbf{x}))^2] \,,$$

*where $g^{[t]}$ denotes the degree-$t$ Hermite part of $g$.*

*Proof.* To simplify notation, write $g_1(\mathbf{x}) = g(\mathbf{U}\mathbf{x})$ and $g_2(\mathbf{x}) = g(\mathbf{V}\mathbf{x})$. Moreover, we will write $g_1(\mathbf{x}) \sim \sum_{k=0}^{\infty} g_1^{[k]}(\mathbf{x})$ and $g_2(\mathbf{x}) \sim \sum_{k=0}^{\infty} g_2^{[k]}(\mathbf{x})$. Using Fact 1.4.1, we obtain

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[g_1(\mathbf{x})g_2(\mathbf{x})] &= \sum_{k=0}^{\infty} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[g_1^{[k]}(\mathbf{x})g_2^{[k]}(\mathbf{x})] = \sum_{k=0}^{\infty} \frac{1}{k!}\nabla^k g_1^{[k]}(\mathbf{x}) \cdot \nabla^k g_2^{[k]}(\mathbf{x}) \\ &= \sum_{k=0}^{\infty} \frac{1}{k!}\nabla^k g^{[k]}(\mathbf{U}\mathbf{x}) \cdot \nabla^k g^{[k]}(\mathbf{V}\mathbf{x})\,. \end{aligned} \tag{11.1}$$

Denote by $\mathcal{U} \subseteq \mathbb{R}^n$ the image of the linear map $\mathbf{U}^\intercal$. Now observe that, using the chain rule, for any function $h(\mathbf{U}\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ it holds $\nabla h(\mathbf{U}\mathbf{x}) = \partial_i h(\mathbf{U}\mathbf{x})\mathbf{U}_{ij} \in \mathcal{U}$, where we used Einstein's summation notation for repeated indices. Applying the above rule $k$ times, we have that

$$\nabla^k h(\mathbf{U}\mathbf{x}) = \partial_{i_k} \dots \partial_{i_1} h(\mathbf{U}\mathbf{x})\mathbf{U}_{i_1 j_1} \dots \mathbf{U}_{i_k j_k} \ \in \mathcal{U}^{\otimes k}\,.$$

We denote $\mathbf{R} = \nabla^k g^{[k]}(\mathbf{x})$ and observe that this tensor does not depend on $\mathbf{x}$. Moreover, denote $\mathbf{M} = \mathbf{U}\mathbf{V}^\intercal$, $\mathbf{S} = \nabla^k g^{[k]}(\mathbf{U}\mathbf{x}) = (\mathbf{U}^\intercal)^{\otimes k}\mathbf{R} \in \mathcal{U}^{\otimes k}$, and $\mathbf{T} = \nabla^k g^{[k]}(\mathbf{V}\mathbf{x}) = (\mathbf{V}^\intercal)^{\otimes k}\mathbf{R} \in \mathcal{V}^{\otimes k}$. We have that

$$\mathbf{S} \cdot \mathbf{T} = (\mathbf{U}^\intercal)^{\otimes k}\mathbf{R} \cdot (\mathbf{V}^\intercal)^{\otimes k}\mathbf{R} = \mathbf{R} \cdot \mathbf{M}^{\otimes k}\mathbf{R} \le \left\|\mathbf{M}^{\otimes k}\right\|_2 \|\mathbf{R}\|_2^2 = k!\,\|\mathbf{M}\|_2^k \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[(g^{[k]}(\mathbf{x}))^2]\,,$$

where to get the last equality we used again Fact 1.4.1. To finish the proof, we combine this inequality with Equation (11.1). □

We consider high-dimensional distributions that encode a function in a subspace and are Gaussian in the orthogonal complement. Using Lemma 11.2.3, we can bound

their pairwise correlations.

**Corollary 11.2.4.** *Let $d \geq 2$ and $D$ be a distribution over $\mathbb{R}^m$ such that the first $(d-1)$ moments of $D$ match the corresponding moments of $\mathcal{N}_m$. Let $G(\mathbf{x}) = D(\mathbf{x})/\phi_m(\mathbf{x})$ be the ratio of the corresponding probability density functions. For matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m \times n}$ such that $\mathbf{U}\mathbf{U}^\intercal = \mathbf{V}\mathbf{V}^\intercal = \mathbf{I}_m$, define $D_{\mathbf{U}}$ and $D_{\mathbf{V}}$ to have probability density functions $G(\mathbf{U}\mathbf{x})\phi_n(\mathbf{x})$ and $G(\mathbf{V}\mathbf{x})\phi_n(\mathbf{x})$, respectively. Then, we have that $|\chi_{\mathcal{N}_n}(D_{\mathbf{U}}, D_{\mathbf{V}})| \leq \|\mathbf{U}\mathbf{V}^\intercal\|_2^d \chi^2(D, \mathcal{N}_m)$.*

*Proof.* We compute

$$\begin{aligned}\chi_{\mathcal{N}_n}(D_{\mathbf{U}}, D_{\mathbf{V}}) &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}\left[\frac{(D_{\mathbf{U}}(\mathbf{x}) - \phi_n(\mathbf{x}))(D_{\mathbf{V}}(\mathbf{x}) - \phi_n(\mathbf{x}))}{\phi_n^2(\mathbf{x})}\right] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}\left[(G(\mathbf{U}\mathbf{x}) - 1)(G(\mathbf{V}\mathbf{x}) - 1)\right] .\end{aligned}$$

We then apply Lemma 11.2.3 to the function $g(\mathbf{x}) = G(\mathbf{x}) - 1$. Note that the assumption that $D$ matches the first $d-1$ moments with $\mathcal{N}_m$ is equivalent to saying that $g^{[t]} = 0$ for $t < d$. Thus, Lemma 11.2.3 implies that

$$\begin{aligned}|\chi_{\mathcal{N}_n}(D_{\mathbf{U}}, D_{\mathbf{V}})| &\leq \|\mathbf{U}\mathbf{V}^\intercal\|_2^d \sum_{t=0}^{\infty} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[(g^{[t]}(\mathbf{x}))^2] = \|\mathbf{U}\mathbf{V}^\intercal\|_2^d \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[g^2(\mathbf{x})] \\ &\leq \|\mathbf{U}\mathbf{V}^\intercal\|_2^d \chi^2(D, \mathcal{N}_m) ,\end{aligned}$$

where the equality is Parseval's identity and in the last inequality we used the definition of $G$. □

Note that $D_{\mathbf{U}}$ and $D_{\mathbf{V}}$ are copies of $D$ in the subspaces defined by $\mathbf{U}$ and $\mathbf{V}$ respectively, and independent Gaussians in the orthogonal component.

In order to create our hard family of distributions, we will need the following lemma which states that there exist exponentially many linear operators from $\mathbb{R}^n$ to $\mathbb{R}^m$ that are nearly orthogonal.

**Lemma 11.2.5.** *Let $0 < a, c < 1/2$ and $m, n \in \mathbb{Z}_+$ such that $m \leq n^a$. There exists a set $S$ of $2^{\Omega(n^c)}$ matrices in $\mathbb{R}^{m \times n}$ such that every $\mathbf{U} \in S$ satisfies $\mathbf{U}\mathbf{U}^\intercal = \mathbf{I}_m$ and every pair $\mathbf{U}, \mathbf{V} \in S$ with $\mathbf{U} \neq \mathbf{V}$ satisfies $\|\mathbf{U}\mathbf{V}^\intercal\|_F \leq O(n^{2c-1+2a})$.*

*Proof.* Our proof relies on the following fact that there exist exponentially many nearly orthogonal unit vectors.

**Fact 11.2.6** (see, e.g., Lemma 3.7 of [DKS17])**.** *For any $0 < c < 1/2$ there exists a set $S'$ of $2^{\Omega(n^c)}$ unit vectors in $\mathbb{R}^n$ such that any pair $\mathbf{u}, \mathbf{v} \in S'$, with $\mathbf{u} \neq \mathbf{v}$, satisfies $|\langle \mathbf{u}, \mathbf{v} \rangle| < O(n^{c-1/2})$.*

Let $S'$ be the set of unit vectors that Fact 11.2.6 constructs. We group them into sets of size $m$ and use the vectors of each group as rows for each matrix that we make. Thus, we create at least $|S'|/n^a = 2^{\Omega(n^c)}$ many matrices. Next, we ortho-normalize each matrix $\mathbf{V} \in S'$ using the Gram-Schmidt process, in order to get $\mathbf{V}\mathbf{V}^\intercal = \mathbf{I}_m$. In every row of $\mathbf{V}$, the Gram-Schmidt algorithm adds at most $m$ orthogonal vectors, each having norm $O(n^{c-1/2})$. Thus, the total correction term for each row has norm at most $\sqrt{m}O(n^{c-1/2})$. Putting everything together, we have that for all $\mathbf{U}, \mathbf{V}$ obtained that way,

$$\|\mathbf{U}\mathbf{V}^\intercal\|_F \leq \left(m^2 m^2 O(n^{4(c-1/2)})\right)^{1/2} = O\left(n^{2c-1+2a}\right) .$$

□

We now formally define the family of distributions that we use to prove our hardness result.

**Definition 11.2.7.** *Given a function $g : \mathbb{R}^m \to [-1, 1]$, we define $\mathcal{D}_g$ to be the class of distributions over $\mathbb{R}^n \times \{\pm 1\}$ of the form $(\mathbf{x}, y)$ such that $\mathbf{x} \sim \mathcal{N}_n$ and $\mathbf{E}[y|\mathbf{x} = \mathbf{z}] = g(\mathbf{U}\mathbf{z})$, where $\mathbf{U} \in \mathbb{R}^{m\times n}$ with $\mathbf{U}\mathbf{U}^\intercal = \mathbf{I}_m$.*

In the following, we show that if $g$ has low-degree moments equal to zero, then distinguishing $\mathcal{D}_g$ from the distribution $(\mathbf{x}, y)$ with $\mathbf{x} \sim \mathcal{N}_n$, $y \sim \mathcal{U}(\{\pm 1\})$ is hard in the SQ model.

**Proposition 11.2.8.** *Let $g : \mathbb{R}^m \to [-1, 1]$ be such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})p(\mathbf{x})] = 0$, for every polynomial $p : \mathbb{R}^m \to \mathbb{R}$ of degree less than $d$, and $\mathcal{D}_g$ be the class of distributions from Definition 11.2.7. Then, if $m \le n^a$, for some constant $a < 1/2$, any SQ algorithm that solves the decision problem $\mathcal{B}(\mathcal{D}_g, \mathcal{N}_n \times \mathcal{U}(\{\pm 1\}))$ must either use queries of tolerance $n^{-\Omega(d)}$ or make at least $2^{n^{\Omega(1)}}$ queries.*

*Proof.* Consider the set of matrices $S$ of Lemma 11.2.5, for an appropriately small value of $c > 0$. Each matrix $\mathbf{U} \in S$ is associated with a unique element of $\mathcal{D}_g$. For every pair of distinct $\mathbf{U}, \mathbf{V} \in S$, we have that

$$\|\mathbf{U}\mathbf{V}^\intercal\|_2 \le \|\mathbf{U}\mathbf{V}^\intercal\|_F \le O(n^{2c-1+2a}) \le n^{-\Omega(1)} ,$$

where for the last inequality we chose $c$ to be a sufficiently small constant, e.g., $c = (1 - 2a)/4$.

Note that the distribution in $\mathcal{D}_g$ associated to a matrix $\mathbf{U}$ has probability density $(1 + g(\mathbf{U}\mathbf{x}))\phi_n(\mathbf{x})$ when conditioned on $y = 1$, and density $(1 - g(\mathbf{U}\mathbf{x}))\phi_n(\mathbf{x})$ when conditioned on $y = -1$. Let $D_{\mathbf{U}}$ be the distribution associated to $\mathbf{U}$ and $D_{\mathbf{V}}$ the distribution associated to $\mathbf{V}$. Denote by $A_{\mathbf{U}}$ the distribution $D_{\mathbf{U}}$ conditioned on the event $y = 1$ and $B_{\mathbf{U}}$ the same distribution conditioned on $y = -1$. Similarly, let $A_{\mathbf{V}}$

and $B_{\mathbf{V}}$ denote the conditional distributions associated with $\mathbf{V}$. Using the definition of pairwise correlation and the fact that $y$ gets each label with equal probability, it follows directly that

$$\chi_{\mathcal{N}_n \times \mathcal{U}(\{\pm 1\})}(D_{\mathbf{U}}, D_{\mathbf{V}}) = \frac{1}{2}\left(\chi_{\mathcal{N}_n}(A_{\mathbf{U}}, A_{\mathbf{V}}) + \chi_{\mathcal{N}_n}(B_{\mathbf{U}}, B_{\mathbf{V}})\right) \ .$$

By Corollary 11.2.4 applied to $A_{\mathbf{U}}, A_{\mathbf{V}}$ and $B_{\mathbf{U}}, B_{\mathbf{V}}$, we obtain

$$\chi_{\mathcal{N}_n}(A_{\mathbf{U}}, A_{\mathbf{V}}) + \chi_{\mathcal{N}_n}(B_{\mathbf{U}}, B_{\mathbf{V}}) \leq \|\mathbf{U}\mathbf{V}^{\intercal}\|_2^d \left(\chi^2(A, \mathcal{N}_m) + \chi^2(B, \mathcal{N}_m)\right) \ ,$$

where $A$ is the distribution of the random variable $\mathbf{U}\mathbf{x}$ for $\mathbf{x} \sim A_{\mathbf{U}}$ (and similarly for $B$). For the $\chi^2$-divergence terms, we have that

$$\begin{aligned}\chi^2(A, \mathcal{N}_m) &= \int_{\mathbb{R}^m} \frac{A^2(\mathbf{z})}{\phi_m(\mathbf{z})} \mathrm{d}\mathbf{z} - 1 = \int_{\mathbb{R}^m} \frac{\phi_m^2(\mathbf{z})\, \mathbf{Pr}^2[y = 1 | \mathbf{x} = \mathbf{z}]}{\phi_m(\mathbf{z})\, \mathbf{Pr}^2[y = 1]} \mathrm{d}\mathbf{z} - 1 \\ &\leq 4 \int_{\mathbb{R}^m} \phi_m(\mathbf{z})\, \mathbf{Pr}[y = 1 | \mathbf{x} = \mathbf{z}] \mathrm{d}\mathbf{z} - 1 = 4\, \mathbf{Pr}[y = 1] - 1 = 1 \ ,\end{aligned}$$

where we used the definition of $A$, Bayes' rule and the fact that $\mathbf{Pr}[y = 1] = 1/2$. Combining the above, we get that $|\chi_{\mathcal{N}_n \times \mathcal{U}(\{\pm 1\})}(D_{\mathbf{U}}, D_{\mathbf{V}})| \leq n^{-\Omega(d)}$. This inequality implies that $\mathrm{SD}(\mathcal{B}, \gamma, \beta) = 2^{\Omega(n^c)}$, for $\gamma = n^{-\Omega(d)}$ and $\beta = O(1)$. Using Lemma 1.4.9, with $\gamma' = \gamma$, completes the proof. □

*Proof of Theorem 11.1.1.* Let $\mathcal{A}$ be an agnostic learner for $\mathcal{C}$. We use $\mathcal{A}$ to solve the decision problem $\mathcal{B}(\mathcal{D}_g, \mathcal{N}_n \times \mathcal{U}(\{\pm 1\}))$, where $g : \mathbb{R}^m \to [-1, 1]$ is the function from Proposition 11.2.1 and $\mathcal{D}_g$ the family of Definition 11.2.7. Let $D'$ be the target distribution, i.e., $D' = \mathcal{N}_n \times \mathcal{U}(\{\pm 1\})$ if the null hypothesis is true or $D' \in \mathcal{D}_g$ otherwise. We feed $\mathcal{A}$ examples drawn from $D'$ and it outputs a hypothesis $h : \mathbb{R}^n \to \{\pm 1\}$ such

that $\mathbf{Pr}_{(\mathbf{x},y)\sim D'}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \frac{\epsilon}{2}$. If $D' \in \mathcal{D}_g$, then for a matrix $\mathbf{U} \in \mathbb{R}^{m\times n}$ with $\mathbf{U}\mathbf{U}^\intercal = \mathbf{I}_m$, we have that $\mathrm{OPT} \leq \mathbf{Pr}_{(\mathbf{x},y)\sim D'}[f(\mathbf{U}\mathbf{x}) \neq y] = \frac{1}{2}\|f - g\|_1 \leq \frac{1}{2}(1-2\epsilon)$, where in the equality we used the fact that the expectation of $y$ conditioned on $\mathbf{x}$ is $g(\mathbf{x})$ and the last inequality is due to Proposition 11.2.1. Combining the above, we get that $\mathbf{Pr}_{(\mathbf{x},y)\sim D'}[h(\mathbf{x}) \neq y] \leq (1-\epsilon)/2$, or equivalently that $\mathbf{E}_{(\mathbf{x},y)\sim D'}[h(\mathbf{x})y] \geq \epsilon$. On the other hand, if the labels were drawn uniformly at random, this correlation would be exactly 0. Therefore, we can distinguish between the two cases by performing a final query of tolerance $\epsilon/2$ for the correlation of $h$ with $y$. □

## 11.3 Explicit SQ Lower Bounds for Boolean Concept Classes

### 11.3.1 LTFs and Degree-$k$ PTFs

Linear threshold functions (LTFs) are Boolean functions of the form $F(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}\rangle + \theta)$, where $\mathbf{w} \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$. A degree-$k$ PTF is any Boolean function of the form $F(\mathbf{x}) = \mathrm{sign}(q(\mathbf{x}))$, where $q : \mathbb{R}^n \to \mathbb{R}$ is a real degree-$k$ polynomial. In this section, we show:

**Theorem 11.3.1** (Degree Lower Bound for PTFs)**.** *There exists a degree-$k$ PTF $f : \mathbb{R} \to \{\pm 1\}$ such that any degree-$d$ polynomial $p : \mathbb{R} \to \mathbb{R}$ with $\|f - p\|_1 < \epsilon$ must have $d = \Omega(k^2/\epsilon^2)$.*

Theorems 11.1.1 and 11.3.1 imply that any SQ algorithm that agnostically learns the class of degree-$k$ PTFs on $\mathbb{R}^n$ under the Gaussian distribution must have complexity at least $n^{\Omega(k^2/\epsilon^2)}$.

We now elaborate on these contributions.

**Lower Bound for LTFs** The $L^1$-regression algorithm [KKMS08] is known to be an agnostic learner for LTFs under Gaussian marginals with complexity $n^{O(1/\epsilon^2)}$. This upper bound uses the known fact that the $L^1$ polynomial $\epsilon$-approximate degree of LTFs under the Gaussian distribution is $d = O(1/\epsilon^2)$ (see, e.g., [DKN10]). This upper bound is tight. Specifically, known results in approximation theory (see Corollary 11.5.2) imply that, any polynomial that $\epsilon$-approximates the function $\mathrm{sign}(t)$ in $L^1$-norm, under the standard Gaussian distribution, requires degree $\Omega(1/\epsilon^2)$. Given this structural result, an application of Theorem 11.1.1, for $m = 1$ and $f(t) = \mathrm{sign}(t)$ gives the tight SQ lower bound of $n^{\Omega(1/\epsilon^2)}$. This bound improves on the best previous bound of $n^{\Omega(1/\epsilon)}$ [GGK20, DKZ20]. Importantly, our approach is much simpler and generalizes to any concept class satisfying the mild assumptions of Theorem 11.1.1.

**Lower Bound for Degree-$k$ PTFs** The $L^1$-regression algorithm is known to be an agnostic learner for degree-$k$ PTFs under Gaussian marginals with complexity $n^{O(k^2/\epsilon^4)}$. This upper bound uses the known upper bound of $O(k\sqrt{\epsilon})$ on the Gaussian noise sensitivity of degree-$k$ PTFs [Kan11], which implies an upper bound of $O(k^2/\epsilon^2)$ on the $L^2$ polynomial $\epsilon$-approximate degree, and therefore an upper bound of $O(k^2/\epsilon^4)$ on the $L^1$ polynomial $\epsilon$-approximate degree. This degree upper bound is not known to be optimal (in fact, it is provably sub-optimal for $k = 1$) and it is a plausible conjecture that the right answer is $\Theta(k^2/\epsilon^2)$. Here we prove a lower bound of $\Omega(k^2/\epsilon^2)$, which applies even for the univariate case.

**Proposition 11.3.2.** *There exists a $(k+1)$-piecewise-constant function $f : \mathbb{R} \to \{0, 1\}$ such that any degree-$d$ polynomial $p : \mathbb{R} \to \mathbb{R}$ that satisfies $\|f - p\|_1 < \epsilon$ must have*

$d = \Omega(k^2/\epsilon^2)$.

An application of Theorem 11.1.1, for $m = 1$ and $f(t)$ being the piecewise constant function of Proposition 11.3.2, implies an SQ lower bound of $n^{\Omega(k^2/\epsilon^2)}$.

Before we provide the formal proof, we sketch the proof of Proposition 11.3.2. The hard function $f$ consists of $k/2$ intervals with the same carefully chosen length; we split each interval in half and we let $f = 0$ in the first half, and $f = 1$ in the second half. We construct a distribution $D$ that puts almost all of its mass in the first half of each interval, matches the first $d$ moments with the standard Gaussian, and $D(x) \leq 2\phi(x)$ for all $\mathbf{x} \in \mathbb{R}$. Then, by construction $\mathbf{E}_{x\sim\mathcal{N}}[f(x)]$ is much larger than the same expectation under $D$. We show that, in fact, this difference bounds from below the error of any degree-$d$ polynomial approximation to the function $f$.

The main technical lemma we establish in this context is the following:

**Lemma 11.3.3.** *There exists a univariate distribution $D$ that (i) matches its first $d$ moments with $\mathcal{N}$, (ii) the pdf of $D$ is at most 2 times the pdf of $\mathcal{N}$ pointwise in $\mathbb{R}$, and (iii) for some $\alpha = \Theta(1/\sqrt{d})$ it holds that $\mathbf{Pr}[(X \mod a) \in (a/2, a)] = 2^{-\Omega(d)}$.*

We defer the proof of Lemma 11.3.3 to Section (11.3.2) and show how it implies Proposition 11.3.2 below.

*Proof of Proposition 11.3.2.* We can assume that $k$ is even. Let $f$ be 1 on the $k/2$ intervals $(ia + a/2, (i+1)a)$, for $i = 0, \ldots, k/2 - 1$, and zero elsewhere. Denote by $D$ the distribution of Lemma 11.3.3. From property (iii), we have that $\mathbf{E}_{x\sim D}[f(x)] = 2^{-\Omega(d)}k$. On the other hand, assuming that $k = O(\sqrt{d})$, we have that $\mathbf{E}_{x\sim\mathcal{N}}[f(x)] = \Omega(k/\sqrt{d})$. This is because the regions where $f$ is 1 are contained in the interval $[0, \Theta(k/\sqrt{d})] \subseteq [0, O(1)]$, where the pdf of the standard Gaussian is bounded below by some constant.

Let $D(x)$ and $\phi(x)$ denote the density on point $x$ of the distribution $D$ and $\mathcal{N}$ respectively. For every polynomial $p : \mathbb{R} \to \mathbb{R}$ of degree at most $d$, it holds

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[f(x)] - \mathop{\mathbf{E}}_{x\sim D}[f(x)] &= \mathop{\mathbf{E}}_{x\sim\mathcal{N}}\left[f(x)\left(1-\frac{D(x)}{\phi(x)}\right)\right] = \mathop{\mathbf{E}}_{x\sim\mathcal{N}}\left[(f(x)-p(x))\left(1-\frac{D(x)}{\phi(x)}\right)\right] \\ &\leq \mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|f(x)-p(x)|]\,,\end{aligned}$$

where the second equality follows from the fact that $D$ matches its first $d$ moments with $\mathcal{N}$, and in the last inequality we used that $0 \leq D(x) \leq 2\phi(x)$ for all $x \in \mathbb{R}$. Thus, if $f$ could be $L^1$-approximated to error $\epsilon$ by a degree-$d$ polynomial, then $\mathbf{E}_{x\sim\mathcal{N}}[f(x)] - \mathbf{E}_{x\sim D}[f(x)]$ would be at most $\epsilon$. But we already showed that this is $\Omega(k/\sqrt{d})$, which implies that $d = \Omega(k^2/\epsilon^2)$. □

### 11.3.2 Proof of Lemma 11.3.3

First, we need the following lemma.

**Lemma 11.3.4.** *There is a $d$-wise independent family of $t = O(d)$ standard Gaussians $X_1, X_2, \ldots, X_t$ such that $\left(\sum_{i=1}^t X_i\right) \mod 1 \in [0, 1/2]$ with probability $1 - 2^{-\Omega(d)}$. Furthermore, such a distribution can be obtained by rejection sampling a set of independent standard Gaussians, where a sample is rejected with probability $1/2$.*

*Proof.* The standard Gaussian distribution can be decomposed into a uniform component and a remaining term. That is, $\mathcal{N} = c\mathcal{U}([0,1]) + (1-c)E$, where $\mathcal{U}([0,1])$ is the uniform distribution in $[0,1]$, $E$ is another distribution, and $c > 0$ is a constant. Let $t \in \mathbb{N}$ such that $t > d/c$. We generate this $d$-wise independent family $X_1, \ldots, X_t$ as follows.

First, we sample $Y_1, \ldots, Y_t$ independent standard Gaussians, writing each $Y_i$ either as a sample from $\mathcal{U}([0,1])$ or a sample from $E$. Then, two complementary cases are

considered.

**Case 1**. The number of $Y_i$'s that came from $\mathcal{U}([0,1])$ is at most $d$. In this case, the sample is rejected with probability $1/2$.

**Case 2**. Otherwise, the sample is rejected if and only if $(\sum_{i=1}^{t} Y_i) \mod 1 \in (1/2, 1]$.

Let $X_1, \ldots, X_t$ be the output of this rejection sampling procedure. The probability that the sample is generated by the first case of the algorithm is exponentially small. To see this, define $Z_i \in \{0,1\}$ to be one if and only $Y_i$ is drawn from $\mathcal{U}([0,1])$. If $C_1$ denotes the event of being in Case 1, then by standard Chernoff bounds we have that

$$\begin{aligned}\mathbf{Pr}[C_1] &= \mathbf{Pr}\left[\sum_{i=1}^{t} Z_i \leq d\right] = \mathbf{Pr}\left[\sum_{i=1}^{t} Z_i \leq \mathbf{E}\left[\sum_{i=1}^{t} Z_i\right]\left(1 - \left(1 - \frac{d}{tc}\right)\right)\right] \\ &\leq \exp\left(-\frac{(1-d/(tc))^2 tc}{2}\right) = 2^{-\Omega(d)}\,,\end{aligned}$$

where we used that $t > d/c$. Therefore, the probability that $(\sum_{i=1}^{t} X_i) \mod 1 \in [0, 1/2]$ is $1 - 2^{-\Omega(d)}$.

Moreover, the probability of accepting the sample is exactly $1/2$ independently of the $Y_i$'s. To see this, let $C_1, C_2 = \overline{C_1}$ be the events of Case 1 and Case 2 being true respectively, and $A$ be the event of accepting the sample. For Case 1, we have $\mathbf{Pr}[A|C_1] = 1/2$. In Case 2, we know that at least one element is drawn from $\mathcal{U}([0,1])$, which means that the $(\sum_{i=1}^{t} X_i) \mod 1$ is going to be uniform in $[0,1]$. Thus, $\mathbf{Pr}[A|C_2] = 1/2$. Therefore, $\mathbf{Pr}[C_1|A] = \mathbf{Pr}[A|C_1]\,\mathbf{Pr}[C_1]/\,\mathbf{Pr}[A] = \mathbf{Pr}[C_1]$ and $\mathbf{Pr}[C_2|A] = \mathbf{Pr}[A|C_2]\,\mathbf{Pr}[C_2]/\,\mathbf{Pr}[A] = \mathbf{Pr}[C_2]$, i.e., accepting is independent of $C_1, C_2$, and thus independent of the sample itself. This means that the output $X_1, \ldots, X_t$ remains Gaussian.

For the $d$-wise independence of the variables $X_1, \ldots, X_t$, let $\mathcal{I}$ be an arbitrary set of at most $d$ indices from $\{1, \ldots, t\}$. We claim that $\{X_i\}_{i \in \mathcal{I}}$ are independent. Case 1 is trivial, since we accept independently of the values of the $Y_i$'s. For Case 2, note that in that case there are more than $d$ $Y_i$'s drawn from $\mathcal{U}([0,1])$. This means that there exists one $j \notin \mathcal{I}$ such that $Y_j$ is uniform and forces the $(\sum_{i=1}^{t} X_i)$ to be uniform in $[0,1]$. Thus, the event $(\sum_{i=1}^{t} X_i) \in [0, 1/2]$ is independent of $\{Y_i\}_{i \in \mathcal{I}}$, and therefore $\{X_i\}_{i \in \mathcal{I}}$ is a set of independent random variables. □

*Proof of Lemma 11.3.3.* Consider the random variable $X = \sum_{i=1}^{t} X_i/\sqrt{t}$ for the $X_i$'s of Lemma 11.3.4. For (i), note that the $d$-th moment involves the expectation of at most $d$ of the $X_i$'s, which are independent. Note that (ii) holds because the distribution of $X$ puts almost all of its mass on half of the real line, and (iii) follows from our scaling of $1/\sqrt{t}$. □

### 11.3.3 Intersections of Halfspaces: Degree Lower Bound via Gaussian Noise Sensitivity

An intersection of $k$ halfspaces on $\mathbb{R}^n$ is any function $f : \mathbb{R}^n \to \{\pm 1\}$ such that there exist $k$ LTFs $h_i : \mathbb{R}^n \to \{\pm 1\}$, $i \in [k]$, such that $f(\mathbf{x}) = 1$ if and only if $h_i(\mathbf{x}) = 1$ for all $i \in [k]$.

The $L^1$-regression algorithm [KKMS08] is known to be an agnostic learner for intersection of $k$ halfspaces on $\mathbb{R}^n$ under Gaussian marginals with complexity $n^{O((\log k)/\epsilon^4)}$. This upper bound uses the known tight upper bound of $O(\sqrt{\epsilon \log k})$ on the Gaussian noise sensitivity of this concept class [KOS08], which implies an upper bound of $O(\log k/\epsilon^4)$ on the $L^1$ polynomial $\epsilon$-degree. This degree upper bound is not known to be optimal (in fact, it is provably suboptimal for $k = 1$) and it is a plausible conjecture that the right

answer is $\Theta(\sqrt{\log k}/\epsilon^2)$. Here we prove a lower bound of $\tilde{\Omega}(\sqrt{\log k}/\epsilon)$, which applies even for $k$-dimensional functions.

**Theorem 11.3.5** (Degree Lower Bound for Intersections of Halfspaces)**.** *There exists an intersection of $k$ halfspaces $f$ on $\mathbb{R}^k$ such that the following holds: Any degree-$d$ polynomial $p : \mathbb{R}^k \to \mathbb{R}$ that satisfies $\|f - p\|_1 < \epsilon$ must have $d = \tilde{\Omega}(\sqrt{\log k}/\epsilon)$.*

Theorem 11.3.5 combined with Theorem 11.1.1, applied for $m = k$ and $f$ being the function from Theorem 11.3.5, implies that any SQ algorithm that agnostically learns intersections of $k$ halfspaces on $\mathbb{R}^n$ under the Gaussian distribution must have complexity at least $n^{\tilde{\Omega}(\sqrt{\log k}/\epsilon)}$.

To prove Theorem 11.3.5, we make essential use of our structural result, Theorem 11.1.2, combined with the following tight lower bound on the Gaussian noise sensitivity of a well-chosen family of intersection of halfspaces.

**Lemma 11.3.6.** *There exists an intersection of $k$ halfspaces on $\mathbb{R}^k$, $f : \mathbb{R}^k \to \{\pm 1\}$, such that $\mathrm{GNS}_\epsilon(f) = \Omega(\sqrt{\epsilon \log k})$.*

*Proof.* We will exhibit a family of $k$ halfspaces whose intersection has the claimed Gaussian noise sensitivity. In particular, these halfspaces will be orthogonal. For $i \in [k]$, let $f_i : \mathbb{R}^n \to \{\pm 1\}$ with $f_i(\mathbf{x}) = \mathrm{sign}(-\langle \mathbf{e}_i, \mathbf{x} \rangle + \theta)$, where $\mathbf{e}_i$ is the vector having 1 in the $i$-th coordinate and 0 elsewhere, and $\theta > 0$ is the bias. That is, $f_i$ is 1 if and only if the $i$-th coordinate is less than $\theta$.

Fix an index $i \in [k]$. The Gaussian noise sensitivity of a single halfspace is $\mathrm{GNS}_\epsilon(f_i) = \Omega(e^{-\frac{\theta^2}{2(1-\epsilon/2)}}\sqrt{\epsilon})$ (see, e.g., [DJS+15, Lemma 3.4] for a proof). Let $\mathbf{x}, \mathbf{y}$ be two $(1-\epsilon)$-correlated $n$-dimensional standard Gaussian random variables. Then, the inner products $\langle \mathbf{e}_i, \mathbf{x} \rangle$ and $\langle \mathbf{e}_i, \mathbf{y} \rangle$ are $(1-\epsilon)$-correlated univariate Gaussians. Since the Gaussian noise

sensitivity of $f_i$ is proportional to the probability that $\langle \mathbf{e}_i, \mathbf{x} \rangle < \theta < \langle \mathbf{e}_i, \mathbf{y} \rangle$, we have that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^{1-\epsilon}}[\langle \mathbf{e}_i, \mathbf{x} \rangle < \theta < \langle \mathbf{e}_i, \mathbf{y} \rangle] = \Omega(e^{-\frac{\theta^2}{2(1-\epsilon/2)}}\sqrt{\epsilon}) \; .$$

Let $\theta$ be the threshold for which $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}_n}[\langle \mathbf{e}_i, \mathbf{x} \rangle > \theta] = 1/k$. The standard bound for the Gaussian tail is $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}_n}[\langle \mathbf{e}_i, \mathbf{x} \rangle > \theta] = \Theta(e^{-\theta^2/2}/\theta)$. Therefore, for the $\theta$ that we selected it holds $\mathbf{Pr}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^{1-\epsilon}}[\langle \mathbf{e}_i, \mathbf{x} \rangle < \theta < \langle \mathbf{e}_i, \mathbf{y} \rangle] = \Omega(\theta\sqrt{\epsilon}/k) = \Omega(\sqrt{\epsilon \log k}/k)$.

Let $f : \mathbb{R}^n \to \{\pm 1\}$ be 1 if and only if $f_i$ is 1 for all $i \in [k]$. Then, we have that

$$\begin{aligned}\mathrm{GNS}_\epsilon(f) &= 2 \mathop{\mathbf{Pr}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^{1-\epsilon}}[f(\mathbf{x}) = 1, f(\mathbf{y}) = -1] = \mathbf{Pr}[f(\mathbf{x}) = 1] - \mathop{\mathbf{Pr}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^{1-\epsilon}}[f(\mathbf{x}) = f(\mathbf{y}) = 1] \\ &= \left(1 - \frac{1}{k}\right)^k - \left(1 - \frac{1}{k} - \Omega\left(\frac{\sqrt{\epsilon \log k}}{k}\right)\right)^k \; ,\end{aligned}$$

where the $k$-th powers are due to the fact that $\langle \mathbf{e}_i, \mathbf{x} \rangle$ and $\langle \mathbf{e}_j, \mathbf{x} \rangle$ are independent for $i \neq j$. We can use the Taylor expansion to show that the above difference is $\Omega(\sqrt{\epsilon \log k})$. Let the function $h(t) = (1 - 1/k + t)^k$. By Taylor's theorem, $h(0) - h(t) = -h'(0)t - h''(\xi)t^2/2$, for some $\xi$ between $t$ and 0. By calculating the derivatives, setting $t = -\Omega(\sqrt{\epsilon \log k}/k)$ and noting that the second term of the approximation is less than the first one, we get that $h(0) - h(t) = \Omega\left(\frac{\sqrt{\epsilon \log k}}{k}\right) k \left(1 - \frac{1}{k}\right)^{k-1}$. □

### 11.3.4 Proof of Theorem 11.1.2

We require the following proposition.

**Proposition 11.3.7.** *Let $p(\theta)$ be a degree-$d$ polynomial on the circle, i.e., a degree at most $d$ polynomial in $\sin\theta$ and $\cos\theta$, and let $B(\theta)$ be a Boolean-valued function that is*

*periodic modulo* $2\pi$*. Then, for* $t$ *being a sufficiently small multiple of* $\log d/d$*, it holds*

$$\frac{1}{2\pi}\int_0^{2\pi} |p(\theta) - B(\theta)|\mathrm{d}\theta = \tilde{\Omega}(1/\log d) \mathop{\mathbf{Pr}}_{\phi\sim\mathcal{U}([0,2\pi])}[B(\phi - t) \neq B(\phi + t)] \,.$$

*Proof.* We can assume that $\frac{1}{2\pi}\int_0^{2\pi} |p(\theta)|\mathrm{d}\theta$ is at most 2, since otherwise the $\frac{1}{2\pi}\int_0^{2\pi} |p(\theta) - B(\theta)|\mathrm{d}\theta$ is at least 1. Let $k$ be an odd integer proportional to $\log d$. We start with the following technical claim.

**Claim 11.3.8.** *For any* $\theta \in [0, 2\pi]$*, it holds* $|p^{(k)}(\theta)| = O(d)^k$*.*

*Proof.* Using $\cos\theta = \left(e^{i\theta} + e^{-i\theta}\right)/2$ and $\sin\theta = \left(e^{i\theta} - e^{-i\theta}\right)/2$, we write $p(\theta) = \sum_{n=-\infty}^{\infty} a_n e^{ni\theta}$, for some coefficients $a_n$, where $a_n = \frac{1}{2\pi}\int_0^{2\pi} p(\phi)e^{-ni\theta}\mathrm{d}\phi$. Since $p$ has degree at most $d$, it holds that $a_n = 0$, for all $n > d$ and $n < -d$. Therefore, we have that $p(\theta) = \sum_{n=-d}^{d} \frac{1}{2\pi}\int_0^{2\pi} p(\phi)e^{ni(\theta-\phi)}\mathrm{d}\phi$. Taking the $k$-th derivative (using Leibniz's rule) gives

$$p^{(k)}(\theta) = \sum_{n=-d}^{d} \frac{1}{2\pi}\int_0^{2\pi} p(\phi)(ni)^k e^{ni(\theta-\phi)}\mathrm{d}\phi \,.$$

This implies that

$$|p^{(k)}(\theta)| \leq \sum_{n=-d}^{d} \frac{1}{2\pi}\int_0^{2\pi} |p(\phi)|n^k\mathrm{d}\phi \leq 2\sum_{n=-d}^{d} n^k = O(d^{k+1}) \,.$$

Moreover, $k$ is proportional to $\log d$, thus $|p^{(k)}(\theta)| = O(d)^k$, for all $\theta \in [0, 2\pi]$. □

We next pick $t$ to be a small multiple of $\log d/d$ and $\phi \in [0, 2\pi]$. Let $z_m = t\cos(\pi m/k) + \phi$, for $m = 0, 1, \ldots, k$, and let $q(z) = \sum_{j=0}^{k} c_j z^j$ be the unique degree-$k$ polynomial such that $q(z_m) = p(z_m)$, for $m = 0, 1, \ldots, k$. Observe that $q - p$ has $k + 1$ zeroes. Therefore, iterating Rolle's theorem, we obtain that there is a point

$\phi - t \leq z \leq \phi + t$ such that $p^{(k)}(z) = q^{(k)}(z)$, and thus $|q^{(k)}(z)| = O(d)^k$, or equivalently $c_k = 2^k O(d/k)^k$.

Let $R(\theta) = q(t\cos\theta + \phi)$. For some constants $b_n$ (which depend on $t$ and $\phi$), we have that $R(\theta) = \sum_{n=-k}^{k} b_n e^{ni\theta}$. Since $R(\theta)$ is an even function, its Fourier coefficients are real numbers. The following claim provides an upper bound on the coefficient $b_k$.

**Claim 11.3.9.** *It holds that* $|b_k| \leq 1/(4k)$.

*Proof.* Note that $b_k = (1/2\pi)\int_0^{2\pi} R(\theta)e^{-ki\theta}\mathrm{d}\theta$. Using the orthogonality of the trigonometric polynomials, only terms containing $\cos(k\theta)$ are non-zero. Moreover, $\cos^k\theta = \sum_{j=0}^{k} u_j \cos(j\theta)$ with $u_k = 2^{-k+1}$, which can be verified using the identity $\cos\theta = (e^{i\theta} + e^{-i\theta})/2$. Therefore, we have that

$$b_k = \frac{1}{2\pi}\int_0^{2\pi} R(\theta)e^{-ki\theta}\mathrm{d}\theta = \frac{1}{2\pi}\int_0^{2\pi} c_k u_k t^k \cos(k\theta)e^{-ki\theta}\mathrm{d}\theta = c_k u_k \frac{t^k}{2\pi}\pi = \left(\frac{t}{2}\right)^k c_k \ ,$$

where we used that $R(\theta) = \sum_{j=0}^{k} c_j (t\cos\theta + \phi)^j$. Since $c_k = 2^k O(d/k)^k$, we have that $b_k = O(td/k)^k$; this is at most $1/(4k)$, if $t$ is a small enough multiple of $\log d/d$. □

On the other hand, by doing a filtering using the $(2k)$-th roots of unity, we get that $\sum_{m=0}^{2k-1} R(2\pi m/(2k)) = 2kb_k$, and this is equivalent to $\sum_{m=-k+1}^{k} q(t\cos(\pi m/k) + \phi)(-1)^m = 2kb_k$. Therefore,

$$b_k = \frac{1}{2k}\sum_{m=-k+1}^{k} q(t\cos(\pi m/k) + \phi)(-1)^m = \frac{1}{2k}\sum_{m=-k+1}^{k} p(z_{|m|})(-1)^m =$$
$$\frac{1}{2k}\Big(\sum_{m=-k+1}^{k} (p(z_{|m|}) - B(z_{|m|}))(-1)^m + \sum_{m=-k+1}^{k} B(z_{|m|})(-1)^m + (B(\phi+t) - B(\phi-t))\Big) \ .$$

Since $k-1$ is even and $B$ is Boolean, $2\sum_{m=1}^{k-1} B(z_m)(-1)^m$ is a multiple of 4. If $B(\phi+t) \neq B(\phi-t)$, the reverse triangle ineq. gives $\left|B(\phi+t) - B(\phi-t) + 2\sum_{m=1}^{k-1} B(z_m)(-1)^m\right| \geq$

2. Therefore, in this case, we have that $\frac{1}{4k} > |b_k| \geq \frac{1}{2k}\left(2 - \sum_{m=-k+1}^{k} |p(z_{|m|}) - B(z_{|m|})|\right)$ or in other words,

$$\sum_{m=-k+1}^{k} |p(z_{|m|}) - B(z_{|m|})| \geq \mathbb{1}\{B(\phi+t) \neq B(\phi-t)\} .$$

Integrating this over $\phi$ from 0 to $2\pi$ gives

$$\int_0^{2\pi} |p(\theta) - B(\theta)| d\theta \geq \frac{\pi}{k} \Pr_{\phi \sim \mathcal{U}([0,2\pi])}[B(\phi-t) \neq B(\phi+t)] .$$

The result follows from our assumption that $k$ is proportional to $\log d$. □

Using Proposition 11.3.7, we can prove the main theorem of this section.

*Proof of Theorem 11.1.2.* The latter statement follows from the fact that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - p(\mathbf{x})|] \geq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - \mathrm{sign}(p(\mathbf{x}))|/2]$. On the other hand, we can write

$$\mathrm{GNS}_\epsilon(f) - \mathrm{GNS}_\epsilon(\mathrm{sign}(p)) = \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_d^\epsilon}[f(\mathbf{x}) \neq f(\mathbf{y})] - \Pr_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_d^\epsilon}[\mathrm{sign}(p(\mathbf{x})) \neq \mathrm{sign}(p(\mathbf{y}))]$$

$$\leq \Pr_{\mathbf{x}\sim\mathcal{N}_n}[f(\mathbf{x}) \neq \mathrm{sign}(p(\mathbf{x}))] + \Pr_{\mathbf{x}\sim\mathcal{N}_n}[f(\mathbf{y}) \neq \mathrm{sign}(p(\mathbf{y}))] = 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - \mathrm{sign}(p(\mathbf{x}))| .$$

Combining these, we find that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - p(\mathbf{x})|] \geq \mathrm{GNS}_\epsilon(f)/4 - \mathrm{GNS}_\epsilon(\mathrm{sign}(p))/4$. The result then follows from noting that $\mathrm{sign}(p)$ is a degree-$d$ PTF, and therefore by [Kan11] it holds that $\mathrm{GNS}_\epsilon(\mathrm{sign}(p)) = O(d\sqrt{\epsilon})$.

For the first statement, let $\mathbf{y}$ and $\mathbf{z}$ be independent Gaussians and let $\mathbf{x}(\phi) = \cos\phi\,\mathbf{y} + \sin\phi\,\mathbf{z}$. Let $a$ be a sufficiently small multiple of $\log d/d$. For any $\phi \in [0, 2\pi]$, $\mathbf{x}(\phi - a)$ and $\mathbf{x}(\phi + a)$ are $(1-\delta)$-correlated Gaussian random variables, where $\delta = \Theta(\log d/d)^2$.

We have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x})-p(\mathbf{x})|] &= \mathop{\mathbf{E}}_{\phi\in\mathcal{U}([0,2\pi])}\left[\mathop{\mathbf{E}}_{\mathbf{y},\mathbf{z}\sim\mathcal{N}_n}[|f(\mathbf{x}(\phi))-p(\mathbf{x}(\phi))|]\right] \\ &= \mathop{\mathbf{E}}_{\mathbf{y},\mathbf{z}\sim\mathcal{N}_n}\left[\mathop{\mathbf{E}}_{\phi\in\mathcal{U}([0,2\pi])}[|f(\mathbf{x}(\phi))-p(\mathbf{x}(\phi))|]\right] \\ &\geq \Omega(1/\log(d)) \mathop{\mathbf{E}}_{\mathbf{y},\mathbf{z}\sim\mathcal{N}_n}[\mathop{\mathbf{Pr}}_{\phi\in\mathcal{U}([0,2\pi])}[f(\mathbf{x}(\phi-a))\neq f(\mathbf{x}(\phi+a))]]\,,\end{aligned}$$

where in the inequality we used Proposition 11.3.7. Moreover, using Fubini's theorem, we have

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x})-p(\mathbf{x})|] \geq \Omega(1/\log(d)) \mathop{\mathbf{E}}_{\phi\in\mathcal{U}([0,2\pi])}\left[\mathop{\mathbf{Pr}}_{\mathbf{y},\mathbf{z}\sim\mathcal{N}_n}[f(\mathbf{x}(\phi-a))\neq f(\mathbf{x}(\phi+a))]\right] \\ &= \Omega(1/\log(d)) \mathop{\mathbf{E}}_{\phi\in\mathcal{U}([0,2\pi])}[\mathrm{GNS}_\delta(f)] = \Omega(1/\log(d))\mathrm{GNS}_\delta(f) \\ &= \Omega(1/\log(d))\mathrm{GNS}_{(\log(d)/d)^2}(f)\,.\end{aligned}$$

□

## 11.4 Lower Bound for Real-Valued Functions

### 11.4.1 CSQ Lower Bound: Proof of Theorem 11.1.3

To prove our CSQ lower bound, we need to find a hard function $g:\mathbb{R}^m\to\mathbb{R}$ that is uncorrelated with low-degree polynomials and, at the same time, is close to $f$ in the $L^2$-sense. Instead of using duality to establish the existence of such a function $g$, we let $g$ be the orthogonal component of the truncated Hermite expansion of $f$.

*Proof of Theorem 11.1.3.* Let an algorithm $\mathcal{A}$ that agnostically learns $\mathcal{C}$ up to $L^2$-error

$\epsilon$. Let $g(\mathbf{x}) = f(\mathbf{x}) - \sum_{i=0}^{d-1} f^{[i]}(\mathbf{x})$, i.e., $g$ is the same as the function $f$ without the low-degree moments up to $d-1$. Note that $\|g\|_2 \geq \epsilon$. Let $C = 2/(\epsilon \|g\|_2)$ and let $S$ be the set of nearly orthogonal matrices of Lemma 11.2.5. Consider the class $\mathcal{C}_g$ that consists of all functions from $\mathbb{R}^n$ to $\mathbb{R}$ of the form $G_{\mathbf{V}}(\mathbf{x}) = Cg(\mathbf{V}\mathbf{x})$, for any matrix $\mathbf{V} \in S$. Every $G_{\mathbf{V}} \in \mathcal{C}_g$ is orthogonal to all polynomials of degree less than $d$, and also $\|G_{\mathbf{V}}\|_2 = 2/\epsilon$. We feed $\mathcal{A}$ with samples $(\mathbf{x}, G_{\mathbf{V}}(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}_n$, $\mathbf{V} \in S$. Let $\epsilon' > 0$ be the accuracy parameter used with $\mathcal{A}$. Then, $\mathcal{A}$ returns a hypothesis $h$ satisfying

$$\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[(h(\mathbf{x}) - G_{\mathbf{V}}(\mathbf{x}))^2]} \leq \mathrm{OPT} + \epsilon' \,. \tag{11.2}$$

For our choice of $C$, the optimal error becomes

$$\begin{aligned}\mathrm{OPT} &\leq \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[(f(\mathbf{x}) - G_{\mathbf{V}}(\mathbf{x}))^2]} = \sqrt{1 + C^2\|g\|_2^2 - 2C \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})]} \\ &\leq \sqrt{1 + \frac{4}{\epsilon^2} - \frac{2\|g\|_2}{\epsilon}} \leq \sqrt{\frac{4}{\epsilon^2} - 1} \leq \frac{2}{\epsilon}\sqrt{1 - \frac{\epsilon^2}{4}} \leq \frac{2}{\epsilon} - \frac{\epsilon}{4} \,,\end{aligned}$$

where in the second inequality we used that $2\,\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})] \geq \|g\|_2^2$. By choosing $\epsilon' = \epsilon/8$, Equation (11.2) becomes $\|h - G_{\mathbf{V}}\|_2 \leq 2/\epsilon - \epsilon/8$.

It remains to bound from above the pairwise correlation of the class $\mathcal{C}_g$. For any two different $\mathbf{U}, \mathbf{V} \in S$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[G_{\mathbf{U}}(\mathbf{x})G_{\mathbf{V}}(\mathbf{x})] &\leq C^2 \sum_{t=0}^{\infty} \|\mathbf{U}\mathbf{V}^\intercal\|_2^t \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[(g^{[t]}(\mathbf{x}))^2] \leq C^2 \|\mathbf{U}\mathbf{V}^\intercal\|_2^d \sum_{t=d}^{\infty} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[(g^{[t]}(\mathbf{x}))^2] \\ &\leq 4\epsilon^{-2}\|\mathbf{U}\mathbf{V}^\intercal\|_F^d \leq \epsilon^{-2} n^{-\Omega(d)} \leq n^{-\Omega(d)} \,,\end{aligned}$$

where in the first inequality we used Lemma 11.2.3, in the second inequality we used

the fact that $g$ is uncorrelated with all polynomials of degree less than $d$, the third inequality follows from Parseval's identity and the fact that $\|g\|_2 C = 2/\epsilon$, the next one follows from Lemma 11.2.5, and the last one from our assumption $\epsilon > n^{-c}$ for an appropriate constant $c$. As a note, we extend our class $\mathcal{C}_g$ to include the identically zero function, which does not increase the pairwise correlations. Using Lemma 1.4.13 with $\gamma' = \gamma$, we have that $\mathrm{CSDA}_{\mathcal{N}_n}(\mathcal{C}_g, 2\gamma) = 2^{n^{\Omega(1)}}$ for $\gamma = n^{-\Omega(d)}$. An application of Lemma 1.4.14 with $\eta = 2/\epsilon$ concludes the proof. □

### 11.4.2 SQ Lower Bound: Proof of Theorem 11.1.4

To prove lower bounds for the general SQ model, we require our hard function $g$ to be pointwise bounded. This allows us to define a learning problem with Boolean labels, for which we have SQ lower bounds ready to be used. Because of our $L^\infty$ constraint on $g$, the resulting lower bound is expressed in terms of the degrees of polynomials that approximate $f$ in $L^1$ rather than $L^2$ sense.

Our duality argument will now use the pair of dual norms $L^1, L^\infty$.

**Proposition 11.4.1.** *Let $f \in L^2(\mathbb{R}^m)$ be such that for any degree at most $d-1$ polynomial $p : \mathbb{R}^m \to \mathbb{R}$, it holds $\|f - p\|_1 \geq \epsilon$. Then, there exists a function $g : \mathbb{R}^m \to [-1, 1]$ such that:*

1. $\mathbf{E}_{x \sim \mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})] \geq \epsilon$, *and,*
2. $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0$, *for any polynomial $P : \mathbb{R}^m \to \mathbb{R}$ with degree less than $d$.*

*Proof.* The function $g$ is a solution to the infinite system:

$$(*)\begin{cases} \mathop{\mathbf{E}}_{x\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})] \geq \epsilon \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0 & \forall P \in \mathcal{P}^m_{d-1} \\ \|g\|_\infty \leq 1 \end{cases}$$

This is equivalent to the following LP:

$$(**)\begin{cases} -\mathop{\mathbf{E}}_{x\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})] + \epsilon \leq 0 \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0 & \forall P \in \mathcal{P}^m_{d-1} \\ \mathop{\mathbf{E}}_{x\sim\mathcal{N}_m}[g(\mathbf{x})h(\mathbf{x})] \leq \|h\|_1 & \forall h \in L^1(\mathbb{R}^m) \end{cases}$$

From Corollary 11.6.2, the above LP is feasible unless the following is infeasible:

$$(**')\begin{cases} \|h\|_1 - \lambda\epsilon < 0 \\ h(\mathbf{x}) + P(\mathbf{x}) - \lambda f(\mathbf{x}) = 0, & \forall \mathbf{x} \in \mathbb{R}^m \\ \lambda \geq 0, h \in L^1(\mathbb{R}^m), P \in \mathcal{P}^m_{d-1} \end{cases}$$

Let $(h, P, \lambda)$ be a solution to $(**')$. Note that we can assume that $\lambda = 1$ since all constraints are homogeneous. Then, the constraints become $h = f - P$ and

$$\|f - P\|_1 < \epsilon\,,$$

which is a contradiction. Therefore, the original system $(*)$ is feasible. □

We conclude with the proof of the main theorem for this section.

*Proof of Theorem 11.1.4.* Suppose that we have such an agnostic learner $\mathcal{A}$. Let $g$ :

$\mathbb{R}^m \to [-1, 1]$ be the function of Proposition 11.4.1, for a parameter $\epsilon' > 0$ to be specified. Let $\mathcal{D}_g$ be the family of distributions over $\mathbb{R}^n \times \{\pm 1\}$ from Definition 11.2.7. We use $\mathcal{A}$ to solve the problem of distinguishing between a distribution from $\mathcal{D}_g$ and the distribution where the labels are drawn uniformly at random. That is, we convert $\mathcal{A}$ into an algorithm for $\mathcal{B}(\mathcal{D}_g, \mathcal{N}_n \times \mathcal{U}(\{\pm 1\}))$, and the hardness result will follow from the hardness of that decision problem, as established by Proposition 11.2.8.

Let $D'$ be a distribution that is either $D' = \mathcal{N}_n \times \mathcal{U}(\{\pm 1\}))$ or $D' \in \mathcal{D}_g$. We feed $\mathcal{A}$ a set of i.i.d. samples of the form $(\mathbf{x}, Cy)$, where $(\mathbf{x}, y) \sim D'$ and $C = 1/\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})]$. Let $\epsilon' > 0$ be the accuracy parameter used when running $\mathcal{A}$ and $h$ be the returned hypothesis. We have that

$$\sqrt{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'}[(h(\mathbf{x}) - Cy)^2]} \leq \mathrm{OPT} + \epsilon' \,. \tag{11.3}$$

If $D' \in \mathcal{D}_g$, for the optimal error we have that

$$\mathrm{OPT} \leq \sqrt{1 + C^2 - 2C \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})]} \leq \sqrt{C^2 - 1} = C\sqrt{1 - 1/C^2} \leq C - 1/(2C) \,.$$

If we choose $\epsilon' = 1/(4C)$, Equation (11.3) becomes $\sqrt{\mathbf{E}_{(\mathbf{x},y)\sim D'}[(h(\mathbf{x}) - Cy)^2]} \leq C - 1/(4C)$. On the other hand, we write $\sqrt{\mathbf{E}_{(\mathbf{x},y)\sim D'}[(h(\mathbf{x}) - Cy)^2]} \geq \sqrt{C^2 - 2C\,\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[h(\mathbf{x})y]}$. Combining these two, we obtain

$$2C \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[h(\mathbf{x})y] \geq C^2 - (C - 1/(4C))^2 \geq 1/3 \,,$$

which gives that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[h(\mathbf{x})y] \geq 1/(6C) = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})]/6 \geq \epsilon/6$ from Proposition 11.4.1.

Note that if $D' = \mathcal{N}_n \times \mathcal{U}(\{\pm 1\})$, then $\mathbf{E}_{(\mathbf{x},y)\sim D'}[h(\mathbf{x})y] = 0$. Therefore, by performing a query of tolerance $\Omega(\epsilon)$ for the correlation of $h$ with the labels, we can distinguish between the two cases of our hypothesis testing problem. By Proposition 11.2.8, this requires either $2^{n^{\Omega(1)}}$ queries or queries of tolerance $n^{-\Omega(d)}$. □

## 11.5 Applications for Classes of Real-Valued Functions

### 11.5.1 ReLU Activation

The class of Rectified Linear Unit (ReLU) functions consists of all functions of the form $\mathrm{ReLU}(\langle \mathbf{w}, \mathbf{x} \rangle)$, where $\mathbf{w} \in \mathbb{R}^n$ is any vector with $\|\mathbf{w}\|_2 = 1$ and $\mathrm{ReLU} : \mathbb{R} \to \mathbb{R}$ is defined as $\mathrm{ReLU}(t) = \max\{0, t\}$.

Upper and lower bounds for agnostically learning ReLUs were given in [GGK20, DKZ20]. [DKZ20] established an SQ lower bound of $n^{\Omega(1/\epsilon^c)}$, for some constant $c > 0$. This constant $c$ was not explicitly calculated in [DKZ20], but can be shown to be approximately $1/40$. [GGK20] gave an SQ lower bound of $n^{\Omega(1/\epsilon^{1/36})}$ for this problem. We note that [GGK20] considered a correlational type of guarantee, i.e., finding a hypothesis whose correlation with the labels is within $\epsilon$ of the optimal, as opposed to $L^2$-error. For this correlational guarantee, the upper bound of [GGK20] is an $L^2$-regression algorithm with complexity $n^{O(\epsilon^{-4/3})}$, and the lower bound states that any SQ algorithm needs to perform queries with tolerance $\tau < n^{-\Omega(\epsilon^{-1/12})}$ or at least $2^{n^{\Omega(1)}}\epsilon$ queries. Furthermore, [GGK20] showed that any agnostic learner with the square loss guarantee can be run with increased accuracy to satisfy the correlational guarantee. This reduction costs a "third root" in the exponent, yielding an $n^{\Omega(\epsilon^{-1/36})}$ SQ lower

bound for the square loss guarantee. As a note, [GGK20] assumes bounded labels. In this setting, agnostically learning within $L^2$-error $\mathrm{OPT} + \epsilon$ is equivalent to agnostically learning in squared $L^2$-error $\mathrm{OPT} + \epsilon'$, for $\epsilon' = \Theta(\epsilon)$.

Given the context of prior work, we can now present our results. To apply our generic lower bound theorems, we bound from below the degree of any polynomial that $\epsilon$-approximates the univariate ReLU function. This can be done by appealing to a known powerful theorem from the approximation theory literature by Ganburg [Gan02, Gan08]. This result can be used to derive tight polynomial degree lower bounds for the ReLU function and the sign function (see Corollary 11.5.2).

Let $A_\sigma(f)_p = \inf_{g \in B_\sigma} \|f - g\|_p$, where $B_\sigma$, $\sigma > 0$ is the class of all entire functions of exponential type $\sigma$, i.e., the class consisting of every entire function $g$ such that for every $\epsilon > 0$ there exists a $C$ for which $|g(z)| \leq Ce^{\sigma(1+\epsilon)|z|}$.

**Fact 11.5.1.** *For any function $f : \mathbb{R} \to \mathbb{R}$ of polynomial growth*

$$\lim_{n\to\infty} \left(\frac{b_n}{\sigma}\right)^{1/p} \inf_{p \in \mathcal{P}_n} \left\| f\left(\frac{b_n}{\sigma}x\right) - p(x) \right\|_p = A_\sigma(f)_p \,,$$

*where $b_n = 2\sqrt{n}$, $p \in [1, 2]$ and $A_\sigma(f)_p$ is the error of the best approximation of $f$ by entire functions of exponential type $\sigma$ in $L^p(\mathbb{R})$.*

By selecting $f(t) = \mathrm{sign}(t)$ and $p = 1$ in Fact 11.5.1, we get that any polynomial that achieves error at most $\epsilon$ with respect to the $L^1$-norm must have degree at least $\Omega(1/\epsilon^2)$.

**Corollary 11.5.2.** *Let $f : \mathbb{R} \to \{\pm 1\}$ with $f(t) = \mathrm{sign}(t)$. Any polynomial $p : \mathbb{R} \to \mathbb{R}$ satisfying $\|f - p\|_1 \leq \epsilon$ must have degree $d = \Omega(1/\epsilon^2)$.*

We also obtain:

**Corollary 11.5.3.** *Let $f : \mathbb{R} \to \mathbb{R}$ be the ReLU function* $\mathrm{ReLU}(t) = \max\{0, t\}$ *and $p \in [1, 2]$. The minimum integer $d$ for which there exists a degree-$d$ polynomial $P : \mathbb{R} \to \mathbb{R}$ such that $\|\mathrm{ReLU} - P\|_p \leq \epsilon$ is $d = \Theta\left(\epsilon^{-\frac{2}{1+1/p}}\right)$.*

Therefore, Theorems 11.1.3 and 11.1.4 imply a complexity of at least $n^{\Omega(\epsilon^{-4/3})}$ for any agnostic CSQ learner; and $n^{\Omega(\epsilon^{-1})}$ for any agnostic SQ learner respectively.

### 11.5.2 Sigmoid Activation

#### CSQ Lower Bound

We now let $f$ be the standard sigmoid function, defined as $f(t) = 1/(1 + e^{-t})$, $t \in \mathbb{R}$. We first focus on bounding the degree of polynomials that approximate $f$ in $L^2$-norm. This can be done via Hermite analysis, in particular, based on the fact that the polynomial of degree $d$ being closest to $f$ in $L^2$-norm is the truncated Hermite expansion $p_d(t) = \sum_{i=0}^{d} \hat{f}(i) H_i(t)$. The error of this approximation is $\|p_d - f\|_2^2 = \sum_{i=d+1}^{\infty} \hat{f}^2(i)$. For the asymptotic behavior of the Hermite coefficients, we use the following fact (see [GGJ+20] and the references therein).

**Fact 11.5.4** (Lemma A.9 from [GGJ+20])**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be the standard sigmoid function $f(t) = 1/(1 + e^{-t})$ and $\hat{f}(i)$ be its Hermite coefficients for $i \in \mathbb{Z}_+$. Then, $\hat{f}(0) = 0.5$, $\hat{f}(2i) = 0$ and $\hat{f}(2i - 1) = e^{-\Theta(\sqrt{i})}$, for $i \geq 1$.*

From this fact, we get the bound on the $L^2$-error of the best polynomials of degree $d$.

**Corollary 11.5.5** ($L^2$-Degree Lower Bound for Sigmoid)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be the standard sigmoid function $f(t) = 1/(1 + e^{-t})$ and $d$ be the smallest integer for which there exists a degree-$d$ polynomial $p : \mathbb{R} \to \mathbb{R}$ such that $\|f - p\|_2 < \epsilon$. Then $d = \Theta(\log^2(1/\epsilon))$.*

*Proof.* Fix a degree $k$. From Fact 11.5.4, the best $k$-degree polynomial $p_k$ achieves error

$$\|f - p_k\|_2^2 = \sum_{i=k+1}^{\infty} \hat{f}^2(i) = \sum_{i>k, i \text{ odd}} e^{-\Theta(\sqrt{i})} = \sqrt{k} e^{-\Theta(\sqrt{k})} .$$

This becomes $\epsilon^2$ when $k$ becomes $\Theta(\log^2(1/\epsilon))$. □

By Theorem 11.1.3, we get that any CSQ agnostic learner for sigmoids has complexity $n^{\Omega(\log^2(1/\epsilon))}$.

### SQ Lower Bound

The approach to derive lower bounds for the degrees of $L^1$-approximating polynomials will be to relate the $L^1$-norm to the $L^2$-norm and use the lower bounds for the latter. In particular, we will use the following fact about polynomials under the Gaussian measure.

**Theorem 11.5.6** (Hypercontractivity [Bog98, Nel73])**.** *If $p$ is a $d$-degree polynomial and $t > 2$, then*

$$\|p\|_t \leq (t-1)^{d/2} \|p\|_2 .$$

**Claim 11.5.7.** *Let $r \in L^4(\mathbb{R})$. Then, $\|r\|_2 \leq \|r\|_1^{1/3} \|r\|_4^{2/3}$.*

*Proof.* The proof follows from two applications of the Cauchy-Schwartz inequality.

$$\mathop{\mathbf{E}}_{t\sim\mathcal{N}}[r^2(t)] \leq \mathop{\mathbf{E}}_{t\sim\mathcal{N}}[|r(t)|]^{1/2} \mathop{\mathbf{E}}_{t\sim\mathcal{N}}\left[|r(t)|^3\right]^{1/2} \leq \mathop{\mathbf{E}}_{t\sim\mathcal{N}}[|r(t)|]^{1/2} \mathop{\mathbf{E}}_{t\sim\mathcal{N}}\left[|r(t)|^2\right]^{1/4} \mathop{\mathbf{E}}_{t\sim\mathcal{N}}\left[|r(t)|^4\right]^{1/4} .$$

Rearranging the above, yields the claimed inequality. □

We can now show our $L^1$ polynomial degree lower bound.

**Theorem 11.5.8** ($L^1$-Degree Lower Bound for Sigmoid)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be the standard sigmoid function $f(t) = 1/(1+e^{-t})$ and $0 < \epsilon < 1$. Any degree-$d$ polynomial $p : \mathbb{R} \to \mathbb{R}$ that satisfies $\|f - p\|_1 < \epsilon$ must have $d = \Omega(\log(1/\epsilon))$.*

*Proof.* Let $p : \mathbb{R} \to \mathbb{R}$ be a degree-$d$ polynomial such that $\|f - p\|_1 < \epsilon$. Using Theorem 11.5.6 with $t = 4$ and then Claim 11.5.7 with $r(t) = p(t)$, we get that

$$\|p\|_4 \leq 3^{d/2} \|p\|_2 \leq 3^{d/2} \|p\|_1^{1/3} \|p\|_4^{2/3} .$$

After dividing both sides by $\|p\|_4^{2/3}$, we have that $\|p\|_4 \leq 3^{3d/2} \|p\|_1$. Furthermore, using the triangle inequality, $\|p\|_1 \leq \epsilon + \|f\|_1 = O(1)$. Therefore, $\|p\|_4 \leq 2^{O(d)}$. Furthermore, Claim 11.5.7 for $r(t) = f(t) - p(t)$ gives

$$\|f - p\|_2 \leq \|f - p\|_1^{1/3} \|f - p\|_4^{2/3} \leq \epsilon^{1/3} 2^{O(d)} .$$

On the other hand, for the $L^2$-error we have that $\|f - p\|_2 \geq \sqrt{d} e^{-\Theta(\sqrt{d})}$ (Corollary 11.5.5). Combining the two bounds, it follows that $d = \Omega(\log(1/\epsilon))$. □

We note that [GGK20] showed an $n^{\Omega(\log^2(1/\epsilon))}$ SQ lower bound for the correlational guarantee.

## 11.6 Duality in Infinite-Dimensional LP

We start with some basic definitions.

**Dual Norms** Consider a vector space $V$ with inner product $\langle \cdot, \cdot \rangle$ and a norm $\|\cdot\|$ on $V$. The *dual norm* $\|f\|_*$, $f \in V$, is defined as $\|f\|_* = \sup\{\langle f, h \rangle : \|h\| \leq 1\}$. Hölder's

inequality states that for any $f, h \in V$ it holds $\langle f, h\rangle \leq \|f\| \, \|h\|_*$.

**Basics on Duality of Infinite-Dimensional LPs** For succinctness, we will use the following notation. We use $(\tilde{h}, t)$ for the inequality $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})\tilde{h}(\mathbf{x})] + t \leq 0$, where $\tilde{h} \in \mathcal{X}$ and $t \in \mathbb{R}$. Here $\mathcal{X}$ is an appropriate space of functions that in our context will be $L^p(\mathbb{R}^m)$. Let $\mathcal{S}$ be the set of all such tuples that describe the target LP. For the set $\mathcal{S}$, the closed convex cone over $\mathcal{X} \times \mathbb{R}$ is the smallest closed set $\mathcal{S}_+$ satisfying the following: if $A \in \mathcal{S}_+$ and $B \in \mathcal{S}_+$ then $A + B \in \mathcal{S}_+$; and if $A \in \mathcal{S}_+$ then $\lambda A \in \mathcal{S}_+$, for all $\lambda \geq 0$.

In our arguments, we need to prove that there exists a function $g : \mathbb{R}^m \to \mathbb{R}$, such that for any function $h \in L^p(\mathbb{R}^m)$ and at most $(d-1)$-degree polynomial $P : \mathbb{R}^m \to \mathbb{R}$, it holds

$$(*) \begin{cases} -\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})f(\mathbf{x})] + c \leq 0 & 0 < c < \|f\| \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0 & \forall P \in \mathcal{P}^m_{d-1} \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})h(\mathbf{x})] - \|h\|_p \leq 0 & \forall h \in L^p(\mathbb{R}^m) \end{cases}$$

This is in fact an infinite dimensional linear system with respect to the unknown function $g \in (L^1(\mathbb{R}^m))^* = L^\infty(\mathbb{R}^m)$, for $p = 1$ and $g \in L^{p/(p-1)}(\mathbb{R}^m)$ for $1 \leq p < \infty$. We are going to denote $\mathcal{X}$ the metric space $L^p(\mathbb{R}^m)$.

For succinctness, we will use the following notation. We use $(\tilde{h}, t)$ for the inequality $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})\tilde{h}(\mathbf{x})] + t \leq 0$, where $\tilde{h} \in \mathcal{X}$ and $t \in \mathbb{R}$. Moreover, let $\mathcal{S}$ be the set that contains all such tuples that describe the target system. For the set $\mathcal{S}$, the closed convex cone over $\mathcal{X} \times R$ is the smallest closed set $\mathcal{S}_+$ satisfying, if $A \in \mathcal{S}_+$ and $B \in \mathcal{S}_+$ then $A + B \in \mathcal{S}_+$ and, if $A \in \mathcal{S}_+$ then $\lambda A \in \mathcal{S}_+$ for all $\lambda \geq 0$. Note that the $\mathcal{S}_+$ contains the same feasible solutions as $\mathcal{S}$. The set $\mathcal{S} = \{(h, -\|h\|_p) : h \in L^p\} \cup \{(P, 0) : P \in \mathcal{P}^m_{d-1}\} \cup \{(-f, c)\}$.

In the finite-dimensional case, we can always prove the feasibility of an LP by applying the standard Farkas' lemma (aka theorem of the alternative). However, when the system is infinite-dimensional, Farkas' lemma does not hold in general. We are going to use the following result from [Fan68].

**Lemma 11.6.1** (Theorem 1 of [Fan68])**.** *If $\mathcal{X}$ is a locally convex, real separated vector space then, a linear system described by $\mathcal{S}$ for which $\mathcal{S}_+$ is closed is feasible (i.e., there exists a $g \in \mathcal{X}^*$) if and only if $(0,1) \notin \mathcal{S}_+$ and $\mathcal{S}_+$ is closed.*

One direction is trivial, but the other one needs an application of Hahn-Banach theorem which is where the assumption on $\mathcal{X}$ to be a separated space is used.

**Corollary 11.6.2.** *If $\mathcal{X} = L^p$ for $1 \leq p < \infty$ then, the LP described by $\mathcal{S}$ is feasible if only if $(0,1) \notin \mathcal{S}_+$.*

*Proof.* It is not hard to see that the positive cone is defined by

$$S_+ \;=\; \Big\{ \Big(P + h - yf,\; -\|h\|_p + yc - t\Big) : P \in \mathcal{P}^m_{d-1},\; h \in L^p(\mathbb{R}^m),\; y,t \in \mathbb{R},\; y,t \geq 0 \Big\}.$$

If $S_+$ were closed, we could simply apply Lemma 11.6.1. However, it is not immediately clear whether this is the case. Instead, note that Lemma 11.6.1 applies to the closure of $S_+$. Consequently, the linear program (LP) is solvable unless, for every $\epsilon > 0$, there exist $P, h, y, t$ such that

$$\|P + h - yf\|_p < \epsilon, \qquad \Big|-\|h\|_p + yc - t - 1\Big| < \epsilon.$$

Set $h' \;=\; yf - P \;=\; h - (P + h - yf)$. Then $S_+$ contains

$$\Big(P + h' - yf,\; -\|h'\|_p + yc - t\Big) \;=\; \Big(0,\; -\|h'\|_p + yc - t\Big).$$

We observe that

$$\|h'\|_p \;\leq\; \|h\|_p + \|P + h - yf\|_p \;=\; \|h\|_p + \epsilon.$$

Hence

$$-\|h'\|_p + yc - t \;\geq\; -\|h\|_m + yc - y - \epsilon \;\geq\; 1 - 2\epsilon \;\geq\; \tfrac{1}{2} \quad (\text{provided } \epsilon < \tfrac{1}{4}).$$

Because $S_+$ is scale-invariant, this implies $(0, 1) \in S_+$. Therefore, the LP is solvable unless $(0, 1) \in S_+$. □

## 11.7 Statistical Query Lower Bounds for Learning Massart Halfspaces

### 11.7.1 SQ Lower Bound for Learning Halfspaces with Constant-Bounded Massart Noise

The problem of learning homogeneous halfspaces with Massart Noise under the Gaussian distribution is by now well understood. All previous algorithms fit in the SQ framework and show that the SQ complexity of learning halfspaces with Massart noise is polynomial in the dimension $d$, the accuracy $\epsilon$, and the noise rate $\eta$. In this section, we show that if the optimal halfspace $f$ is $\gamma$-biased, i.e., $\mathbf{Pr}[f(\mathbf{x}) = +1] = \gamma$, the SQ complexity of learning $f$ is quasi-polynomial in the bias $\gamma$, that is $d^{\Omega(\log(1/\gamma))}$ SQ queries are required. We prove the following theorem.

**Theorem 11.7.1.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ with standard normal* $\mathbf{x}$-

*marginal that satisfies the $\eta$-Massart noise condition with parameter $\eta \in (0, 1/2)$ with respect to some unknown $(1-\gamma)$-biased optimal halfspace $f(\mathbf{x})$, for some $\gamma > 0$ less than a sufficiently small constant. Any SQ algorithm that, for any such distribution $\mathcal{D}$, learns a hypothesis $h : \mathbb{R}^d \mapsto \{\pm 1\}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$ for $\epsilon \leq \gamma(1-2\eta)$, either requires queries with tolerance at most $d^{-\Omega(\log(\eta/\gamma))}$ or makes at least $2^{d^{\Omega(1)}} d^{-\log(\eta/\gamma)}$ queries.*

**Remark 11.7.2.** We remark that when the bias $\gamma$ is less than $\epsilon$ one can output the constant guess $-1$ and obtain error at most $\epsilon$. Therefore, reasonable values for $\gamma$ are $\Omega(\epsilon)$. In the extreme case where $\gamma = \Theta(\epsilon)$, Theorem 11.7.1 implies a quasi-polynomial SQ lower bound in the accuracy parameter $\epsilon$. However, our result is fine-grained with respect to $\gamma$: we show that this quasi-polynomial dependency on the bias is required across the whole regime of $\gamma$.

The main structural result of the proof of Theorem 11.7.1 is the proposition that follows. We prove that we can construct a distribution $\mathcal{D}$ over labeled pairs $(z, y) \in \mathbb{R} \times \{\pm 1\}$ that satisfies the $\eta$-Massart noise assumption with respect to some (biased) halfspace $f$ and whose low-order moments match the moments of the product distribution of the marginals $\mathcal{D}_z$ and $\mathcal{D}_y$. In other words, we show that we can construct and instance $\mathcal{D}$ whose $z$-marginal is a one dimensional Gaussian distribution and $y$ is uncorrelated with $z^k$ for any $k$ less than a sufficiently small multiple of $\log(\eta/\gamma)$.

**Proposition 11.7.3.** *Fix $\eta \in (0, 1/2)$ and $\gamma > 0$ less than a sufficiently small constant. There exists a distribution $\mathcal{D}$ on $(z, y) \in \mathbb{R} \times \{\pm 1\}$ whose $z$-marginal is the standard normal distribution with the following properties.*

- *$\mathcal{D}$ satisfies the Massart noise condition with parameter $\eta$ with respect to a halfspace $f(z)$ with $\mathbf{Pr}_{z\sim\mathcal{D}_z}[f(z) = 1] = \gamma$.*

- *For any integer $k \leq C\log(\eta/\gamma)$ where $C > 0$ is a sufficiently small constant, it holds $\mathbf{E}_{(z,y)\sim\mathcal{D}}[yz^k] = \mathbf{E}_{y\sim\mathcal{D}_y}[y]\,\mathbf{E}_{z\sim\mathcal{D}_z}[z^k]$.*

*Proof.* Our goal is to construct a distribution $\mathcal{D}$ on $\mathbb{R}\times\{\pm 1\}$ with Gaussian $z$-marginal satisfying the $\eta$-Massart noise condition. Recall that, for any such distribution, we denote $\beta = 1-2\eta$ and $\beta(z) = 1-2\eta(z)$. It holds

$$\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}}[y|x=z] = -f(z)\eta(z) + f(z)(1-\eta(z)) = f(z)(1-2\eta(z)) = f(z)\beta(z)\,.$$

Therefore, we need to prove that there exists a "noise" function $\beta(z):\mathbb{R}\mapsto[\beta,1]$ such that for every zero mean polynomial $p(z)$ of degree at most $k$, it holds $\mathbf{E}_{(z,y)\sim\mathcal{D}}[p(z)y] = \mathbf{E}_{z\sim\mathcal{N}}[p(z)\beta(z)f(z)] = 0$. Recall that by $\mathcal{P}_{k,d}$ we denote the space of polynomials of degree at most $k$ over $\mathbb{R}^d$. In what follows, we will be using polynomials on the subspace of univariate mean-zero polynomials (with respect to the Gaussian measure) which we denote by

$$\mathcal{P}_k^0 \stackrel{\text{def}}{=} \{p\in\mathcal{P}_{k,1} : \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[p(z)] = 0\}\,.$$

Using duality, we first show that such a noise function $\beta(z)$ exists when there exists no mean-zero polynomial $p$ such that the expectation of $(f(z)p(z))^+$ is $1/\beta$ times larger than the expectation of $(f(z)p(z))^-$, where $z^+ \stackrel{\text{def}}{=} \max(0,z)$ and $z^- \stackrel{\text{def}}{=} |\min(0,z)|$. We prove the following lemma.

**Lemma 11.7.4** (Moment-Matching Duality)**.** *Let $f:\mathbb{R}\mapsto\{\pm1\}$ be any one-dimensional Boolean function. Assume that for any polynomial $p\in\mathcal{P}_k^0$, with $p\neq 0$, it holds that*

$$\beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^+] < \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^-]\,.$$

*Then, there exists a function* $\beta(z) : \mathbb{R} \mapsto \mathbb{R}$ *such that* $\mathbf{Pr}_{z\sim\mathcal{N}}[\beta \le \beta(z) \le 1] = 1$ *and for every polynomial* $p \in \mathcal{P}_k^0$ *it holds that* $\mathbf{E}_{z\sim\mathcal{N}}[f(z)p(z)\beta(z)] = 0$.

**Remark 11.7.5.** Even though we used the assumption that $f$ is a one-dimensional function, this statement is true for any $m$-dimensional function as long as $f$ is biased enough. Therefore, using similar techniques as in [DKPZ21], we can embed this $m$-dimensional subspace to $d$ dimensions and get lower bounds for learning more general classes. In fact, for $m = 2$, one can show an SQ lower bound for intersections of two homogeneous halfspaces under constant-bounded Massart noise.

*Proof.* Treating the function $\beta(z)$ as an (infinite dimensional) variable we can formulate the following feasibility linear program.

$$\begin{aligned} \text{Find} \quad & \beta(z) \in L^\infty(\mathbb{R}) \\ \text{such that} \quad & \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)p(z)\beta(z)] = 0 \qquad \forall p \in \mathcal{P}_k^0 \\ & \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}}[\beta \le \beta(z) \le 1] = 1 \end{aligned} \tag{11.4}$$

We claim that the LP (11.4) is equivalent to the following LP:

$$\begin{aligned} \text{Find} \quad & \beta(z) \in L^\infty(\mathbb{R}) \\ \text{such that} \quad & \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)p(z)\beta(z)] = 0 \qquad \forall p \in \mathcal{P}_k^0 \\ & \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\beta(z)h(z)] \le \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[h(z)] \qquad \forall h \in L_+^1(\mathbb{R}) \\ & \beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[T(z)] \le \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[\beta(z)T(z)] \qquad \forall T \in L_+^1(\mathbb{R}) \end{aligned} \tag{11.5}$$

Recall that $L^1(\mathbb{R})$ are all the functions that have bounded $L_1$-norm, and we denote $L_+^1(\mathbb{R})$ to be the positive functions in $L^1(\mathbb{R})$.

**Claim 11.7.6.** *The LP (11.4) is equivalent to the LP (11.5).*

*Proof.* We now show the equivalence between the two formulations. We claim that the second constraint of LP (11.4) is equivalent with the second and the third constraints of LP (11.5). This follows by introducing the "dual variables" $h : \mathbb{R} \to \mathbb{R}$ and $T : \mathbb{R} \to \mathbb{R}$. First, we so that any valid solution $b(z)$ of LP (11.4) is also a valid solution for the LP (11.5). For any valid solution $b(z)$ of LP (11.4), it should hold that $b(z) \leq 1$ and $b(z) \geq \beta$ with probability 1, thus for $h(z) \in L^1_+(\mathbb{R})$ it should hold that $b(z)h(z) \leq h(z)$ and by taking expectation in both sides, we get the third inequality, similarly we get the fourth inequality from $b(z) \geq \beta$. To prove that any valid solution $b(z)$ of LP (11.5) is also a valid solution for the LP (11.4), first we assume, in order to reach a contradiction, that there is a set $A$ with non-zero probability that $b(z) > 1$. Then, by taking $h(z) = \mathbb{1}\{z \in A\}$, the third inequality becomes

$$\mathop{\mathbf{E}}_{z \sim \mathcal{N}}[h(z)] < \mathop{\mathbf{E}}_{z \sim \mathcal{N}}[\beta(z)h(z)] \leq \mathop{\mathbf{E}}_{z \sim \mathcal{N}}[h(z)] \,,$$

which is a contradiction. For the case, we assume, to reach a contradiction, that there is a set $B$ with non-zero probability that $b(z) < \beta$. Similarly, we have $T(z) = \mathbb{1}\{z \in B\}$ and get that $\beta \, \mathbf{E}_{z \sim \mathcal{N}}[T(z)] \leq \mathbf{E}_{z \sim \mathcal{N}}[\beta(z)T(z)] < \beta \, \mathbf{E}_{z \sim \mathcal{N}}[T(z)]$, which is again a contradiction. □

At this point, we would like to use "LP duality" to argue that LP (11.5) is feasible if and only if its "dual LP" is infeasible. While such a statement turns out to be true, it requires some care to prove since we are dealing with infinite LPs (both in number of variables and constraints). More formally, we have that LP (11.5) is feasible if and only if there is no conical combination that yields the contradicting inequality $1 \leq 0$. We

define the "dual LP" to be the following:

$$\begin{aligned} \text{Find} \quad & h \in L^1_+(\mathbb{R}), T \in L^1_+(\mathbb{R}), p \in \mathcal{P}^0_k \\ \text{such that} \quad & f(z)p(z) + h(z) - T(z) = 0 \qquad \forall z \in \mathbb{R} \\ & \beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[T(z)] - \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[h(z)] > 0 \end{aligned} \tag{11.6}$$

The following lemma states that the sufficiently conditions so that LP (11.5) is feasible.

**Lemma 11.7.7.** *If there is no polynomial $p \in \mathcal{P}^0_k$ such that $\beta\, \mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^+] > \mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^-]$ then, the LP (11.5) is feasible if only if LP (11.6) is infeasible.*

*Proof.* First we introduce some notation. We use $(\tilde{h}, c)$ for the inequality $\mathbf{E}_{z\sim\mathcal{N}}[\beta(z)\tilde{h}(z)] + c \leq 0$, where $\tilde{h} \in L^1(\mathbb{R})$ and $c \in \mathbb{R}$. Moreover, let $\mathcal{S}$ be the set that contains all such tuples that describe the target system. For the set $\mathcal{S}$, the closed convex cone over $L^1(\mathbb{R}) \times R$ is the smallest closed set $\mathcal{S}_+$ satisfying, if $A \in \mathcal{S}_+$ and $B \in \mathcal{S}_+$ then $A+B \in \mathcal{S}_+$ and, if $A \in \mathcal{S}_+$ then $\lambda A \in \mathcal{S}_+$ for all $\lambda \geq 0$. Note that the $\mathcal{S}_+$ contains the same feasible solutions as $\mathcal{S}$. In order to prove this, we need the following functional analysis result from [Fan68].

**Fact 11.7.8** (Theorem 1 of [Fan68])**.** *If $\mathcal{X}$ is a locally convex, real separated vector space then, a linear system described by $\mathcal{S}$ is feasible (i.e., there exists a $g \in \mathcal{X}^*$) if and only if $(0,1) \notin \mathcal{S}_+$.*

Our LP is defined by the following inequalities: $(pf, 0)$ for $p \in \mathcal{P}^0_k$, $(h, -\|h\|_1)$ for all $h \in L^1_+(\mathbb{R})$, $(-T, \beta\|T\|_1)$ for all $T \in L^1_+(\mathbb{R})$. By taking the convex cone defined from the above inequalities, we get the following set

$$\mathcal{S}'_+ = \{(pf + h - T, -\|h\|_1 + \beta\|T\|_1) : p \in \mathcal{P}^0_k, h \in L^1_+(\mathbb{R}), T \in L^1_+(\mathbb{R})\} .$$

We have the following lemma.

**Lemma 11.7.9.** *If there is no polynomial $p \in \mathcal{P}_k^0$ such that $\beta\|(pf)^+\|_1 \geq \|(pf)^-\|_1$ then, the LP described by $\mathcal{S}$ is feasible if only if $(0,1) \notin \mathcal{S}'_+$.*

*Proof.* This proof follows from Fact 11.7.8 by showing that $\mathcal{S}'_+$ is closed, i.e., $\mathcal{S}'_+ = \mathcal{S}_+$. Let $F = fp + h - T$ and $\lambda \geq \|h\|_1 - \beta\|T\|_1$, therefore the above LP is equivalent to $\mathbf{E}_{z\sim\mathcal{N}}[\beta(z)F(z)] \leq \lambda$. Let $(p_n, h_n, T_n, \lambda_n)$ be a sequence that satisfies our constraints, i.e., $(fp_n + h_n - T_n, \lambda_n) \in S'_+$, with $F_n = fp_n + h_n - T_n$ converging with respect the $\ell_1$-norm to $F_0$ and $\lambda_n$ converging to $\lambda_0$.

Note that $\lambda_n \geq \|(F_n - (fp_n))^+\|_1 - \beta\|(F_n - (fp_n))^-\|_1$. This holds because $h_n$ and $T_n$ are positive functions and $F_n - fp_n = h_n - T_n$, therefore $h_n = (F_n - (fp_n))^+$ and $T_n = (F_n - (fp_n))^-$. We have that

$$\|(F_n - (fp_n))^+\|_1 - \beta\|(F_n - (fp_n))^-\|_1 \geq -2\|F_n\|_1 + \|(fp_n)^-\|_1 - \beta\|(fp_n)^+\|_1 \,.$$

Note that from our assumption we have that for any polynomial $p \in \mathcal{P}_k^0$, it holds $\|(pf)^-\|_1 - \beta\|(pf)^+\|_1 > 0$. Because this is homogeneous with respect $p$, we can assume that $\|p\|_1 = 1$. Therefore, using the fact that the set $\|p\|_1 = 1$ and $\mathcal{P}_k^0$ is compact, we have that all the limits are inside $\mathcal{P}_k^0$, thus there exists a $c > 0$, such that $\|(pf)^-\|_1 - \beta\|(pf)^+\|_1 > c$. Therefore, it holds $\|(pf)^-\|_1 - \beta\|(pf)^+\|_1 > c\|p\|_1$.

From the above, we have that $\lambda_n + 2\|F_n\|_1 \geq c\|p_n\|_1$ , and because $F_n \to F_0$ and $\lambda_n \to \lambda_0$, that means that $\|p_n\|_1$ is bounded. Since, an $L^1$ ball in $\mathcal{P}_k^0$ is compact, there is a subsequence such that $p_n \to p_0$. By setting $h_0 = (F_0 - (fp_0))^+$ and $T_0 = (F_0 - (fp_0))^-$, we find appropriate $p_0, h_0, T_0$ that give the appropriate limit points. Therefore, the set $\mathcal{S}'_+$ is closed and the lemma follows from Fact 11.7.8. □

The proof of Lemma 11.7.7, follows from Lemma 11.7.9, by noting that $(0,1) \notin \mathcal{S}'_+$ is equivalent to the infeasibility of LP (11.6). □

Note that if LP (11.6) is feasible for some functions $(h, T, p)$, then it should hold that $h(z) = (f(z)p(z))^-$ and $T(z) = (f(z)p(z))^+$. Therefore, we claim that the LP (11.6) is equivalent to the following LP.

$$\begin{aligned} \text{Find} \quad & p \in \mathcal{P}_k^0 \\ \text{such that} \quad & \beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^+] > \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^-] \end{aligned} \tag{11.7}$$

Observe that if LP (11.7) is infeasible then from Lemma 11.7.7, we get that LP (11.5) is feasible, therefore the proof of Lemma 11.7.4 follows. □

**Lemma 11.7.10.** *Let $f : \mathbb{R} \mapsto \{\pm 1\}$ be any one-dimensional Boolean function, $\beta \in (0,1)$ and $k \in \mathbb{Z}_+$. If $\mathbf{Pr}_{z\sim\mathcal{N}}[f(z) = 1] \leq 2^{-Ck}(1-\beta)$ for some sufficiently large constant $C > 0$, then for any polynomial $p \in \mathcal{P}_k^0$, it holds $\beta\, \mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^+] \leq \mathbf{E}_{z\sim\mathcal{N}}[(f(z)p(z))^-]$.*

*Proof.* First, notice that, since $\mathbf{E}_{z\sim\mathcal{N}}[p(z)] = 0$, it holds that $\mathbf{E}_{z\sim\mathcal{N}}[p(z)^+] = \mathbf{E}_{z\sim\mathcal{N}}[p(z)^-]$. Moreover, we have that for any $z \in \mathbb{R}$ we either have $f(z) = +1$ or $f(z) = -1$ and therefore, it holds that $f(z)^+ + f(z)^- = 1$. Thus, $\mathbf{E}_{z\sim\mathcal{N}}[p(z)^+ f(z)^+] + \mathbf{E}_{z\sim\mathcal{N}}[p(z)^+ f(z)^-] = \mathbf{E}_{z\sim\mathcal{N}}[p(z)^- f(z)^+] + \mathbf{E}_{z\sim\mathcal{N}}[p(z)^- f(z)^-]$. We have that

$$\begin{aligned} \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^+] &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+ p(z)^+] + \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^- p(z)^-]\,, \\ \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[(f(z)p(z))^-] &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+ p(z)^-] + \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^- p(z)^+]\,. \end{aligned}$$

Therefore, LP (11.7) implies the following inequality

$$2\beta \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+p(z)^+] > (1+\beta)\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+p(z)^-] + (1-\beta)\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^-p(z)^+]$$
$$\geq (1-\beta)\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^-p(z)^+]\,, \tag{11.8}$$

where we used the fact that $\mathbf{E}_{z\sim\mathcal{N}}[f(z)^+p(z)^-] \geq 0$. Notice that $\mathbf{E}_{z\sim\mathcal{N}}[|p(z)|] = 2\,\mathbf{E}_{z\sim\mathcal{N}}[p(z)^+] = 2\,\mathbf{E}_{z\sim\mathcal{N}}[p(z)^-]$, because $p(z)$ is a zero mean polynomial. Moreover, using $f(z)^- = 1 - f(z)^+$, it holds $\mathbf{E}_{z\sim\mathcal{N}}[f(z)^-p(z)^+] = \mathbf{E}_{z\sim\mathcal{N}}[|p(z)|]/2 - \mathbf{E}_{z\sim\mathcal{N}}[f(z)^+p(z)^+]$. Thus, by substituting this in the Equation (11.8), we get

$$2\frac{1+\beta}{1-\beta}\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+p(z)^+] > \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[|p(z)|]\,,$$

or $4\,\mathbf{E}_{z\sim\mathcal{N}}[f(z)^+p(z)^+]/(1-\beta) > \mathbf{E}_{z\sim\mathcal{N}}[|p(z)|]$, where we used that $\beta < 1$. From Cauchy–Schwarz inequality, we have

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+p(z)^+] \leq \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+|p(z)|] \leq (\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+])^{1/2}(\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[|p(z)|])^{1/2}\,.$$

Using Cauchy–Schwarz, it holds that $(\mathbf{E}_{z\sim\mathcal{N}}[p(z)^2])^{1/2} \leq (\mathbf{E}_{z\sim\mathcal{N}}[|p(z)|])^{1/3}(\mathbf{E}_{z\sim\mathcal{N}}[p(z)^4])^{1/6}$ and from Lemma 2.3.14, we have that $(\mathbf{E}_{z\sim\mathcal{N}}[p(z)^4])^{1/4} \leq 3^{k/2}(\mathbf{E}_{z\sim\mathcal{N}}[p(z)^2])^{1/2}$. Putting everything together, we have that $(\mathbf{E}_{z\sim\mathcal{N}}[p(z)^2])^{1/2} \leq 2^{Ck}\,\mathbf{E}_{z\sim\mathcal{N}}[|p(z)|]$, for $C$ some large enough positive constant, thus

$$\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[|p(z)|] \leq \frac{4}{1-\beta}\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[f(z)^+p(z)^+] \leq \frac{4}{1-\beta}(\mathbf{Pr}[f(z)=1])^{1/2}2^{Ck}\mathop{\mathbf{E}}_{z\sim\mathcal{N}}[|p(z)|]\,,$$

which is a contradiction if $\mathbf{Pr}[f(z)=1] \leq 2^{-Ck}(1-\beta)$ for large enough positive constant

$C$. Thus, the LP (11.7) is infeasible and thus, the LP (11.4) is feasible. □

Using Lemma 11.7.4 and Lemma 11.7.10, we get that if $\gamma = \mathbf{Pr}_{z\sim\mathcal{N}}[f(z) = 1] \leq 2^{-Ck}(1-\beta)$ for some large enough constant $C > 0$, then there exists a function $\beta : \mathbb{R} \mapsto \mathbb{R}$ such that for every polynomial $p \in \mathcal{P}_k^0$ it holds $\mathbf{E}_{z\sim\mathcal{N}}[f(z)p(z)\beta(z)] = 0$, thus, for any $m \leq k$, we have

$$\mathop{\mathbf{E}}_{(z,y)\sim\mathcal{D}}[yz^m] = \mathop{\mathbf{E}}_{(z,y)\sim\mathcal{D}}[y(z^m - \mathop{\mathbf{E}}_{z\sim\mathcal{D}_z}[z^m])] + \mathop{\mathbf{E}}_{(z,y)\sim\mathcal{D}_y}[y] \mathop{\mathbf{E}}_{z\sim\mathcal{D}_z}[z^m] = \mathop{\mathbf{E}}_{(z,y)\sim\mathcal{D}_y}[y] \mathop{\mathbf{E}}_{z\sim\mathcal{D}_z}[z^m] .$$

Moreover, using the fact that $\beta = 1 - 2\eta$, we get that the degree $k$ is less than a sufficiently small multiply of $\log(\eta/\gamma)$, which completes the proof of Proposition 11.7.3. □

**Proposition 11.7.11** (SQ Complexity of Hypothesis Testing)**.** *Fix $\eta \in (0, 1/2)$ and $\gamma > 0$ less than a sufficiently small constant. There exist:*

- *a family of distributions $\mathfrak{D}$ such that every $\mathcal{D} \in \mathfrak{D}$ is a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$, its $\mathbf{x}$-marginal is the standard normal distribution, and $\mathcal{D}$ satisfies the $\eta$-Massart noise condition with respect to some $\gamma$-biased halfspace, and*
- *a reference distribution $\mathcal{R}$ over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$, whose $\mathbf{x}$-marginal is the standard normal and $y$ is independent of $\mathbf{x}$,*

*such that any SQ algorithm that decides whether the input distribution belongs to $\mathfrak{D}$ or is equal to the reference $\mathcal{R}$ either requires queries with tolerance at most $d^{-\Omega(\log(\eta/\gamma))}/\sqrt{\gamma}$ or makes at least $2^{d^{\Omega(1)}} d^{-\log(\eta/\gamma)}$ queries.*

Proposition 11.7.11 follows from the lemma below showing that we can construct a family of distributions with small pairwise correlation

**Lemma 11.7.12** (Correlated Family of Distributions)**.** *Fix $t \in \mathbb{R}$ such that $|t|$ is larger than some absolute constant. Define $\gamma(t) \stackrel{\text{def}}{=} \mathbf{Pr}_{z\sim\mathcal{N}}[z \geq |t|]$. There exist:*

- *a set $\mathfrak{D}_t$ of $2^{d^{\Omega(1)}}$ distributions on $\mathbb{R}^d \times \{\pm 1\}$, such that every $\mathcal{D} \in \mathfrak{D}_t$ is a distribution over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$, its $\mathbf{x}$-marginal is the standard normal distribution, and $\mathcal{D}$ satisfies the $\eta$-Massart condition with $\eta \in (0, \frac{1}{2})$ with respect to a halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{v} \cdot \mathbf{x} + t)$ where $\mathbf{v}$ is a unit vector in $\mathbb{R}^d$. Moreover, all $\mathcal{D} \in \mathfrak{D}_t$ have the same $y$-marginal,*
- *a reference distribution $\mathcal{R}_t$ in $\mathbb{R}^d \times \{\pm 1\}$, where for $(\mathbf{x}, y) \sim \mathcal{R}_t$ we have that $\mathbf{x}$ is distributed according to the standard normal $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $y$ is independent of $\mathbf{x}$, and the distribution of $y$ is equal with the $y$-marginal of any $\mathcal{D} \in \mathfrak{D}_t$.*

*Moreover, the set $\mathfrak{D}_t$ is $(d^{-\Omega(\log(\eta/\gamma(t)))}/\gamma(t), 4/\gamma(t))$-correlated with respect to the reference distribution $\mathcal{R}_t$.*

*Proof.* From Proposition 11.7.3, we know that for every $t$, with $|t|$ larger than a sufficiently large constant, there exists a distribution $\mathcal{D}_t$ on $\mathbb{R} \times \{\pm 1\}$ whose $z$-marginal is a standard normal, $\mathcal{D}$ satisfies the $\eta$-Massart noise distribution with respect to a $\gamma$-biased halfspace $f : \mathbb{R} \mapsto \{\pm 1\}$, with $f(z) = \mathrm{sign}(z + t)$, and for every $k = \Theta(\log(\eta/\gamma))$ it holds that $\mathbf{E}_{(z,y)\sim\mathcal{D}_t}[yz^k] = \mathbf{E}_{y\sim\mathcal{D}_{ty}}[y]\,\mathbf{E}_{z\sim\mathcal{N}}[z^k]$. Let $\beta_t(z)$ be the noise function corresponding to $\mathcal{D}_t$ and $\phi(z)$ be the density function of the single dimensional standard normal distribution. We define the following densities on $\mathbb{R}$: $A_t(z) = (1 + \beta(z)f(z))\phi(z)/(1 + c)$ and $B_t(z) = (1 - \beta(z)f(z))\phi(z)/(1 - c)$ where $c = \mathbf{E}_{z\sim\mathcal{N}}[\beta(z)f(z)] = \mathbf{E}_{y\sim\mathcal{D}_{ty}}[y]$. It holds

that

$$\begin{aligned}\mathop{\mathbf{E}}_{z\sim A_t}[z^k] &= \frac{\int z^k\phi(z) + z^k\beta(z)f(z)\mathrm{d}z}{1+c} = \frac{\mathbf{E}_{z\sim\mathcal{N}}[z^k] + \mathbf{E}_{(z,y)\sim\mathcal{D}_t}[z^k y]}{1+c} \\ &= \frac{\mathbf{E}_{z\sim\mathcal{N}}[z^k] + \mathbf{E}_{z\sim\mathcal{D}_{tz}}[z^k]\,\mathbf{E}_{y\sim\mathcal{D}_{ty}}[y]}{1+\mathbf{E}_{y\sim\mathcal{D}_{ty}}[y]} = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}[z^k]\,.\end{aligned}$$

Similarly, we have that $\mathbf{E}_{z\sim B_t}[z^k] = \mathbf{E}_{z\sim\mathcal{N}}[z^k]$. Therefore, the distributions $A_t$ and $B_t$ match the first $\Theta(\log(\eta/\gamma))$ moments with $\mathcal{N}$.

Moreover, we have that

$$\chi^2(A_t,\mathcal{N}) = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\frac{1+\beta(z)f(z)}{1+c}\right)^2 - 1\right] \le \frac{4}{(1+c)^2}\,,$$

and

$$\chi^2(B_t,\mathcal{N}) = \mathop{\mathbf{E}}_{z\sim\mathcal{N}}\left[\left(\frac{1-\beta(z)f(z)}{1+c}\right)^2 - 1\right] \le \frac{4}{(1-c)^2}\,,$$

which means that $\max((1+c)\chi^2(A_t,\mathcal{N}),(1-c)\chi^2(B_t,\mathcal{N})) \lesssim 1/\gamma$. Let S be as in Fact 11.2.6. Choose $p=(1+c)/2$ and consider the mixture distributions $\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}$, for $\mathbf{v}\in S$. We set the hard family of distributions $\mathfrak{D}_t = \{\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p} : \mathbf{v}\in S\}$. First, we show that $\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}$ corresponds to a distribution that satisfies the $\eta$-Massart noise condition. We have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}}[y|\mathbf{x}] &= \frac{(1+c)}{2}(1+f(\mathbf{x}\cdot\mathbf{v})\beta(\mathbf{x}\cdot\mathbf{v}))\frac{\phi(\mathbf{x})}{1+c} - \frac{(1-c)}{2}(1-f(\mathbf{x}\cdot\mathbf{v})\beta(\mathbf{x}\cdot\mathbf{v}))\frac{\phi(\mathbf{x})}{1-c} \\ &= f(\mathbf{x}\cdot\mathbf{v})\beta(\mathbf{x}\cdot\mathbf{v})\,,\end{aligned}$$

thus, indeed the distribution $\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}$ satisfies the $\eta$-Massart noise condition, because

from Proposition 11.7.3, we know that $\beta = 1 - 2\eta \leq \beta(z) \leq 1$ almost surely for all $z \in \mathbb{R}$.

Let $\mathcal{R}_t$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that if $(\mathbf{x}, y) \sim \mathcal{R}_t$ then $\mathbf{x} \sim \mathcal{N}_d$, $y$ is independent of $\mathbf{x}$, and $y = 1$ with probability $p$ and $y = -1$ otherwise. We need to show that for $\mathbf{u}, \mathbf{v} \in S$ we have that $|\chi_{\mathcal{R}_t}(\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}, \mathbf{P}_{\mathbf{u}}^{A_t,B_t,p})|$ is small. Since $\mathcal{R}_t, \mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}$, and $\mathbf{P}_{\mathbf{u}}^{A_t,B_t,p}$ all assign $y = 1$ with probability $p$, we have that

$$\begin{aligned}\chi_{\mathcal{R}_t}(\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}, \mathbf{P}_{\mathbf{u}}^{A_t,B_t,p}) &= p\, \chi_{\mathcal{R}_t|y=1}\left((\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p} \mid y = 1), (\mathbf{P}_{\mathbf{u}}^{A_t,B_t,p} \mid y = 1)\right) + \\ &\qquad (1-p)\, \chi_{\mathcal{R}_t|y=-1}\left((\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p} \mid y = -1), (\mathbf{P}_{\mathbf{u}}^{A_t,B_t,p} \mid y = -1)\right) \\ &= p\, \chi_{\mathcal{N}_d}(\mathbf{P}_{\mathbf{v}}^{A_t}, \mathbf{P}_{\mathbf{u}}^{A_t}) + (1-p)\, \chi_{\mathcal{N}_d}(\mathbf{P}_{\mathbf{v}}^{B_t}, \mathbf{P}_{\mathbf{u}}^{B_t}).\end{aligned}$$

By Fact 1.4.10, it follows that

$$\begin{aligned}\chi_{\mathcal{R}_t}(\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}, \mathbf{P}_{\mathbf{u}}^{A_t,B_t,p}) &\leq d^{-\Omega(\log(\eta/\gamma))}(p\,\chi^2(A_t, N(0,1)) + (1-p)\chi^2(B_t, N(0,1))) \\ &= d^{-\Omega(\log(\eta/\gamma))}/\gamma\ .\end{aligned}$$

A similar computation shows that

$$\chi_{\mathcal{R}_t}(\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}, \mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}) = \chi^2(\mathbf{P}_{\mathbf{v}}^{A_t,B_t,p}, \mathcal{R}_t) \leq p\,\chi^2(A_t, N(0,1)) + (1-p)\chi^2(B_t, N(0,1) \leq 4/\gamma\ .$$

Thus, the set $\mathfrak{D}_t$ is $(d^{-\Omega(\log(\eta/\gamma))}/\gamma, 4/\gamma)$-correlated with respect the reference distribution $\mathcal{R}_t$. □

*Proof of Proposition 11.7.11.* Fix a $\gamma > 0$ less than a sufficiently small constant, and let $t \in \mathbb{R}$ such that $\mathbf{Pr}_{z\sim\mathcal{N}}[z \geq |t|] = \gamma$ Moreover, because $|\mathfrak{D}_t| = 2^{d^{\Omega(1)}}$ and the set $\mathfrak{D}_t$ is $(d^{-\Omega(\log(\eta/\gamma))}/\gamma, 4/\gamma)$-correlated with respect $\mathcal{R}_t$ we have that $\mathrm{SD}(\mathcal{B}, d^{-\Omega(\log(\eta/\gamma))}/\gamma, 4/\gamma) =$

$2^{d^{\Omega(1)}}$, thus an application of Lemma 1.4.9 completes the proof. □

*Proof of Theorem 11.7.1.* Let $\mathfrak{D}$ and $\mathcal{R}$ be as in Proposition 11.7.11. Note that from the construction, for any $\mathcal{D} \in \mathfrak{D}$, it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{R}}[y = i] = \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[y = i] = p_i$, for $i \in \{\pm 1\}$ Let $\mathcal{A}$ an algorithm that outputs a hypothesis $h$ with respect a distribution $\mathcal{D}$ that satisfies the $\eta$-Massart Noise condition such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, where OPT is the error achieved by the best classifier. Moreover, from for any classifier $h'$ it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{R}}[h'(\mathbf{x}) \neq y] \geq \min_{i\in\{\pm1\}} p_i$. Thus, it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{R}}[h'(\mathbf{x}) \neq y] - \mathrm{OPT} \geq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})f(\mathbf{x})] \geq (1-2\eta)\gamma$. Thus, algorithm $\mathcal{A}$ for $\epsilon \leq \gamma\beta$ would solve the decision problem $\mathcal{B}(\mathfrak{D}, \mathcal{R})$ (with one additional query $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})y]$ up to accuracy $\gamma\beta/2$) and from Proposition 11.7.11 the result follows. □

### 11.7.2 SQ Lower Bound for Learning Homogeneous Halfspaces with General Massart Noise

Learning homogeneous halfspaces under $\eta$-Massart noise for any constant $\eta < 1/2$ is known to be solvable with polynomially many statistical queries [DKTZ20a]. When $\eta = 1/2$, we show that any SQ algorithm for homogeneous Massart halfspaces requires $d^{\Omega(\log(1/\epsilon))}$ queries. For comparison, we recall that the SQ complexity of learning halfspaces under adversarial noise is $d^{\Omega(1/\epsilon^2)}$ [DKPZ21].

Our formal result is the following theorem.

**Theorem 11.7.13** (1/2-Massart Noise SQ-Lower Bound)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the Massart noise condition for $\eta = 1/2$, for homogeneous halfspaces, where the $\mathbf{x}$-marginal is distributed according to the standard normal. Any SQ algorithm that, for any such distribution $\mathcal{D}$, finds a hypothesis $h$ such that*

$\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$ *either requires queries with tolerance at most* $d^{-\Omega(\log(1/\epsilon))}$ *or makes at least* $2^{d^{\Omega(1)}}$ *queries.*

To prove the theorem above, we are going to use a "reduction" to the problem of learning general halfspaces with constant-bounded Massart noise.

**Proposition 11.7.14** (SQ Complexity of Hypothesis Testing)**.** *Let* $\epsilon > 0$ *less than a sufficiently small constant. There exist:*

- *a family of distributions* $\mathfrak{D}_B$ *such that every* $\mathcal{D} \in \mathfrak{D}$ *is a distribution over* $(\mathbf{x}', y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$*, its* $\mathbf{x}'$*-marginal is the standard normal distribution, and* $\mathcal{D}$ *satisfies the Massart noise condition for* $\eta = 1/2$*, with respect to some unbiased halfspace, and*
- *a reference distribution* $\mathcal{R}$ *over* $((\mathbf{x}, z), y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$*, whose* $(\mathbf{x}, z)$*-marginal is the standard normal and* $y$ *depends only on* $z$*,*

*such that any SQ algorithm that decides whether the input distribution belongs to* $\mathfrak{D}$ *or is equal to the reference* $\mathcal{R}$ *either requires queries with tolerance at most* $d^{-\Omega(\log(1/\epsilon))}$ *or makes at least* $2^{d^{\Omega(1)}} d^{-\log(1/\epsilon)}$ *queries.*

*Proof.* Proposition 11.7.14 follows from the lemma below showing that we can construct a family of distributions with small pairwise correlation

**Lemma 11.7.15** (Correlated Family of Distributions)**.** *Let* $\epsilon > 0$ *less than a sufficiently small constant. There exist:*

- *a set* $\mathfrak{D}_B$ *of* $2^{d^{\Omega(1)}}$ *distributions on* $\mathbb{R}^{d+1} \times \{\pm 1\}$*, such that every* $\mathcal{D} \in \mathfrak{D}_B$ *is a distribution over* $(\mathbf{x}', y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$*, its* $\mathbf{x}'$*-marginal is the standard normal distribution, and* $\mathcal{D}$ *satisfies the Massart noise condition for* $\eta = 1/2$ *with respect*

*to a halfspace* $f(\mathbf{x}') = \text{sign}((\mathbf{v}, 1) \cdot \mathbf{x}')$ *where* $\mathbf{v}$ *is a unit vector in* $\mathbb{R}^d$*. Moreover, all* $\mathcal{D} \in \mathfrak{D}_B$ *have the same* $y$*-marginal.*

- *A reference distribution* $\mathcal{R}$ *in* $\mathbb{R}^{d+1} \times \{\pm 1\}$*, where for* $((\mathbf{x}, z), y) \sim \mathcal{R}$ *we have that* $\mathbf{x}'$ *is a standard Gaussian* $\mathcal{N}(\mathbf{0}, \mathbf{I})$*,* $y$ *depends only on* $z$*, and for any function* $h : \mathbb{R}^{d+1} \mapsto \{\pm 1\}$ *and any* $\mathcal{D} \in \mathfrak{D}_B$*, it holds that* $\mathbf{Pr}_{(\mathbf{x}', y) \sim \mathcal{R}}[h(\mathbf{x}') \neq y] - \min_{f \in \mathcal{C}} \mathbf{Pr}_{(\mathbf{x}', y) \sim \mathcal{D}}[f(\mathbf{x}') \neq y] \geq 2\epsilon$*.*

*Moreover, the set* $\mathfrak{D}_B$ *is* $(d^{-\Omega(\log(1/\epsilon))}, 16\epsilon)$*-correlated with respect to the reference distribution* $\mathcal{R}$*.*

*Proof.* Fix $t_0 > 0$ and $\zeta > 0$ such that $\mathbf{Pr}_{z \sim \mathcal{N}}[z \geq t_0] = \gamma$ for $\gamma = 4\sqrt{\epsilon}$ and $\mathbf{Pr}_{z \sim \mathcal{N}}[t_0 + \zeta > z \geq t_0] = \gamma/2$. We are going to construct a new set of distributions $\mathfrak{D}_B$ as follows:

1. For $\eta = 1/4$ and some threshold $t$ denote $(\mathfrak{D}_t, \mathcal{R}_t)$ be the family of distributions and their corresponding reference distributions from Lemma 11.7.12. Recall that the distributions in the family $\mathfrak{D}_t$ are indexed by a set of unit vectors $S$, i.e., $\mathfrak{D}_t = \{\mathcal{D}_{\mathbf{u},t} : \mathbf{u} \in S\}$. Every $\mathcal{D}_{\mathbf{u},t} \in \mathfrak{D}_t$ is a distribution in $\mathbb{R}^d$ that satisfies the $\eta$-Massart noise condition with $\eta = 1/4$ and with respect to the halfspace $f(\mathbf{x}) = \text{sign}(\mathbf{u} \cdot \mathbf{x} + t)$.

2. Fix some direction $\mathbf{u} \in S$. We define a new distribution $\mathcal{D}'_{\mathbf{u}}$ on $((\mathbf{x}, z), y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$ where the "extra" coordinate $z$ is drawn from the standard normal distribution $\mathcal{N}(0, 1)$. When $z$ falls inside some thin interval $[t_0, t_0 + \zeta]$ we sample $(\mathbf{x}, y)$ from the 1/4-Massart noise distribution $\mathcal{D}_{\mathbf{u},z}$. This corresponds to the blue/red region in Figure 2.3. When $z$ falls outside $[t_0, t_0 + \zeta]$, we draw $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and set $y$ to be $\pm 1$ with probability 1/2, independently of $\mathbf{x}$. This "high-noise"

area corresponds to the gray area of Figure 2.3. More formally, we define

$$\mathcal{D}'_{\mathbf{u}}((\mathbf{x},z),y) = \begin{cases} \mathcal{D}'_{\mathbf{u},z}(\mathbf{x},y)\phi(z) & \text{if } z \in [t_0, t_0+\zeta] \\ \frac{1}{2}\phi_{d+1}((\mathbf{x},z)) & \text{otherwise .} \end{cases}$$

Let $\mathfrak{D}_B$ be the set of these distributions.

3. We define a reference distribution $\mathcal{R}$ on $((\mathbf{x},z),y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$ similarly. The "extra" coordinate $z$ is again drawn from the standard normal distribution $\mathcal{N}(0,1)$. When $z$ falls inside some thin interval $[t_0, t_0+\zeta]$, i.e., the blue/red region of Figure 2.3 we sample $(\mathbf{x},y)$ from the reference distribution $\mathcal{R}_t$. When $z$ falls outside $[t_0, t_0+\zeta]$, i.e., in the gray area of Figure 2.3, we draw $\mathbf{x} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and set $y$ to be $\pm 1$ with probability $1/2$, independently of $\mathbf{x}$. We have

$$\mathcal{R}((\mathbf{x},z),y) = \begin{cases} \mathcal{R}_z(\mathbf{x},y)\phi(z) & \text{if } z \in [t_0, t_0+\zeta] \\ \frac{1}{2}\phi_{d+1}((\mathbf{x},z)) & \text{otherwise .} \end{cases}$$

Our first step is to prove that any "hidden-direction" $(d+1)$-dimensional distribution $\mathcal{D}'_{\mathbf{u}}$ that we create out of the Massart-noise instances $\mathcal{D}_{\mathbf{u},t}$ satisfies the $1/2$-Massart noise condition with a homogeneous optimal halfspace. We show the following claim.

**Claim 11.7.16.** *For the distribution $\mathcal{D}'_{\mathbf{u}}$, the optimal hypothesis is $f'(\mathbf{x}') = \operatorname{sign}(\mathbf{x}' \cdot (\mathbf{u},1))$, thus $\mathcal{D}'_{\mathbf{u}}$ satisfies the Massart noise condition for $\eta = 1/2$ with respect to a homogeneous $(d+1)$-dimensional halfspace.*

*Proof.* In what $A$ be the event that $z \in [t_0, t_0+\zeta]$. Assume in order to reach to a contradiction that $f'(\mathbf{x}')$ is not the optimal hypothesis. Let $h(\mathbf{x}')$ to be the optimal

hypothesis for the distribution $\mathcal{D}_{\mathbf{u}'}$. First, observe that for any hypothesis $h'(\mathbf{x}')$ it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[h'(\mathbf{x}') \neq y, A^c] = \mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[A^c]/2 \; .$$

Thus, we need only to consider the error inside the $A$. For the function $f'(\mathbf{x}')$, we have that

$$\begin{aligned}\mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[f'(\mathbf{x}') \neq y, A] &= \int_{[t_0,t_0+\zeta]} \mathop{\mathbf{Pr}}_{((\mathbf{x},t),y)\sim\mathcal{D}'_{\mathbf{u}}}[f'((\mathbf{x},t)) \neq y | t = z]\phi(z)\mathrm{d}z \\ &= \int_{[t_0,t_0+\zeta]} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathbf{u},z}}[\mathrm{sign}(\mathbf{u}\cdot\mathbf{x}+z) \neq y]\phi(z)\mathrm{d}z \; . \end{aligned} \tag{11.9}$$

Similarly, for any other hypothesis $h'(\mathbf{x}')$, we have that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[h'(\mathbf{x}') \neq y, A] = \int_{[t_0,t_0+\zeta]} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathbf{u},z}}[h'((\mathbf{x},z)) \neq y]\phi(z)\mathrm{d}z \; . \tag{11.10}$$

Moreover, for the $h'(\mathbf{x}')$ it should hold that $\mathbf{Pr}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[f'(\mathbf{x}') \neq y, A] - \mathbf{Pr}_{(\mathbf{x}',y)\sim\mathcal{D}'_{\mathbf{u}}}[h(\mathbf{x}') \neq y, A] > 0$, thus combining with eq. (11.9) and eq. (11.10), we have that there exists $z \in [t_0, t_0+\zeta]$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathbf{u},z}}[\mathrm{sign}(\mathbf{u}\cdot\mathbf{x}+z) \neq y] > \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathbf{u},z}}[h'((\mathbf{x},z)) \neq y]$, which is a contradiction, because for each $\mathcal{D}_{\mathbf{u},t}$ the optimal classifier is $f(\mathbf{x}) = \mathrm{sign}(\mathbf{u}\cdot\mathbf{x}+t)$. □

We next have to show that the family of "hidden-direction" distributions $\mathfrak{D}_B$ is pairwise correlated. We have the following claim.

**Claim 11.7.17.** *For every $\mathcal{D} \in \mathfrak{D}_B$ it holds $\chi_{\mathcal{R}}(\mathcal{D},\mathcal{D}) \leq 4$ and for every distinct $\mathcal{D}_1, \mathcal{D}_2 \in \mathfrak{D}_B$ it holds $\chi_{\mathcal{R}}(\mathcal{D}_1,\mathcal{D}_2) \leq d^{-\Omega(\log(1/\gamma))}$. In other words, the family $\mathfrak{D}_B$ is $(d^{-\Omega(\log(1/\gamma))}, 4)$-correlated with respect to the reference distribution $\mathcal{R}$.*

*Proof.* Note that by the construction of the distributions $\mathcal{D}_1, \mathcal{D}_2$, for a point $\mathbf{x}' =$

$(\mathbf{x}, z) \in \mathbb{R}^{d+1}$, such that $z \notin [t_0, t_0 + \zeta]$ the conditional distributions $\mathcal{D}_1 \mid z$, $\mathcal{D}_2 \mid z$ and $\mathcal{R} \mid z$ coincide. Therefore, since $\mathcal{D}_1, \mathcal{D}_2, \mathcal{R}$ give the same distribution on $z$, for $z \notin [t_0, t_0 + \zeta]$ it holds $\chi_{\mathcal{R}|z}((\mathcal{D}_1 \mid z), (\mathcal{D}_2 \mid z)) = 0$, for $z \notin [t_0, t_0 + \zeta]$. Thus, to compute the correlation of $\mathcal{D}_1, \mathcal{D}_2$ it suffices to consider $z \in [t_0, t_0 + \zeta]$, i.e.,

$$\begin{aligned}\chi_{\mathcal{R}}(\mathcal{D}_1, \mathcal{D}_2) &= \mathop{\mathbf{E}}_{z \sim \mathcal{N}}[\chi_{\mathcal{R}_z}((\mathcal{D}_1 \mid z), (\mathcal{D}_2 \mid z))] \\ &= \int_{t_0}^{t_0+\zeta} \chi_{\mathcal{R}_z}((\mathcal{D}_1 \mid z), (\mathcal{D}_2 \mid z))\phi(z)\mathrm{d}z \ ,\end{aligned}$$

where $(\mathcal{D}_1 \mid z)$ (resp. $(\mathcal{D}_2 \mid z)$) is the pdf $\mathcal{D}_1$ (resp. $\mathcal{D}_2$) conditioned on $z$. From Lemma 11.7.12, it holds that

$$\chi_{\mathcal{R}}(\mathcal{D}_1, \mathcal{D}_2) \leq \frac{d^{-\Omega(\log(1/\gamma))}}{\gamma} \int_{t_0}^{t_0+\zeta} \phi(z)\mathrm{d}z = d^{-\Omega(\log(1/\gamma))} \ ,$$

where we used that $\mathbf{Pr}_{z \sim \mathcal{N}}[t_0 + \zeta > z \geq t_0] = \gamma/2$. The second part follows similarly. □

Moreover, if $r$ is the best hypothesis for $\mathcal{R}$ and $f$ for $\mathcal{D}'_{\mathbf{u}}$, then from Lemma 11.7.12, we have

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}', y) \sim \mathcal{R}}[r(\mathbf{x}') \neq y] - \mathop{\mathbf{Pr}}_{(\mathbf{x}', y) \sim \mathcal{D}'_{\mathbf{u}}}[f(\mathbf{x}') \neq y] \geq \frac{1}{3}\gamma \int_{t_0}^{t_0+\zeta} \phi(z)\mathrm{d}z \geq \frac{\gamma^2}{6} \ ,$$

where in the last inequality we used that $t_0, \zeta$ are chosen such that $\mathbf{Pr}_{z \sim \mathcal{N}}[t_0 + \zeta > z \geq t_0] = \gamma/2$, by substituting $\gamma = 4\sqrt{\epsilon}$ the result follows. □

To prove Proposition 11.7.14, from Lemma 11.7.15, we have that the set $\mathfrak{D}_B$ is $(d^{-\Omega(\log(1/\epsilon))}, 16)$-correlated with respect $\mathcal{R}$. Thus, we have that $\mathrm{SD}(\mathcal{B}, d^{-\Omega(\log(1/\epsilon))}, 16) = 2^{d^{\Omega(1)}}$ and an application of Lemma 1.4.9 completes the proof. □

*Proof of Theorem 11.7.13.* Fix $\epsilon > 0$ less than a sufficiently small constant. Let $\mathfrak{D}_B$ and $\mathcal{R}$ as in Lemma 11.7.15. Let $\mathcal{A}$ be an algorithm that given $\epsilon' > 0$ and $\mathcal{D}$ that satisfies the 1/2-Massart Noise Condition with respect the halfspace $f'(\mathbf{x}')$ computes a hypothesis $h$ such that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}}[h(\mathbf{x}') \neq y] \leq \mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}}[f(\mathbf{x}') \neq y] + \epsilon' \; .$$

We show that $\mathcal{A}$ can solve the Decision Problem $\mathcal{B}(\mathfrak{D}_B, \mathcal{R})$. Let $\mathcal{D} \in \mathfrak{D}_B$ with optimal halfspace $f(\mathbf{x})$, then from Lemma 11.7.15, we have that for any hypothesis $h$ it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{R}}[h(\mathbf{x}') \neq y] - \mathop{\mathbf{Pr}}_{(\mathbf{x}',y)\sim\mathcal{D}}[f(\mathbf{x}') \neq y] \geq 2\epsilon \; .$$

The algorithm $\mathcal{A}$ for $\epsilon' \leq 2\epsilon$ would solve the decision problem $\mathcal{B}(\mathfrak{D}_B, \mathcal{R})$ (with one additional query $\mathbf{E}_{(\mathbf{x}',y)\sim\mathcal{D}}[h(\mathbf{x}')y]$ up to accuracy $\epsilon$), thus from Proposition 11.7.14, we get our result. □

# Chapter 12

# SQ Lower Bounds for Learning Multi-Index Models

## 12.1 Introduction

In this section we prove Theorems 10.1.1 and 10.1.4. First, in Chapter 12 we present SQ lower bounds for general classes of Multi-Index Models that do not satisfy Definition 10.1.2. Essentially showing that the algorithm of Section 10.3 is quantitatively optimal in the SQ model. Second, in Section (12.1.2), we present an SQ lower bound for the problem of learning multiclass linear classifiers under RCN. For an introduction to the SQ model, we refer the reader to Section 1.4.

### 12.1.1 SQ Lower Bound for General Multi-Index Models

We show that if a Multi-Index Model class $\mathcal{F}$ does not satisfy Definition 10.3.2, any SQ algorithm that learns $\mathcal{F}$ to an accuracy comparable to that of Theorem 10.3.3 requires

super-polynomial complexity. Specifically, we establish the following theorem:

**Theorem 12.1.1** (SQ Optimality of General Algorithm)**.** *Let $d, K, m \in \mathbb{Z}_+$ with $d^c \geq \max(m, K)$ for a sufficiently small absolute constant $c > 0$, and let $\mathcal{Y}$ be a set of finite cardinality. Additionally, let $\tau, \mathrm{OPT} \in (0, 1)$ with $\tau, \mathrm{OPT} \geq d^{-O(m)}$. Assume the existence of (i) a function $f : \mathbb{R}^d \mapsto \mathcal{Y}$ (ii) a random variable $(\mathbf{x}, y)$ over $\mathbb{R}^d \times \mathcal{Y}$ with $\mathbf{x}$-marginal the standard Gaussian such that $\mathbf{Pr}[f(\mathbf{x}) \neq y] = \mathrm{OPT}$ (iii) a subspace $W$ of dimension at most $K$, and iv) a subspace $V$ such that:*

1. *The function $f$ depends only on the projection onto $W$, i.e., $f(\mathbf{x}) = f(\mathbf{x}^W)$.*
2. *For every function $g : V \to \mathcal{Y}$ we have that $\mathbf{Pr}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[g(\mathbf{x}^V) \neq f(\mathbf{x})] \geq \tau$.*
3. *For any $\mathbf{x}_0 \in \mathbb{R}^d, i \in [K]$, and for any zero-mean polynomial $p : U \to \mathbb{R}$ of degree at most $m$, it holds that $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}[p(\mathbf{x}^U) \mathbb{1}(y = i) \mid \mathbf{x}^V = \mathbf{x}_0^V] = 0$, where $U = (V + W) \cap V^\perp$.*

*Let $\mathcal{F}$ denote the class of MIMs defined as $\mathcal{F} = \{f(\mathbf{x}^{\perp U}, \mathbf{U}\mathbf{x}^U) : \text{for all } \mathbf{U} \in \mathbb{R}^{k \times d} : \mathbf{U}\mathbf{U}^\intercal = \mathbf{I}_k\}$, where $k = \dim(U)$. Then, any SQ algorithm that learns $\mathcal{F}$ in the presence of $\mathrm{OPT}$ adversarial label noise up to an error of smaller than $\tau - 2\mathrm{OPT}$ with probability at least $2/3$ requires either $2^{d^{\Omega(1)}}$ queries or at least one query to $\mathrm{VSTAT}\left(d^{\Omega(m)}/|\mathcal{Y}|\right)$.*

*Proof.* Let $U = (V + W) \cap V^\perp$ with $K' = \dim(U)$ and let $U^\perp$ be the orthogonal complement of $U$.

Let $\mathcal{Q} \subseteq \mathbb{R}^{d \times K'}$ be the set of matrices from Lemma 11.2.5 for which it holds that $|\mathcal{Q}| = 2^{\Omega(d^c)}, \mathbf{Q}\mathbf{Q}^\intercal = \mathbf{I}_{K'}$ and $\|\mathbf{Q}\mathbf{P}^\intercal\|_F \leq O(d^{2c-1+2\alpha})$ for all $\mathbf{Q}, \mathbf{P} \in \mathcal{Q}$, where $\alpha = \log K'/\log d$. For each $\mathbf{Q} \in \mathcal{Q}$, we define the function $f_{\mathbf{Q}}(\mathbf{x}) := f(\mathbf{x}^{\perp U}, \mathbf{Q}\mathbf{x}^U)$ along with the coupled random variables $y_{\mathbf{Q}}$ satisfying $\mathbf{Pr}[y_{\mathbf{Q}} = i | \mathbf{x} = \mathbf{z}] = \mathbf{Pr}[y = i | \mathbf{x} = \mathbf{z}^{\perp U} + \mathbf{Q}\,\mathbf{z}^U]$, for all

$i \in \mathcal{Y}, \mathbf{z} \in \mathbb{R}^d$. We consider the class of functions $\mathcal{F}_\mathcal{Q} = \{f_\mathbf{Q}(\mathbf{x}) : \mathbf{Q} \in \mathcal{Q}\}$ and note that by definition $\mathcal{F}_\mathcal{Q} \subseteq \mathcal{F}$.

For each $\mathbf{Q} \in \mathcal{Q}$, let $D_\mathbf{Q}$ be the joint distribution of $(\mathbf{x}, y_\mathbf{Q})$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Denote by $\mathcal{D} = \{D_\mathbf{Q} : \mathbf{Q} \in \mathcal{Q}\}$ the class of these distributions. Denote by $D_0$ the distribution of pairs $(\mathbf{x}, y_0)$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $y_0$ a random variable supported on $\mathcal{Y}$ that depends only on $W^V$ such that $\mathbf{Pr}_{(\mathbf{x},y_0)\sim D_0}[y_0 = i \mid \mathbf{x}^{\perp U} = \mathbf{z}^{\perp U}] = \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[y = i \mid \mathbf{x}^{\perp U} = \mathbf{z}^{\perp U}]$, for all $i \in \mathcal{Y}, \mathbf{z} \in \mathbb{R}^d$. Denote by $y_i(\mathbf{x}^{\perp U}) = \mathbf{Pr}_{(\mathbf{x},y)\sim D_0}[y_0 = i \mid \mathbf{x}^{\perp U}]$.

We show that is SQ-Hard to determinate whether we have access to distribution $D_0$ or to a distribution $D \in \mathcal{D}$. To this end, we show that for all $D_\mathbf{P}, D_\mathbf{Q} \in \mathcal{D}$ with $\mathbf{P} \neq \mathbf{Q}$, the correlation $\chi_{D_0}(D_\mathbf{P}, D_\mathbf{Q})$ (see Definition 1.4.7) is small. Using the law of total expectation, we have that

$$\chi_{D_0}(D_\mathbf{P}, D_\mathbf{Q}) = \sum_{i\in\mathcal{Y}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\frac{1}{y_i(\mathbf{x}^{\perp U})}(\mathbb{1}(y_\mathbf{P} = i) - y_i(\mathbf{x}^{\perp U}))(\mathbb{1}(y_\mathbf{Q} = i) - y_i(\mathbf{x}^{\perp U}))\right]$$
$$= \sum_{i\in\mathcal{Y}} \mathop{\mathbf{E}}_{\mathbf{x}^{\perp U}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\frac{1}{y_i(\mathbf{x}^{\perp U})} \mathop{\mathbf{E}}_{\mathbf{x}^{U}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbb{1}(y_\mathbf{P} = i) - y_i(\mathbf{x}^{\perp U}))(\mathbb{1}(y_\mathbf{Q} = i) - y_i(\mathbf{x}^{\perp U})) \mid \mathbf{x}^{\perp U}]\right].$$

Considering each term separately, by Lemma 11.2.3 and Assumption (3) of the theorem statement, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}^{U}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[(\mathbb{1}(y_\mathbf{P} = i) - y_i(\mathbf{x}^{\perp U}))(\mathbb{1}(y_\mathbf{Q} = i) - y_i(\mathbf{x}^{\perp U})) \mid \mathbf{x}^{\perp U}]$$
$$\leq \sum_{t=m}^{\infty} \|\mathbf{P}\mathbf{Q}^\intercal\|_2^t (y^{[t]})^2 \leq \|\mathbf{P}\mathbf{Q}^\intercal\|_2^m \sum_{t=m}^{\infty} (y^{[t]})^2 \,,$$

where $y^{[t]}$ is the degree-$t$ Hermite part of the random variable $\mathbb{1}(y = i) \mid \mathbf{x}^{\perp U}$ and that $\|\mathbf{P}\mathbf{Q}^\intercal\|_2 \leq 1$. Furthermore, using Parseval's identity we get that $\sum_{t=m}^{\infty}(y^{[t]})^2 \leq \mathbf{E}[\mathbb{1}(y =$

$i)^2 \mid \mathbf{x}^{\perp U}] = \mathbf{E}[\mathbb{1}(y = i) \mid \mathbf{x}^{\perp U}]$. Therefore, we have that

$$\chi_{D_0}(D_{\mathbf{P}}, D_{\mathbf{Q}}) \leq \|\mathbf{P}\mathbf{Q}^\intercal\|_2^m \sum_{i \in \mathcal{Y}} \mathop{\mathbf{E}}_{\mathbf{x}^{\perp U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[ \frac{\mathbf{E}[\mathbb{1}(y = i) \mid \mathbf{x}^{\perp U}]}{y_i(\mathbf{x}^{\perp U})} \right] = \|\mathbf{P}\mathbf{Q}^\intercal\|_2^m |\mathcal{Y}| \,,$$

where we used that $\mathbf{E}_{\mathbf{x}^{\perp U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[ (\mathbf{E}[\mathbb{1}(y = i) \mid \mathbf{x}^{\perp U}])/y_i(\mathbf{x}^{\perp U}) \right] = 1$, by the definition of $y_i(\mathbf{x}^{\perp U})$. Hence, $\chi_{D_0}(D_{\mathbf{P}}, D_{\mathbf{Q}}) = d^{-\Omega(m)}|\mathcal{Y}|$. Similarly, we can show that $\chi_{D_0}(D_{\mathbf{P}}, D_{\mathbf{P}}) \leq |\mathcal{Y}|$ for all $D_{\mathbf{P}} \in \mathcal{D}$. Therefore, we have that $\mathcal{D}$ is $(d^{-\Omega(m)}|\mathcal{Y}|, |\mathcal{Y}|)$-correlated relative to $D_0$ (see Definition 1.4.8), and thus from Lemma 1.4.9 we have that any SQ algorithm that solves the decision problem $\mathcal{B}(\mathcal{D}, D_0)$ with probability at least $2/3$ requires $2^{d^{\Omega(1)}}$ queries to the VSTAT$(d^{\Omega(m)}/|\mathcal{Y}|)$ oracle.

Hence, in order to complete the statement of the theorem, we need to show the following standard reduction from the testing problem to the learning problem (see, e.g. Chapter 8 of [DK23a]).

**Claim 12.1.2.** *Any SQ algorithm that learns $\mathcal{F}$ with agnostic noise with error better than $\tau - 2\mathrm{OPT}$ can be used to solve the decision problem $\mathcal{B}(\mathcal{D}, D_0)$ using one more query of accuracy* OPT.

*Proof.* Assume there exists an SQ algorithm $\mathcal{A}$ that, when given SQ access to any distribution $D'$ over $\mathbb{R}^d \times \mathcal{Y}$ with standard Gaussian $\mathbf{x}$-marginals, outputs a hypothesis $h$ satisfying $\mathbf{Pr}_{(\mathbf{x},y)\sim D'}[h(\mathbf{x}) \neq y] < \tau - 2\mathrm{OPT}$, where $\mathrm{OPT} = \inf_{f \in \mathcal{F}} \mathbf{Pr}_{(\mathbf{x},y)\sim D'}[f(\mathbf{x}) \neq y]$. Using such an algorithm $\mathcal{A}$, we show that we can solve the testing problem.

Suppose we run $\mathcal{A}$ on a distribution $D_{\mathbf{P}} \in \mathcal{D}$. Then, by definition, $\mathcal{A}$ returns a hypothesis $h$ with error $\mathbf{Pr}_{(\mathbf{x},y)\sim D_{\mathbf{P}}}[h(\mathbf{x}) \neq y] < \tau - 2\mathrm{OPT}$. In contrast, when we run $\mathcal{A}$ with SQ oracle access to $D_0$ no hypothesis can achieve error lower than $\tau - \mathrm{OPT}$, as $D_0$ is a product distribution for each fixed projection of $\mathbf{x}$ in $V$. Since $\tau - \mathrm{OPT} > \tau - 2\mathrm{OPT}$,

an additional query of tolerance less than OPT, would distinguish between $D_{\mathbf{P}}$ and $D_0$, thus solving $\mathcal{B}(\mathcal{D}, D_0)$. □

□

### 12.1.2 SQ Lower Bounds for Learning Multiclass Linear Classifiers with RCN

In this section, we present an SQ hardness result for learning Multiclass Linear Classifiers in the presence of RCN under the Gaussian distribution. In particular, we show that achieving $0-1$ error of $\mathrm{OPT}+\epsilon$ in the SQ model must either make a super-polynomial (with respect the dimension $d$) number of queries or queries with accuracy that depends on the spectral properties of the confusion matrix (see Definition 12.1.3), rather than on the amount of noise.

**Definition 12.1.3** (RCN Distribution)**.** *Let $D$ be a distribution over $\mathbb{R}^d \times \mathcal{Y}$. We say that $D$ is an RCN distribution for the class $\mathcal{F} \subseteq \{f : \mathbb{R}^d \to \mathcal{Y}\}$ if there exists a function $f \in \mathcal{F}$ and a row-stochastic matrix $\mathbf{H} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ with $\mathbf{H}_{i,i} > \mathbf{H}_{i,j} + \gamma$ for all $i, j \in \mathcal{Y}$, for some $\gamma > 0$, such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[y = i \mid \mathbf{x}] = \mathbf{H}_{f(\mathbf{x}),i}$. We refer to $\mathbf{H}$ as the confusion matrix of the distribution.*

The main result of this section is Theorem 10.1.1.

**Theorem 12.1.4** (SQ Lower Bound for Learning under MLC RCN)**.** *Any SQ algorithm that learns $K$-Multiclass Linear Classifiers under the $d$-dimensional standard normal in the presence of RCN to error $\mathrm{OPT} + \mathrm{poly}(1/K)$, requires either $2^{d^{\Omega(1)}}$ queries or at least one query to VSTAT$\left(d^{\Omega(K)}\right)$.*

**SQ-Hard Instances** To construct our family of hard instances, we first define a two-dimensional classifier $f : \mathbb{R}^2 \to [K]$ as follows:

$$\mathbf{w}^{(k)} = \begin{bmatrix} \cos\left(\frac{2\pi k}{K}\right) \\ \sin\left(\frac{2\pi k}{K}\right) \end{bmatrix}, k \in [K], \text{ and } f(\mathbf{x}) = \operatorname*{argmax}_{k \in [K]} (\mathbf{w}^{(k)} \cdot \mathbf{x}) \,. \tag{12.1}$$

Geometrically, the classifier $f$ partitions the plane into $K$ equal angular sectors, assigning each sector to a different class. Given a set of $2 \times d$ matrices $\mathcal{Q}$, we can define a class of functions from $\mathbb{R}^d$ to $[K]$ as follows

$$\mathcal{F}_{\mathcal{Q}} = \{f(\mathbf{Q}\mathbf{x}) : \mathbf{Q} \in \mathcal{Q}\} \,. \tag{12.2}$$

We show that it is hard to distinguish a distribution labeled by a classifier in $\mathcal{F}_{\mathcal{Q}}$ corrupted with RCN from a distribution with random labels independent of $\mathbf{x}$. Before proving Theorem 12.1.4, we first need to show that we can match the appropriate number of moments with the standard Gaussian.

**Lemma 12.1.5** (Moment matching under RCN)**.** *Let $K \in \mathbb{Z}_+$ be such that $K \geq 8$ and $K$ is divisible by 4. Then there exists a $K \times K$ doubly stochastic matrix $\mathbf{H}$ and a distribution $D$ over $\mathbb{R}^2 \times [K]$, whose $\mathbf{x}$-marginal is $\mathcal{N}(\mathbf{0}, \mathbf{I})$, satisfying Definition 12.1.3 for the confusion matrix $\mathbf{H}$ and the function $f$ defined in Equation (12.1), such that:*

*i)* $\min_{i \neq j} \mathbf{H}_{i,i} - \mathbf{H}_{i,j} = \Omega(1/K^3)$.

*ii) For every polynomial $p : \mathbb{R}^2 \to \mathbb{R}$ of degree at most $K/2 - 2$, it holds that*
$\mathbf{E}_{\mathbf{x} \sim \mathcal{N}^2} \left[ \left( \mathbf{Pr}_{(\mathbf{x},y) \sim D} [y = i \mid \mathbf{x}] - 1/K \right) p(\mathbf{x}) \right] = 0$ *for all $i \in [K]$ .*

*Proof.* Let $\mathbf{H} \in \mathbb{R}^{K \times K}$ be a doubly stochastic matrix and define the functions $\tilde{f}_i$ for $i \in [K]$ as $\tilde{f}_i(\mathbf{x}) = \sum_{j=1}^{K} \mathbf{H}_{j,i} \mathbb{1}(f(\mathbf{x}) = j)$. For any function $p : \mathbb{R}^2 \to \mathbb{R}$, we denote by $\bar{p}$

the zero mean centering of $p$, i.e., $\bar{p}(\mathbf{x}) = p(\mathbf{x}) - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^2}[p(\mathbf{x})]$, and by $\mathbf{v}^{(p)} \in \mathbb{R}^K$ the vector defined as $\mathbf{v}_i^{(p)} = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^2}[\bar{p}(\mathbf{x})\mathbb{1}(f(\mathbf{x}) = i)]$. We will refer to $\mathbf{v}^{(p)}$ as the vector of expectations of $p$. Also denote by $\boldsymbol{\omega}^{(k)}$ the root of unity vector of rate $k$, i.e. $\boldsymbol{\omega}^{(k)} \in \mathbb{C}^K$ such that $\boldsymbol{\omega}_j^{(k)} = e^{\frac{2\pi kji}{K}}$, where $i$ is the imaginary unit.

We start by showing that a function $p$ has zero correlation with $\tilde{f}_i - 1/K$ for all $i \in [K]$ under the standard Gaussian if and only if the vector of expectations of $p$ over the level sets belongs in the kernel of $\mathbf{H}^\intercal$.

**Claim 12.1.6.** *Let $p : \mathbb{R}^2 \to \mathbb{R}$. We have that $\mathbf{v}^{(p)} \in \ker(\mathbf{H}^\intercal)$ if and only if for all $i \in [K]$ it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^2}[(\tilde{f}_i(\mathbf{x}) - 1/K)p(\mathbf{x})] = 0$.*

*Proof of Claim 12.1.6.* By the rotational invariance of the standard Gaussian we have that $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}^2}[f(\mathbf{x}) = i] = 1/K$, for all $i \in [K]$. From the assumption that $\mathbf{H}$ is column stochastic, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^2}[\tilde{f}_i(\mathbf{x})] = \sum_{j=1}^K \mathbf{H}_{ji}\,\mathbf{E}[\mathbb{1}(f(\mathbf{x}) = i)] = 1/K$. Thus, for all $i \in [K]$ we have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}\left[\left(\tilde{f}_i(\mathbf{x}) - \frac{1}{K}\right)p(\mathbf{x})\right] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}\left[\tilde{f}_i(\mathbf{x})p(\mathbf{x})\right] - \frac{1}{K}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}[p(\mathbf{x})] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}\left[\tilde{f}_i(\mathbf{x})p(\mathbf{x})\right] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}[p(\mathbf{x})]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}\left[\tilde{f}_i(\mathbf{x})\right] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}\left[\tilde{f}_i(\mathbf{x})\bar{p}(\mathbf{x})\right] .
\end{aligned}$$

Therefore, substituting the definition of $\tilde{f}_i$ we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}\left[\tilde{f}_i(\mathbf{x})\bar{p}(\mathbf{x})\right] = \sum_{j=1}^K \mathbf{H}_{j,i}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}[\bar{p}(\mathbf{x})\mathbb{1}(f(\mathbf{x}) = j)] = (\mathbf{H}^\intercal)_i\mathbf{v}^{(p)} ,$$

which concludes the proof. □

We next observe that all polynomials $\mathbb{R}^2 \mapsto \mathbb{R}$ of degree at most $m$ can be expressed as

a linear combination of polynomials of the form $(x+iy)^a(x-iy)^b$, where $a, b > 0, a+b \leq m$ and $i$ is the imaginary unit. Specifically, we show that the vectors associated with each polynomial of this basis are precisely the root of unity vectors.

**Claim 12.1.7.** *Let $p(x,y) = (x+iy)^a(x-iy)^b, a, b \in \mathbb{Z}_+$. If $a \neq b$, then there exists a constant $c \in \mathbb{C}$ such that $\mathbf{v}^{(p)} = c\boldsymbol{\omega}^{(a-b)}$; and if $a = b$, then $\mathbf{v} = \mathbf{0}$.*

*Proof of Claim 12.1.7.* By the definition of $f$, the region $\{(x,y) \in \mathbb{R}^2 : f(x,y) = j\}$ is equivalent to $\theta \in [2\pi j/K - \pi/K, 2\pi j/K + \pi/K]$, where $\theta = \arctan(y/x)$. Fix $a, b \in \mathbb{Z}_+, a \neq b$, and let $\theta_{1,j} = 2\pi j/K - \pi/K$, $\theta_{2,j} = 2\pi j/K + \pi/K$, $j \in [K]$. We have that

$$\begin{aligned}
\mathbf{v}_j^{(p)} = \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{N}^2}[p(x,y)\mathbb{1}(f(x,y)=j)] &= \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{N}^2}[\left(x^2+y^2\right)^{\frac{a+b}{2}} e^{i\theta(a-b)}\mathbb{1}(f(x,y)=j)] \\
&= \int_0^\infty \int_{\theta_{1,j}}^{\theta_{2,j}} r^{a+b+1} e^{i\theta(a-b)} g(r) d\theta dr \\
&= \left(\int_0^\infty r^{a+b+1} g(r) dr\right)\left(\int_{\theta_{1,j}}^{\theta_{2,j}} e^{i\theta(a-b)} d\theta\right) \\
&= c\frac{1}{i(a-b)}\left(e^{i\theta_{2,j}(a-b)} - e^{i\theta_{1,j}(a-b)}\right) ,
\end{aligned}$$

where $g : \mathbb{R} \to \mathbb{R}_+$ denotes the pdf of the Gaussian distribution in polar coordinates and $c = \left(\int_0^\infty r^{a+b+1} g(r) dr\right)$. Notice that $c$ depends only on $a+b$, and hence is the same for all coordinates of the vector $\mathbf{v}^{(p)}$. Also note that for $a = b$ we have that the vector $\mathbf{v}^{(p)}$ is parallel to the all ones vector, as the expectation is constant.

Therefore, it holds that $\mathbf{v}^{(p)}_{(j+1) \mod K} = \mathbf{v}^{(p)}_j e^{\frac{2\pi i}{K}(a-b)}$ for all $j \in \{0, \ldots, K-1\}$, since $\theta_{1,(j+1)} = \theta_{1,j} + 2\pi/K$ and $\theta_{2,(j+1)} = \theta_{2,j} + 2\pi/K$. Moreover, if $a \neq b$, we observe that the integral of $p(x,y)$ over $\mathbb{R}^2$ is zero, which concludes the proof of the claim. □

By Claim 12.1.7, in order to have 0 correlation with all polynomials of degree at

most $m$, we need to have that $\boldsymbol{\omega}^{(k)} \in \ker(\mathbf{H}^\intercal)$ for all $k \in \{1, \ldots, m, K-m, \ldots, K-1\}$, as the roots of unity are symmetric, i.e., $\boldsymbol{\omega}^{(-k)} = \boldsymbol{\omega}^{(K-k)}$.

We first construct a stochastic vector $\mathbf{h} \in \mathbb{R}^K$ with $\mathbf{h}_0 - \mathbf{h}_j = \Omega(1/K^3)$, for all $j \in \{1, \ldots, K-1\}$ that is also orthogonal to the aforementioned root of unity vectors for $m = \Omega(K)$. Next, we will extend this vector to a matrix. Define the vectors $\mathbf{h}^{(1)} = \boldsymbol{\omega}^{(0)}/K$, $\mathbf{h}^{(2)} = (\boldsymbol{\omega}^{(K/2-1)} + \boldsymbol{\omega}^{(K/2+1)})/K$ and $\mathbf{h}' = 2\mathbf{h}^{(1)} + \mathbf{h}^{(2)}$. Note that $\mathbf{h}^{(1)}_j = 1/K$ and $\mathbf{h}^{(2)}_j = (2/K)(-1)^j \cos(2\pi j/K)$ for all $j \in \{0, \ldots, K-1\}$. Hence, we have that all the coordinates of $\mathbf{h}'$ are nonnegative. Moreover, $\mathbf{h}'$ attains its maximum value at $\mathbf{h}'_0 = 4/K$, since $\mathbf{h}'_{K/2} = 0$ (because $(-1)^j \cos(2\pi j/K) = -1$ when $j = K/2$ and $K$ is divisible by 4). Furthermore, it holds that $\max_{j\in[K-1]} \mathbf{h}'_j \leq 2/K + \max_{j\in[K-1]} |\mathbf{h}^{(2)}_j| = 2/K + 2/K \cos(2\pi/K)$, and hence we have

$$\mathbf{h}'_0 - \max_{j\in[K-1]} \mathbf{h}'_j \geq \frac{2}{K}\left(1 - \cos\left(\frac{2\pi}{K}\right)\right) \geq \frac{2}{K}\left(\left(\frac{2\pi}{K}\right)^2/2 - \left(\frac{2\pi}{K}\right)^4/24\right) \gtrsim \frac{1}{K^3}\,,$$

where the second inequality follows from the Taylor series of $\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$, which is an alternating decreasing series when $x = 2\pi/K \leq 1$ (thus it can be upper bounded by the first three terms). Let $\mathbf{h}$ be the normalization of $\mathbf{h}'$ such that the sum of its coordinates equals 1. Given that each element of $\mathbf{h}'$ is at most $4/K$, we have that the normalization factor is at most 4. Hence, $\mathbf{h}_0 - \max_{j\in[K-1]} \mathbf{h}_j \gtrsim 1/K^3$. Moreover, from the well-known fact that $\boldsymbol{\omega}^{(0)}, \ldots, \boldsymbol{\omega}^{(K-1)}$ form an orthogonal basis of $\mathbb{C}^K$, we have that $\mathbf{h}$ is orthogonal to $\boldsymbol{\omega}^{(k)}$ for all $k \in \{1, \ldots, K/2-2, K/2, K/2+2, \ldots, K-1\}$, since it is a linear combination of $\boldsymbol{\omega}^{(0)}, \boldsymbol{\omega}^{(K/2-1)}, \boldsymbol{\omega}^{(K/2+1)}$.

Now consider $\mathbf{H}$, the circulant matrix generated by row shifts of the vector $h$. By definition, we have that $\mathbf{H}$ is a row stochastic matrix and that $\min_{i\neq j} \mathbf{H}_{ii} - \mathbf{H}_{i,j} = \Omega(1/K^3)$. Moreover, $\mathbf{H}$ is symmetric, because it holds that $\mathbf{H}_{i,j} = \mathbf{h}_{j-i \bmod K} =$

$\mathbf{h}_{i-j \bmod K} = \mathbf{H}_{j,i}$ and $\mathbf{h}_j = \mathbf{h}_{K-j}, j \in [K-1]$. This equality is justified by the identity $(-1)^j \cos(2\pi j/K) = (-1)^{K-j} \cos(2\pi(K-j)/K)$, which holds since $K$ is an even integer. Hence $\mathbf{H}$ is also column stochastic.

By Fact 10.1.12, the eigenvectors of any circulant matrix are exactly $\boldsymbol{\omega}^{(k)}$ with eigenvalues $\mathbf{h} \cdot \boldsymbol{\omega}^{(k)}, k \in \{0, \ldots, K-1\}$. Hence, $\boldsymbol{\omega}^{(k)} \in \ker(\mathbf{H}) = \ker(\mathbf{H}^\intercal)$ for all $k \in \{1, \ldots, K/2-2, K/2, K/2+2, \ldots, K-1\}$. Therefore, by applying Claim 12.1.6, we conclude the proof of Lemma 12.1.5. □

### Proof of Theorem 12.1.4

*Proof of Theorem 12.1.4.* Consider the case where $K$ is divisible by 4 and let $\mathbf{H}$ be a matrix that satisfies the statement of Lemma 12.1.5 Let $f : \mathbb{R}^2 \to [K]$ be the linear classifier defined in Equation (12.1), and let $\mathcal{F}_\mathcal{Q}$ be the class of functions defined in Equation (12.2), where $\mathcal{Q} \subseteq \mathbb{R}^{d \times 2}$ is the family that satisfies the statement of Lemma 11.2.5. We define the family of distributions $\mathcal{D}$ such that $D_\mathbf{Q} \in \mathcal{D}$ if and only if $D_\mathbf{Q}$ is a joint distribution $(\mathbf{x}, y_\mathbf{Q})$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $y_\mathbf{Q}$ is a random variable supported on $[K]$ defined as $\mathbf{Pr}[y_\mathbf{Q} = i \mid \mathbf{x}] = \mathbf{H}_{f(\mathbf{Q}\mathbf{x}),i}$ for some $\mathbf{Q} \in Q$. Note that all distributions in $\mathcal{D}$ satisfy Definition 12.1.3 for confusion matrix $\mathbf{H}$ and some function in $\mathcal{F}_\mathcal{Q}$. Denote by $D$ the product distribution with $\mathbf{x}$-marginals distributed as standard normal and the labels $y$ are distributed uniformly on $[K]$. For $D_\mathbf{Q}, D_\mathbf{P} \in \mathcal{D}$ with $\mathbf{Q} \neq \mathbf{P}$,

we have that

$$\begin{aligned}
\chi_D(D_{\mathbf{Q}}, D_{\mathbf{P}}) &= K \sum_{i=1}^{K} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d} \left[ (\mathbf{Pr}[y_{\mathbf{Q}} = i \mid \mathbf{x}] - \frac{1}{K})(\mathbf{Pr}[y_{\mathbf{P}} = i \mid \mathbf{x}] - \frac{1}{K}) \right] \\
&= K \sum_{i=1}^{K} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d} \left[ (\mathbf{H}_{f(\mathbf{Q}\mathbf{x}),i} - \frac{1}{K})(\mathbf{H}_{f(\mathbf{P}\mathbf{x}),i} - \frac{1}{K}) \right] \\
&= K \sum_{i=1}^{K} \sum_{j=1}^{\infty} \|\mathbf{P}\mathbf{Q}^{\intercal}\|_2^j \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2} \left[ ((\mathbf{H}_{f(\mathbf{x}),i} - \frac{1}{K})^{[j]})^2 \right] \\
&= \|\mathbf{P}\mathbf{Q}^{\intercal}\|_2^{K/2-2} \, K \sum_{i=1}^{K} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2} \left[ (\mathbf{H}_{f(\mathbf{x}),i} - \frac{1}{K})^2 \right] \leq d^{\Omega(K)(c-1/2)} K \ ,
\end{aligned}$$

where we used Lemma 11.2.3, Parseval's identity and the fact that $\mathbf{H}$ satisfies the statement of Lemma 12.1.5. Similarly, we have that $\chi_D(D_{\mathbf{Q}}, D_{\mathbf{Q}}) \leq K$ for all $D_{\mathbf{Q}} \in \mathcal{D}$.

Consequently, we have that there exists a family of distributions of size $2^{\Omega(d^c)}$ that is $(d^{\Omega(K)(c-1/2)}K, K)$ correlated. Hence, by Lemma 1.4.9 we have that any SQ algorithm that solves the decision problem $\mathcal{B}(\mathcal{D}, D)$ requires $2^{\Omega(d^c)} d^{\Omega(K)(c-1/2)} = 2^{d^{\Omega(1)}}$ (where we used the assumption that $d/\log(d)$ is much larger that $K$) queries to the VSTAT($d^{\Omega(K)}$.

In order to complete the proof of the theorem, we need to show the following standard reduction from the testing problem to the learning problem.

**Claim 12.1.8.** *Let $\gamma = \min_{i,j} \mathbf{H}_{i,i} - \mathbf{H}_{i,j}$. Any SQ algorithm that learns multiclass linear classifiers under the standard normal in the presence of RCN with error better than* OPT $+ \gamma/2$ *can be used to solve the decision problem $\mathcal{B}(\mathcal{D}, D)$ using one more query of accuracy $\gamma/4$.*

*Proof.* Assume that there exists an SQ algorithm $\mathcal{A}$ that if given access to a distribution $D'$ over $\mathbb{R}^d \times \mathcal{Y}$ whose $\mathbf{x}$-marginal is the standard normal returns a hypothesis $h$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D'}[h(\mathbf{x}) \neq y] \leq$ OPT $+ \gamma/4$. Using $\mathcal{A}$, we can solve the testing problem.

Note that by running $\mathcal{A}$ to distribution $D_{\mathbf{P}} \in \mathcal{D}$, $\mathcal{A}$ returns a hypothesis $h$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim D_{\mathbf{P}}}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \gamma/4$. Moreover, by running $\mathcal{A}$ with SQ oracle access to $D$ results in a hypothesis $h$ that cannot perform better than $1 - 1/K$, as $D$ is a product distribution with the uniform distribution over the labels. Since $\mathrm{OPT} < 1 - 1/K - \gamma/2$, because $\mathbf{H}_{i,i} \geq 1/K + \gamma/2$ as $\sum_j \mathbf{H}_{i,j} = 1, i \in [K]$, we have that with an additional query of tolerance less than $\gamma/4$, we can solve $\mathcal{B}(\mathcal{D}, D)$. □

For $K$ not divisible by 4, we add $r = 4 \mod K$ additional classes whose weight vectors are identical to $\mathbf{w}^{(1)}$, and with no noise, i.e., for these added classes $j$, we have that $\mathbf{H}_{j,j} = 1$ and $\mathbf{H}_{j,i} = \mathbf{H}_{i,j} = 0$ for all $j \neq i$. For the remaining classes, i.e., for $i, j \in [K - r]$, we consider a matrix $\mathbf{H}_{i,j}$ that satisfies the conditions of Lemma 12.1.5. Note that since these additional classes have zero probability (as ties are broken by the smallest index), our SQ hardness result extends to the general case of any $K$. This completes the proof. □

# Bibliography

[AAB+21] E. Abbe, E. Adsera, M. Brennan, G. Bresler, and D. Nagaraj. The staircase property: How hierarchical structure can guide deep learning. *Advances in Neural Information Processing Systems*, 34:26989–27002, 2021.

[AAM22] E. Abbe, E. B. Adsera, and T. Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[AAM23] E. Abbe, E. B. Adsera, and T. Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023.

[ABHU15] P. Awasthi, M. F. Balcan, N. Haghtalab, and R. Urner. Efficient learning of linear separators under bounded noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pages 167–190, 2015.

[ABHZ16] P. Awasthi, M. F. Balcan, N. Haghtalab, and H. Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, pages 152–192, 2016.

[ABL17] P. Awasthi, M. F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63(6):50:1–50:27, 2017.

[AL88] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[Ang87] D. Angluin. Learning Regular Sets from Queries and Counterexamples. *Information and Computation*, 75(2):87–106, 1987.

[ATV23] P. Awasthi, A. Tang, and A. Vijayaraghavan. Agnostic learning of general ReLU activation using gradient descent. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

[AV99] R. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proceedings of the 40$^{th}$ Annual Symposium on Foundations of Computer Science (FOCS)*, pages 616–623, New York, NY, 1999.

[BAGJ21] G. Ben Arous, R. Gheissari, and A. Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

[BBH$^+$12] B. Barak, F. G. S. L. Brandão, A. W. Harrow, J. A. Kelner, D. Steurer, and Y. Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, 2012*, pages 307–326. ACM, 2012.

[BBSS22] A. Bietti, J. Bruna, C. Sanford, and M. Song. Learning single-index models with shallow neural networks. *Advances in Neural Information Processing Systems*, 2022.

[BJL$^+$19] S. Bubeck, Q. Jiang, Y.-T. Lee, Y. Li, and A. Sidford. Complexity of highly parallel non-smooth convex optimization. *Advances in neural information processing systems*, 32, 2019.

[BJW19] A. Bakshi, R. Jayaram, and D. P. Woodruff. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory, COLT 2019*, 2019.

[BK09] E. Beigman and B. B. Klebanov. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47$^{th}$ Annual Meeting of the ACL and the 4$^{th}$ International Joint Conference on Natural Language Processing of the AFNLP*, pages 280–287, 2009.

[BLQT22] G. Blanc, J. Lange, M. Qiao, and L.-Y. Tan. Properly learning decision trees in almost polynomial time. *J. ACM*, 69(6):39:1–39:19, 2022.

[Blu03] A. Blum. Machine learning: My favorite results, directions, and open problems. In *44$^{th}$ Symposium on Foundations of Computer Science (FOCS 2003)*, pages 11–14, 2003.

[BNPS17] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[BNS16] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*, 29, 2016.

[Bog98] V. Bogachev. *Gaussian measures*. Mathematical surveys and monographs, vol. 62, 1998.

[BPS+19] A. Beygelzimer, D. Pal, B. Szorenyi, D. Thiruvenkatachari, C. Wei, and C. Zhang. Bandit multiclass linear classification: Efficient algorithms for the separable case. In *International Conference on Machine Learning*. PMLR, 2019.

[BZ17] M.-F. Balcan and H. Zhang. Sample and computationally efficient learning algorithms under s-concave distributions. In *Advances in Neural Information Processing Systems*, pages 4796–4805, 2017.

[CCK17] V. Chernozhukov, D. Chetverikov, and K. Kato. Detailed proof of Nazarov's inequality. *arXiv preprint arXiv:1711.10696*, 2017.

[CDG+23] S. Chen, Z. Dou, S. Goel, A. Klivans, and R. Meka. Learning narrow one-hidden-layer relu networks. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 2023.

[CKL+06] C.-T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. Bradski, A. Y. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *Proceedings of the 19$^{th}$ International Conference on Neural Information Processing Systems*, NIPS'06, pages 281–288, Cambridge, MA, USA, 2006. MIT Press.

[CKM21] S. Chen, A. R. Klivans, and R. Meka. Efficiently learning one hidden layer relu networks from queries. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 24087–24098, 2021.

[CKM22] S. Chen, A. R. Klivans, and R. Meka. Learning deep relu networks is fixed-parameter tractable. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, 2022.

[CKMY20] S. Chen, F. Koehler, A. Moitra, and M. Yau. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[CKMY21] S. Chen, F. Koehler, A. Moitra, and M. Yau. Online and distribution-free robustness: Regression and contextual bandits with huber contamination. In *FOCS*, 2021.

[CKST24] G. Chandrasekaran, V. Kontonis, K. Stavropoulos, and K. Tian. Learning noisy halfspaces with a margin: Massart is no harder than random. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Cla90] F. H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[CM84] R. S. Chhikara and J. McKeon. Linear discriminant analysis with misallocation in training samples. *Journal of the American Statistical Association*, 79(388):899–906, 1984.

[CM20] S. Chen and R. Meka. Learning polynomials in few relevant dimensions. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 1161–1227. PMLR, 2020.

[CN23] S. Chen and S. Narayanan. A faster and simpler algorithm for learning shallow networks. *CoRR*, abs/2307.12496, 2023.

[CW01] A. Carbery and J. Wright. Distributional and $L^q$ norm inequalities for polynomials over convex bodies in $R^n$. *Mathematical Research Letters*, 8(3):233–248, 2001.

[Dan15] A. Daniely. A PTAS for agnostically learning halfspaces. In *Proceedings of The 28$^{th}$ Conference on Learning Theory, COLT 2015*, pages 484–502, 2015.

[Dan16] A. Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48$^{th}$ Annual Symposium on Theory of Computing, STOC 2016*, pages 105–117, 2016.

[DBW12] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

[DDK+23a] I. Diakonikolas, J. Diakonikolas, D. M. Kane, P. Wang, and N. Zarifis. Information-computation tradeoffs for learning margin halfspaces with random classification noise. In *COLT*, 2023.

[DDK+23b] I. Diakonikolas, J. Diakonikolas, D. M. Kane, P. Wang, and N. Zarifis. Near-optimal bounds for learning gaussian halfspaces with random classification noise. In *NeurIPS*, 2023.

[DFT+15] D. Dachman-Soled, V. Feldman, L.-Y. Tan, A. Wan, and K. Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In Piotr Indyk, editor, *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 498–511. SIAM, 2015.

[DG22] A. Daniely and E. Granot. An exact poly-time membership-queries algorithm for extracting a three-layer relu network. In *The Eleventh International Conference on Learning Representations*, 2022.

[DG24] J. Diakonikolas and C. Guzmán. Optimization on a finer scale: Bounded local subgradient variation perspective. *arXiv preprint arXiv:2403.16317*, 2024.

[DGJ+10] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010.

[DGK+20] I. Diakonikolas, S. Goel, S. Karmalkar, A. R. Klivans, and M. Soltanolkotabi. Approximation schemes for ReLU regression. In *Conference on Learning Theory, COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 1452–1485. PMLR, 2020.

[DGT19] I. Diakonikolas, T. Gouleakis, and C. Tzamos. Distribution-independent pac learning of halfspaces with massart noise. In *Advances in Neural Information Processing Systems 32, NeurIPS*. 2019.

[DH18] R. Dudeja and D. Hsu. Learning single-index models in Gaussian space. In *Conference on Learning Theory, COLT*, volume 75 of *Proceedings of Machine Learning Research*, pages 1887–1930. PMLR, 2018.

[DHK+10] I. Diakonikolas, P. Harsha, A. Klivans, R. Meka, P. Raghavendra, R. A. Servedio, and L. Y. Tan. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *STOC*, pages 533–542, 2010.

[DIK+21] I. Diakonikolas, R. Impagliazzo, D. M. Kane, R. Lei, J. Sorrell, and C. Tzamos. Boosting in the presence of massart noise. In *Proceedings of The 34$^{th}$ Conference on Learning Theory, COLT*, 2021.

[DIKZ25] I. Diakonikolas, G. Iakovidis, D. M. Kane, and N. Zarifis. Robust learning of multi-index models via iterative subspace approximation, 2025.

[DJS08] A. S. Dalalyan, A. Juditsky, and V. Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *The Journal of Machine Learning Research*, 9:1647–1678, 2008.

[DJS+15] I. Diakonikolas, R. Jaiswal, R. A. Servedio, L. Y. Tan, and A. Wan. Noise stable halfspaces are close to very small juntas. *Chicago Journal OF Theoretical Computer Science*, 4:1–13, 2015.

[DK20a] I. Diakonikolas and D. M. Kane. Hardness of learning halfspaces with massart noise. *CoRR*, abs/2012.09720, 2020.

[DK20b] I. Diakonikolas and D. M Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 184–195. IEEE, 2020.

[DK23a] I. Diakonikolas and D. M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.

[DK23b] I. Diakonikolas and D. M. Kane. Efficiently learning one-hidden-layer relu networks via schur polynomials. *CoRR*, abs/2307.12840, 2023.

[DKK+20] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. A polynomial time algorithm for learning halfspaces with tsybakov noise. *arXiv*, 2020.

[DKK+21a] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Agnostic proper learning of halfspaces under gaussian marginals. In *Proceedings of The 34$^{th}$ Conference on Learning Theory, COLT*, 2021.

[DKK+21b] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Efficiently learning halfspaces with tsybakov noise. *STOC*, 2021.

[DKK+22] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Learning general halfspaces with general Massart noise under the Gaussian distribution. In *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 874–885. ACM, 2022.

[DKK+24] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Agnostically learning multi-index models with queries. In *FOCS*, 2024.

[DKKZ20] I. Diakonikolas, D. M. Kane, V. Kontonis, and N. Zarifis. Algorithms and SQ lower bounds for pac learning one-hidden-layer ReLU networks. In *Conference on Learning Theory, COLT*, pages 1514–1539. PMLR, 2020.

[DKMR22a] I. Diakonikolas, D. Kane, P. Manurangsi, and L. Ren. Cryptographic hardness of learning halfspaces with Massart noise. In *Advances in Neural Information Processing Systems*, 2022.

[DKMR22b] I. Diakonikolas, D. Kane, P. Manurangsi, and L. Ren. Hardness of learning a single neuron with adversarial label noise. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[DKN10] I. Diakonikolas, D. M. Kane, and J. Nelson. Bounded independence fools degree-2 threshold functions. In *FOCS*, pages 11–20, 2010.

[DKPZ21] I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. The optimality of polynomial regression for agnostic learning under Gaussian marginals in the SQ model. In *Proceedings of The 34$^{th}$ Conference on Learning Theory, COLT*, 2021.

[DKR23] I. Diakonikolas, D. M. Kane, and L. Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and ReLU regression under Gaussian marginals. In *ICML*, 2023.

[DKRS22] I. Diakonikolas, D. Kane, L. Ren, and Y. Sun. SQ lower bounds for learning single neurons with Massart noise. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022.

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58$^{th}$ Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84, 2017.

[DKS18] I. Diakonikolas, D. M. Kane, and A. Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50$^{th}$ Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1061–1073, 2018.

[DKT21] I. Diakonikolas, D. Kane, and C. Tzamos. Forster decomposition and learning halfspaces with noise. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 7732–7744, 2021.

[DKTZ20a] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory, COLT*, 2020.

[DKTZ20b] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning halfspaces with tsybakov noise. *arXiv*, 2020.

[DKTZ20c] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Non-convex SGD learns halfspaces with adversarial label noise. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[DKTZ22a] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning a single neuron with adversarial label noise via gradient descent. In *Conference on Learning Theory (COLT)*, pages 4313–4361, 2022.

[DKTZ22b] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning general halfspaces with adversarial label noise via online gradient descent. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5118–5141. PMLR, 17–23 Jul 2022.

[DKTZ24] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Online Linear Classification with Massart Noise, 2024. Arxiv eprint: 2405.12958.

[DKZ20] I. Diakonikolas, D. M. Kane, and N. Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under Gaussian marginals. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[DLS22] A. Damian, J. Lee, and M. Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*. PMLR, 2022.

[DMN21] A. De, E. Mossel, and J. Neeman. Robust testing of low dimensional functions. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 584–597. ACM, 2021.

[DPT21] I. Diakonikolas, J. Park, and C. Tzamos. Relu regression with Massart noise. In *Advances in Neural Information Processing Systems*, 2021.

[DPVLB24] A. Damian, L. Pillaud-Vivien, J. Lee, and J. Bruna. Computational-statistical gaps in gaussian single-index models. In *The Thirty Seventh Annual Conference on Learning Theory*. PMLR, 2024.

[DRST14] I. Diakonikolas, P. Raghavendra, R. A. Servedio, and L. Y. Tan. Average sensitivity and noise sensitivity of polynomial threshold functions. *SIAM J. Comput.*, 43(1):231–253, 2014.

[DSTW10] I. Diakonikolas, R. Servedio, L.-Y. Tan, and A. Wan. A regularity lemma, and low-weight approximators, for low-degree polynomial threshold functions. In *CCC*, pages 211–222, 2010.

[DTK22] I. Diakonikolas, C. Tzamos, and D. M. Kane. A strongly polynomial algorithm for approximate forster transforms and its application to halfspace learning. In *STOC*, 2022.

[DV04] J. Dunagan and S. Vempala. Optimal outlier removal in high-dimensional spaces. *J. Computer & System Sciences*, 68(2):335–373, 2004.

[DZ24] I. Diakonikolas and N. Zarifis. A near-optimal algorithm for learning margin halfspaces with massart noise. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Fan68] K. Fan. On infinite systems of linear inequalities. *Journal of Mathematical Analysis and Applications*, 21(3):475–478, 1968.

[FCG20] S. Frei, Y. Cao, and Q. Gu. Agnostic learning of a single neuron with gradient descent. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[FCG21] S. Frei, Y. Cao, and Q. Gu. Agnostic learning of halfspaces with gradient descent via soft margins. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021.

[Fed69] H. Federer. *Geometric Measure Theory*. Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen. Springer, 1969.

[Fel08] V. Feldman. On the power of membership queries in agnostic learning. In *21st Annual Conference on Learning Theory - COLT 2008*, pages 147–156, 2008.

[Fel16] V. Feldman. Statistical query learning. In *Encyclopedia of Algorithms*, pages 2090–2095. 2016.

[Fel17] V. Feldman. A general characterization of the statistical query complexity. In *Proceedings of the 30$^{th}$ Conference on Learning Theory, COLT 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 785–830. PMLR, 2017.

[FGKP06] V. Feldman, P. Gopalan, S. Khot, and A. Ponnuswami. New results for learning noisy parities and halfspaces. In *Proc. FOCS*, pages 563–576, 2006.

[FGR$^+$17] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *J. ACM*, 64(2):8:1–8:37, 2017.

[FGV17] V. Feldman, C. Guzman, and S. S. Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In Philip N. Klein, editor, *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017*, pages 1265–1277. SIAM, 2017.

[FJS81] J. H. Friedman, M. Jacobson, and W. Stuetzle. Projection Pursuit Regression. *J. Am. Statist. Assoc.*, 76:817, 1981.

[FP03] F. Facchinei and J-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

[FPV15] V. Feldman, W. Perkins, and S. Vempala. On the complexity of random satisfiability problems with planted solutions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC, 2015*, pages 77–86, 2015.

[FS97] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[Gan02] M. Ganzburg. Limit theorems for polynomial approximations with Hermite and Freud weights. In *Approximation theory, X (St. Louis, MO, 2001)*, Innov. Appl. Math., pages 211–221. Vanderbilt Univ. Press, Nashville, TN, 2002.

[Gan08] M. Ganzburg. Limit theorems of polynomial approximation with exponential weights. *Mem. Amer. Math. Soc.*, 192, 2008.

[GGJ+20] S. Goel, A. Gollakota, Z. Jin, S. Karmalkar, and A. Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning, ICML*, 2020.

[GGK20] S. Goel, A. Gollakota, and A. R. Klivans. Statistical-query lower bounds via functional gradients. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[GGKS23] A. Gollakota, P. Gopalan, A. R. Klivans, and K. Stavropoulos. Agnostically learning single-index models using omnipredictors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[GHR92] M. Goldmann, J. Håstad, and A. Razborov. Majority gates vs. general weighted threshold gates. *Computational Complexity*, 2:277–300, 1992.

[GKK08a] P. Gopalan, A. Kalai, and A. Klivans. Agnostically learning decision trees. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 527–536, 2008.

[GKK08b] P. Gopalan, A. Kalai, and A. R. Klivans. A query algorithm for agnostically learning dnf? In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 515–516, 2008.

[GKK19] S. Goel, S. Karmalkar, and A. R. Klivans. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, 2019.

[GKKT17] S. Goel, V. Kanade, A. R. Klivans, and J. Thaler. Reliably learning the relu in polynomial time. In *Proceedings of the 30$^{th}$ Conference on Learning Theory, COLT 2017*, pages 1004–1042, 2017.

[GKLW19] R. Ge, R. Kuditipudi, Z. Li, and X. Wang. Learning two-layer neural networks with symmetric inputs. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[GL89] O. Goldreich and L. Levin. A hard-core predicate for all one-way functions. In *Proceedings of the Twenty-First Annual Symposium on Theory of Computing*, pages 25–32, Seattle, Washington, 1989.

[GLM18] R. Ge, J. D. Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[GR06] V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. In *Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 543–552. IEEE Computer Society, 2006.

[GV24] A. Guo and A. Vijayaraghavan. Agnostic learning of arbitrary ReLU activation under Gaussian marginals. *arXiv preprint arXiv:2411.14349*, 2024.

[GWBL20] S. Garg, Y. Wu, S. Balakrishnan, and Z. Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*, 2020.

[Han09] S. Hanneke. Theoretical foundations of active learning. Technical report, 2009.

[Hau92] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[Haz16] E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[HJS01] M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

[HKM14] P. Harsha, A. R. Klivans, and R. Meka. Bounding the sensitivity of polynomial threshold functions. *Theory of Computing*, 10:1–26, 2014.

[HL93] P. Hall and K.-C. Li. On almost Linearity of Low Dimensional Projections from High Dimensional Data. *The Annals of Statistics*, 21(2):867 – 889, 1993.

[HMS+04] W. Härdle, M. Müller, S. Sperlich, A. Werwatz, et al. *Nonparametric and semiparametric models*, volume 1. Springer, 2004.

[Hof52] A. J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49:263–265, 1952.

[HSSV22] D. J. Hsu, C. Sanford, R. A. Servedio, and E. Vlatakis-Gkaragkounis. Near-optimal statistical query lower bounds for agnostically learning intersections of halfspaces with gaussian marginals. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 283–312. PMLR, 2022.

[HTY24] L. Hu, K. Tian, and C. Yang. Omnipredicting single-index models with multi-index models. *arXiv preprint arXiv:2411.13083*, 2024.

[Hub85] P. J. Huber. Projection Pursuit. *The Annals of Statistics*, 13(2):435 – 475, 1985.

[HY15] S. Hanneke and L. Yang. Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16:3487–3602, 2015.

[Ich93] H. Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of econometrics*, 58(1-2):71–120, 1993.

[Jac97] J. Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55(3):414–440, 1997.

[JCB+20] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX Security Symposium, USENIX Security 2020, 2020*, pages 1345–1362. USENIX Association, 2020.

[JGN+17] C. Jin, R. Ge, P. Netrapalli, S. Kakade, and M. Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

[JL84] W. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[JSA15a] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods, 2015.

[JSA15b] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[JWZ20] R. Jayaram, D.P. Woodruff, and Q. Zhang. Span recovery for deep neural networks with applications to input obfuscation. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[Kan11] D. M. Kane. The gaussian surface area and noise sensitivity of degree-$d$ polynomial threshold functions. *Computational Complexity*, 20(2):389–412, 2011.

[Kan14] D. M. Kane. The average sensitivity of an intersection of half spaces. In *Symposium on Theory of Computing, STOC 2014*, pages 437–440, 2014.

[KB09] B. B. Klebanov and E. Beigman. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503, 2009.

[KB10] B. B. Klebanov and E. Beigman. Some empirical evidence for annotation noise in a benchmarked dataset. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 438–446, 2010.

[Kea98] M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

[KIT+23] V. Kontonis, F. Iliopoulos, K. Trinh, C. Baykal, G. Menghani, and E. Vee. Slam: Student-label mixing for distillation with unlabeled examples. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023.

[KKMS08] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008. Special issue for FOCS 2005.

[KKSK11] S. M. Kakade, V. Kanade, O. Shamir, and A. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011.

[KLS09] A. Klivans, P. Long, and R. Servedio. Learning Halfspaces with Malicious Noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009.

[KLT09] A. R. Klivans, P. M. Long, and A. K. Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *13th International Workshop, RANDOM 2009*, pages 588–600, 2009.

[KM93] E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *SIAM J. on Computing*, 22(6):1331–1348, December 1993.

[KNS16] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811, 2016.

[KOS04] A. Klivans, R. O'Donnell, and R. Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer & System Sciences*, 68(4):808–840, 2004.

[KOS08] A. Klivans, R. O'Donnell, and R. Servedio. Learning geometric concepts via Gaussian surface area. In *Proc. 49th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 541–550, Philadelphia, Pennsylvania, 2008.

[KS09] A. T. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

[KSS94] M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2/3):115–141, 1994.

[KSST08] S. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, 2008.

[KTZ19] V. Kontonis, C. Tzamos, and M. Zampetakis. Efficient truncated statistics with unknown truncation. In *2019 IEEE 60$^{th}$ Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595. IEEE, 2019.

[LCP22] J. Liu, Y. Cui, and J-S. Pang. Solving nonsmooth and nonconvex compound stochastic programs with applications to risk measure minimization. *Mathematics of Operations Research*, 2022.

[Led94] M. Ledoux. Semigroup proofs of the isoperimetric inequality in Euclidean and Gauss space. *Bull. Sci. Math.*, 118:485–510, 1994.

[LH22] C. Lu and D. S. Hochbaum. A unified approach for a 1D generalized total variation problem. *Mathematical Programming*, 194(1-2):415–442, 2022.

[Li91] K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[LKDM24] C.-H. Lin, C. Kaushik, E. L. Dyer, and V. Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *Journal of Machine Learning Research*, 25(91):1–85, 2024.

[Łoj63] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[Łoj93] S. Łojasiewicz. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pages 1575–1595, 1993.

[Lon95] P. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

[LV07] L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007.

[LWS18] Z. Lipton, Y. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2018.

[MBM18] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[MN06] P. Massart and E. Nedelec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, October 2006.

[MP68] M. Minsky and S. Papert. *Perceptrons: an introduction to computational geometry*. MIT Press, Cambridge, MA, 1968.

[MR18] P. Manurangsi and D. Reichman. The computational complexity of training ReLU(s). *arXiv preprint arXiv:1810.04207*, 2018.

[MSDH19] S. Milli, L. Schmidt, A. D. Dragan, and M. Hardt. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT 2019, Atlanta, GA, USA, 2019*, pages 1–9. ACM, 2019.

[MT94] W. Maass and G. Turan. How fast can a threshold gate learn? In S. Hanson, G. Drastal, and R. Rivest, editors, *Computational Learning Theory and Natural Learning Systems*, pages 381–414. MIT Press, 1994.

[MT99] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, December 1999.

[MV19] O. Mangoubi and N. K. Vishnoi. Nonconvex sampling with the metropolis-adjusted langevin algorithm. In *Conference on Learning Theory, COLT 2019*, 2019.

[MW06] S. Mukherjee and Q. Wu. Estimation of gradients and coordinate covariation in classification. *The Journal of Machine Learning Research*, 7:2481–2514, 2006.

[MZST06] S. Mukherjee, D. Zhou, and J. Shawe-Taylor. Learning coordinate covariances via gradients. *Journal of Machine Learning Research*, 7(3), 2006.

[NDRT13] N. Natarajan, I. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, 2013.

[Nee14] J. Neeman. Testing surface area with arbitrary accuracy. In *Symposium on Theory of Computing, STOC 2014, 2014*, pages 393–397. ACM, 2014.

[Nel73] E. Nelson. The free markoff field. *Journal of Functional Analysis*, 12(2):211–227, 1973.

[Nov62] A. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

[NS17] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

[NT22] R. Nasser and S. Tiegel. Optimal SQ lower bounds for learning halfspaces with massart noise. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1047–1074. PMLR, 2022.

[O'D14] R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[Pan97] J.-S. Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1):299–332, 1997.

[Pis86] G. Pisier. Probabilistic methods in the geometry of Banach spaces. In *Lecture notes in Math.*, pages 167–241. Springer, 1986.

[PMG+17] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017*, pages 506–519. ACM, 2017.

[PRKM+17] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L/ Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[Rd17] V. Roulet and A. d'Aspremont. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017.

[RK20] D. Rolnick and K. P. Kording. Reverse-engineering deep ReLU networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 8178–8187. PMLR, 2020.

[Ros58] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[RS94] R. Rivest and R. Sloan. A formal model of hierarchical concept learning. *Information and Computation*, 114(1):88–114, 1994.

[RS10] P. Raghavendra and D. Steurer. Graph expansion and the unique games conjecture. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010*, pages 755–764. ACM, 2010.

[Sím02] J. Síma. Training a single sigmoidal neuron is hard. *Neural Computation*, 14(11):2709–2728, 2002.

[SKP+20] Hwanjun S., Minseok K., Dongmin P., Yooju S., and Jae-Gil L. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[Slo88] R. H. Sloan. Types of noise in data for concept learning. In *Proceedings of the First Annual Workshop on Computational Learning Theory*, COLT '88, pages 91–96, San Francisco, CA, USA, 1988. Morgan Kaufmann Publishers Inc.

[Slo92] R. H. Sloan. Corrigendum to types of noise in data for concept learning. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992*, page 450, 1992.

[Slo96] R. H. Sloan. *Pac Learning, Noise, and Geometry*, pages 21–41. Birkhäuser Boston, Boston, MA, 1996.

[Sol17] M. Soltanolkotabi. Learning ReLUs via gradient descent. In *Advances in neural information processing systems*, pages 2007–2017, 2017.

[SSBD14] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[SSG17] Y. Shi, Y. Sagduyu, and A. Grushin. How to steal a machine learning classifier with deep learning. In *2017 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–5, 2017.

[SST10] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

[STC00] J. Shawe-Taylor and N. Cristianini. *An introduction to support vector machines*. Cambridge University Press, 2000.

[Ste81] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

[SZ07] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[Sze67] G. Szegö. *Orthogonal Polynomials*. Number $\tau$. 23 in American Mathematical Society colloquium publications. American Mathematical Society, 1967.

[Szö09] B. Szörényi. Characterizing statistical query learning: Simplified notions and proofs. In Ricard Gavaldà, Gábor Lugosi, Thomas Zeugmann, and Sandra Zilles, editors, *Algorithmic Learning Theory, 20$^{th}$ International Conference, ALT 2009*, volume 5809 of *Lecture Notes in Computer Science*, pages 186–200. Springer, 2009.

[Tie23] S. Tiegel. Hardness of agnostically learning halfspaces from worst-case lattice problems. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 2023.

[Tsy04] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[Tsy08] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

[TZJ+16] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction apis. In *25th USENIX Security Symposium, USENIX Security 16, 2016*, pages 601–618. USENIX Association, 2016.

[Val84] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vap82] V. Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics*. Springer-Verlag, Berlin, Heidelberg, 1982.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

[Vem10a] S. Vempala. Learning convex concepts from gaussian distributions with PCA. In *51$^{th}$ Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 124–130, 2010.

[Vem10b] S. Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *J. ACM*, 57(6):32:1–32:14, 2010.

[Ver18] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[VRW17] Brendan Van R. and Robert C. W. A theory of learning with corrupted labels. *J. Mach. Learn. Res.*, 2017.

[WGMM10] Q. Wu, J. Guinney, M. Maggioni, and S. Mukherjee. Learning gradients: predictive models that infer geometry and statistical dependence. *Journal of Machine Learning Research*, 11:2175–2198, 2010.

[WLT18] R. Wang, T. Liu, and D. Tao. Multiclass learning with partially corrupted labels. *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[WZDD23] P. Wang, N. Zarifis, I. Diakonikolas, and J. Diakonikolas. Robustly learning a single neuron via sharpness. *40th International Conference on Machine Learning*, 2023.

[WZDD24] P. Wang, N. Zarifis, I. Diakonikolas, and J. Diakonikolas. Sample and computationally efficient robust learning of gaussian single-index models. *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

[WZDD25] P. Wang, N. Zarifis, I. Diakonikolas, and J. Diakonikolas. Robustly learning monotone single-index models. 2025.

[Xia08] Y. Xia. A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484):1631–1640, 2008.

[XTLZ02] Y. Xia, H. Tong, W. K. Li, and L. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):363–410, 2002.

[Yao90] A. Yao. On ACC and threshold circuits. In *Proceedings of the Thirty-First Annual Symposium on Foundations of Computer Science*, pages 619–627, 1990.

[YGLS+19] D. Yin, R. Gontijo Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

[YS20] G. Yehudai and O. Shamir. Learning a single neuron with gradient methods. In *Conference on Learning Theory, COLT*, 2020.

[YZ17] S. Yan and C. Zhang. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 1056–1066, 2017.

[ZFG21] D. Zou, S. Frei, and Q. Gu. Provable robustness of adversarial training for learning halfspaces with noise. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021.

[ZL16] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[ZL21] C. Zhang and Y. Li. Improved algorithms for efficient active learning halfspaces with massart and tsybakov noise. In *Proceedings of The 34$^{th}$ Conference on Learning Theory, COLT*, 2021.

[ZLA21] M. Zhang, J. Lee, and S. Agarwal. Learning from noisy labels with no change to the training process. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2021.

[ZLC17] Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Proceedings of the 30$^{th}$ Conference on Learning Theory, COLT 2017*, pages 1980–2022, 2017.

[ZSA20] C. Zhang, J. Shen, and P. Awasthi. Efficient active learning of sparse halfspaces with arbitrary bounded noise. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[ZWDD24] N. Zarifis, P. Wang, I. Diakonikolas, and J. Diakonikolas. Robustly learning single-index models via alignment sharpness. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58197–58243. PMLR, 21–27 Jul 2024.

[ZWDD25] N. Zarifis, P. Wang, I. Diakonikolas, and J. Diakonikolas. Robustly learning monotone generalized linear models via data augmentation. In *COLT*, 2025.

[ZY13] H. Zhang and W. Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

# Appendix A

# Appendix to Chapter 2

## A.1 "Massart" Noise with $\eta > 1/2$ is Equivalent to Agnostic Learning

**Lemma A.1.1** (Learning with $\eta > 1/2$)**.** *Let $\mathcal{C}$ be a class of Boolean functions on $\mathbb{R}^d$ and let $\mathcal{F}$ be a class of distributions over $\mathbb{R}^d$. Fix $\eta \in (1/2, 1]$ and let $\mathcal{A}$ be a learning algorithm that given $m$ samples from a distribution $\mathcal{D}'$ with $\eta$-semi-random noise, learns a hypothesis $h : \mathbb{R}^d \mapsto \{\pm 1\}$ such that*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}'}[h(\mathbf{x}) \neq y] \leq \min_{c\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}'}[c(\mathbf{x}) \neq y] + \epsilon\,.$$

*Then $\mathcal{A}$ can learn $\mathcal{C}$ in the agnostic PAC learning model.*

*Proof.* Let $\mathcal{D}$ be any distribution on $\mathbb{R}^d \times \{\pm 1\}$. Since $\eta > 1/2$, we can create a distribution $\mathcal{D}'$ with $\eta$-semi-random noise as follows:

1. Draw $(\mathbf{x}, y) \sim \mathcal{D}$.
2. With probability $2\eta - 1$ return $(\mathbf{x}, y)$ and with probability $2(1-\eta)$ return $(\mathbf{x}, \hat{y})$ where $\hat{y} \in \{\pm 1\}$ is uniformly random and independent of $\mathbf{x}$.

For any $h \in \mathcal{C}$ (in fact for any classifier in general), it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}'}[h(\mathbf{x}) = y|\mathbf{x}] \geq 2(1-\eta)/2 \geq 1-\eta$ and thus we have that $\eta(\mathbf{x}) \leq \eta$. The classifier $f$ of Definition 1.3.1 can therefore be any classifier in $\mathcal{C}$. For every classifier $h$ it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}'}[h(\mathbf{x}) \neq y] = (2\eta - 1) \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] + 2(\eta - 1)(1/2)\,. \tag{A.1}$$

We can therefore, use $\mathcal{A}$ on the samples from $\mathcal{D}'$ and obtain a classifier $h$ such that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}'}[h(\mathbf{x}) = y] \leq \min_{c\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}'}[c(\mathbf{x}) = y] + \epsilon\,.$$

Using Equation (A.1) we obtain that for the same classifier $h$ it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) = y] \leq \min_{c\in\mathcal{C}} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[c(\mathbf{x}) = y] + \epsilon\,,$$

and therefore $\mathcal{A}$ can learn the class $\mathcal{C}$ with respect to any distribution $\mathcal{D}$, i.e., in the agnostic PAC learning setting. □

## A.2 Benign Noise is Equivalent to $1/2$-Massart Noise

**Fact A.2.1.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$. $\mathcal{D}$ satisfies the $1/2$-Massart noise condition with respect to a halfspace $f(\mathbf{x})$ if only if the distribution $\mathcal{D}$ satisfies the Benign noise condition with respect to a halfspace $f(\mathbf{x})$.*

*Proof.* We prove each direction separately.

**Claim A.2.2.** *If $\mathcal{D}$ satisfies the Massart noise condition with $\eta = 1/2$ with respect to a halfspace $f(\mathbf{x})$, then $\mathcal{D}$ satisfies the Benign noise condition with respect to the halfspace $f(\mathbf{x})$.*

*Proof.* Assume in order to reach in contradiction that the optimal classifier is not $f$ and is $h$ the optimal one. Therefore, it holds $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] < \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y]$. It holds that

$$\begin{aligned}\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1-\eta(\mathbf{x}))] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) = f(\mathbf{x})\}\eta(\mathbf{x})] \\ &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1-2\eta(\mathbf{x}))] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})]\,.\end{aligned}$$

Note that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})]$, therefore, we have

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1-2\eta(\mathbf{x}))] < 0\,,$$

which means that there is a point (a mass with non-zero measure) with $\eta(\mathbf{x}) > 1/2$, which leads to a contradiction. Therefore, if $\eta(\mathbf{x}) \leq 1/2$ then the optimal classifier is $f(\mathbf{x})$.

□

Next, we prove the other direction.

**Claim A.2.3.** *If $\mathcal{D}$ satisfies the Benign noise condition with respect to a halfspace $f(\mathbf{x})$, then $\mathcal{D}$ satisfies the $1/2$-Massart noise condition with respect to the halfspace $f(\mathbf{x})$.*

*Proof.* It suffices to prove that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\eta(\mathbf{x}) > 1/2\}] = 0$. Because $f(\mathbf{x})$ is the optimal classifier, that means that other classifier $h(\mathbf{x})$ gets more error, therefore it holds

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1-2\eta(\mathbf{x}))] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\eta(\mathbf{x})] \,.$$

Because $f(\mathbf{x})$ is the optimal classifier, it holds that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x}) \neq y] \geq \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[f(\mathbf{x}) \neq y]$, therefore

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1-2\eta(\mathbf{x}))] \geq 0 \,.$$

Assume that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\eta(\mathbf{x}) > 1/2\}] = a > 0$, therefore we can set $h(\mathbf{x}) = f(\mathbf{x})$ when $\eta(\mathbf{x}) \leq 1/2$ and $h(\mathbf{x}) \neq f(\mathbf{x})$ otherwise. Hence, we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{h(\mathbf{x}) \neq f(\mathbf{x})\}(1-2\eta(\mathbf{x}))] = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(1-2\eta(\mathbf{x}))\mathbb{1}\{\eta(\mathbf{x}) > 1/2\}] < 0$, which is a contradiction. Therefore, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbb{1}\{\eta(\mathbf{x}) > 1/2\}] = 0$, that means that the measure of points that $\eta(\mathbf{x}) > 1/2$ is 0, hence, it satisfies the 1/2-Massart noise condition. □

□

## A.3 A Non-Continuous Certificate

**Non-Continuous Certificates.** When we do not restrict our search to continuous functions in order to find a certificate we can search over functions of the form $T(x) = \mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq \mathrm{sign}(\mathbf{v}\cdot\mathbf{x} - b)\}$ for some $\mathbf{v} \in \mathbb{R}^d, b \in \mathbb{R}$.

**Fact A.3.1.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the $\eta$-Massart noise condition with respect to the optimal halfspace $f(\mathbf{x})$. Then, for any linear function $\ell(\mathbf{x})$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\ell(\mathbf{x})) \neq y] \geq \mathrm{OPT} + \epsilon$, let $T(\mathbf{x}) = \mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})\}$, then we have that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y\ T(\mathbf{x})] \leq -\Omega(\epsilon^2)\|\ell(\mathbf{x})\|_2 \,.$$

*Proof.* Using the $\eta$-Massart noise condition, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(\mathbf{x})y\ T(\mathbf{x})] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\ell(\mathbf{x})f(\mathbf{x})\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})\}] \\ &= -\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\ell(\mathbf{x})|\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})\}] \,.\end{aligned}$$

From Carbery-Wright [CW01], we have $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\ell(\mathbf{x})| \geq \|\ell(\mathbf{x})\|_2\epsilon/C] \geq 1-\epsilon/2$, for some absolute constant $C > 0$. From the assumptions, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})\}] \geq \epsilon$, therefore it holds

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\{|\ell(\mathbf{x})| \geq \|\ell(\mathbf{x})\|_2\epsilon/c\}\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x})) \neq f(\mathbf{x})\}] \geq \epsilon/2 \,,$$

and we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}&[|\ell(\mathbf{x})|\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x}))\neq f(\mathbf{x})\}]\\ &\geq (\|\ell(\mathbf{x})\|_2\epsilon/C)\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\{|\ell(\mathbf{x})|\geq\|\ell(\mathbf{x})\|_2\epsilon/c\}\beta(\mathbf{x})\mathbb{1}\{\mathrm{sign}(\ell(\mathbf{x}))\neq f(\mathbf{x})\}]\\ &\gtrsim \|\ell(\mathbf{x})\|_2\epsilon^2/C\,.\end{aligned}$$

And this completes the proof. □

## A.4 Learning General Halfspaces with General Massart Noise

In this section, we provide the algorithm of learning biased halfspaces with general Massart Noise.

**Theorem A.4.1** (Learning General Halfspaces with General Massart Noise)**.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta=1/2$ with respect to some optimal (possibly biased) halfspace $f\in\mathcal{C}$. Let $\epsilon,\delta\in(0,1]$. There exists an algorithm that draws $N=d^{O(\log(1/\epsilon))}\log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N,d)2^{\mathrm{poly}(1/\epsilon)}$, and computes a halfspace $h\in\mathcal{C}$ such that with probability at least $1-\delta$,*

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[h(\mathbf{x})\neq y]\leq \mathrm{OPT}+\epsilon\,.$$

The proposition below is similar to Proposition 2.4.2.

**Proposition A.4.2.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition for $\eta=1/2$, with respect to some optimal halfspace $f(\mathbf{x})=\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t^*)$. Let $\mathbf{w}\in\mathbb{R}^d$ be a unit vector and $t\in\mathbb{R}$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t)\neq y]\geq\mathrm{OPT}+\epsilon$ and $\theta(\mathbf{w},\mathbf{w}^*)\leq\pi-\epsilon$, for some $\epsilon\in(0,1]$. There exists an algorithm that draws $N=d^{O(\log(1/\epsilon))}\log(1/\delta)$ samples from $\mathcal{D}$, runs in time $\mathrm{poly}(N,d)$ and, with probability at least $1-\delta$ returns a basis of a subspace $V\subseteq\mathbf{w}^\perp$ such that $\|\mathrm{proj}_V((\mathbf{w}^*)^{\perp\mathbf{w}})\|_2=\mathrm{poly}(\epsilon)$.*

The proof of Proposition A.4.2 is the same as Proposition 2.4.2 with the only difference be that, instead of Lemma 2.4.6 we use the lemma below.

**Lemma A.4.3.** *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d\times\{\pm1\}$, with standard normal $\mathbf{x}$-marginal, that satisfies the Massart noise condition with $\eta=1/2$ with respect to $f(\mathbf{x})=\mathrm{sign}(\mathbf{w}^*\cdot\mathbf{x}+t)$ with $t\in\mathbb{R}$. Fix $\epsilon>0$ and $\rho>0$ such that $\rho\lesssim\epsilon$. Let $\mathbf{w}$ be a unit vector such that for any $t'\in\mathbb{R}$, it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\beta(\mathbf{x})\mathbb{1}\{f(\mathbf{x})\neq\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t')\}]\geq\epsilon$.*

*For $t_1, t_2 \in \mathbb{R}$ consider the band $B = \{t_1 \leq \mathbf{w} \cdot \mathbf{x} \leq t_2\}$. Denote $\mathcal{D}^{\perp} = \mathcal{D}_B^{\perp_{\mathbf{w}}}$, i.e., $\mathcal{D}$ is the orthogonal projection onto $\mathbf{w}^{\perp}$ of the conditional distribution on $B$, and consider the halfspace $f^{\perp} : \mathbf{w}^{\perp} \mapsto \{\pm 1\}$, with $f^{\perp}(\mathbf{x}) = \mathrm{sign}(\mathbf{x} \cdot (\mathbf{w}^*)^{\perp_{\mathbf{w}}} - b)$, for some threshold $b \in \mathbb{R}$. Moreover, define the noise function*

$$\eta^{\perp}(\mathbf{x}) = \mathop{\mathbf{Pr}}_{(\mathbf{z},y)\sim\mathcal{D}^{\perp}}[y \neq f^{\perp}(\mathbf{z}) | \mathbf{z} = \mathbf{x}] \,.$$

*There exist $t_1, t_2 \in \mathbb{R}$ multiples of $\rho$, with $|t_1 - t_2| = \rho$, such that:*

- $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}^{\perp}}[(1 - 2\eta^{\perp}(\mathbf{x}))\mathbb{1}\{f^{\perp}(\mathbf{x})\mathrm{sign}(b) > 0\}] \gtrsim \epsilon/\sqrt{\log(1/\epsilon)} - \rho/\epsilon$,
- *if $\eta^{\perp}(\mathbf{x}) > 1/2$, then $0 < \mathrm{sign}(b)(\mathbf{x} \cdot (\mathbf{w}^*)^{\perp_{\mathbf{w}}} - b) \leq \rho/\epsilon$.*

*Proof.* First, we consider the region $B = \{t_1 \leq \mathbf{w} \cdot \mathbf{x} \leq t_2\}$, for any $t_1, t_2 \in \mathbb{R}$ with $|t_1 - t_2| = \rho$. We denote by $\mathbf{x}^{\perp}$ the projection of $\mathbf{x}$ onto the subspace $\mathbf{w}^{\perp}$. Define the distribution $\mathcal{D}^{\perp} = \mathcal{D}_B^{\mathrm{proj}_{\mathbf{w}^{\perp}}}$, that is the distribution $\mathcal{D}$ conditioned on the set $B$ and projected onto $\mathbf{w}^{\perp}$, the hypothesis $f^{\perp}(\mathbf{x}^{\perp}) = \mathrm{sign}(\mathbf{x}^{\perp} \cdot (\mathbf{w}^*)^{\perp_{\mathbf{w}}} - b)$ where $b \in \mathbb{R}$ is chosen appropriately below, and the noise function $\eta^{\perp}(\mathbf{x}^{\perp}) = \mathbf{Pr}_{(\mathbf{z},y)\sim\mathcal{D}^{\perp}}[y \neq f^{\perp}(\mathbf{z}) | \mathbf{z} = \mathbf{x}^{\perp}]$.

Note that the distribution $\mathcal{D}^{\perp}$ does not satisfy the 1/2-Massart noise condition anymore. We first illustrate how the noise function changes. The orthogonal projection on $\mathbf{w}^{\perp}$ creates a region where the Massart condition is violated, i.e., a region where $\eta^{\perp}(\mathbf{x}^{\perp}) \geq 1/2$, but we can control the probability that we get points inside this region. More formally, we show that$\mathbf{Pr}_{(\mathbf{x}^{\perp},y)\sim\mathcal{D}^{\perp}}[\eta^{\perp}(\mathbf{x}^{\perp}) \geq 1/2] \lesssim \rho/\epsilon$.

To show that first notice that $\theta(\mathbf{v}, \mathbf{w}^*) \geq \epsilon$, otherwise we would have $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w} \cdot \mathbf{x})] \leq \epsilon$. We can assume that $\mathbf{w}^* = \lambda_1 \mathbf{w} + \lambda_2 (\mathbf{w}^*)^{\perp_{\mathbf{w}}}$, where $\lambda_1 = \cos\theta$ and $\lambda_2 = \sin\theta$. Next we set $\mathbf{x} = (\mathbf{x}_{\mathbf{w}}, \mathbf{x}^{\perp})$, where $\mathbf{x}_{\mathbf{w}} = \mathbf{w} \cdot \mathbf{x}$. We show that the hypothesis $f^{\perp}(\mathbf{x}) = \mathrm{sign}((\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{x}^{\perp} + (t + t_1\lambda_1)/\lambda_2)$ is almost as good as the $f(\mathbf{x})$ for the distribution $\mathcal{D}^{\perp}$.

Conditioned on $\mathbf{x} \in B$, i.e., $\mathbf{x}_{\mathbf{w}} \in [t_1, t_1 + \rho]$, it holds that

$$\mathbf{w}^* \cdot \mathbf{x} = \lambda_1 \mathbf{x}_{\mathbf{w}} + \lambda_2 (\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{x}^{\perp} = \lambda_1 t_1 + \lambda_2 (\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{x}^{\perp} + s\rho \,,$$

for some $s \in [-1, 1]$ (recall that $|\lambda_1| \leq 1$). Let $b = -(t + \lambda_1 t_1)/\lambda_2$. Notice that when $0 \leq \mathrm{sign}(b)((\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{x}^{\perp} - b) < \rho/\lambda_2$, $f^{\perp}(\mathbf{x}^{\perp})$ is not equal to the sign of $(\mathbf{w}^* \cdot \mathbf{x} + t)$ (recall that $\lambda_2 > 0$), and therefore we are inside the region that Massart noise is violated. Thus, we need to bound the probability of this event $I^{\xi}$. We have that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x}^{\perp},y)\sim\mathcal{D}^{\perp}}[\eta^{\perp}(\mathbf{x}^{\perp}) > 1/2] = \frac{\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{w}^*)^{\perp_{\mathbf{w}}} \cdot \mathbf{x} \in I^{\xi}\right]}{\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x} \in B]} \lesssim \rho/\lambda_2 \lesssim \rho/\epsilon \,,$$

where we used the anti-concentration property of the Gaussian distribution and the last inequality holds because we have that $\lambda_2 \gtrsim \epsilon$.

It remains to prove that there is a choice $t_1, t_2$ and a band $B = \{t_1 \leq \mathbf{w} \cdot \mathbf{x} \leq t_2\}$ with respect the $t_1, t_2$ such that $\mathbf{E}_{\mathbf{x} \sim (\mathcal{D}_B^{\perp})_{\mathbf{x}}}[(1 - 2\eta^{\perp}(\mathbf{x}))\mathbb{1}\{f(\mathbf{x})^{\perp} \neq \mathrm{sign}(-b)\}] \gtrsim \frac{\epsilon}{\sqrt{\log(1/\epsilon)}}$, where $\mathcal{D}_B^{\perp}$ is the distribution $\mathcal{D}$ conditioned on the set $B$ and projected onto $\mathbf{w}^{\perp}$. We first show the following claim Let $t_i = i\rho$ and $t_{-i} = -i\rho$, for $0 \leq i \leq C \log(1/\epsilon)/\rho$ where $C > 0$ is a large enough constant. We define $B_i = \{t_i \leq \mathbf{v} \cdot \mathbf{x} \leq t_{i+1}\}$ and $B_{-i} = \{-t_{i+1} \leq \mathbf{v} \cdot \mathbf{x} \leq -t_i\}$. For each $B_i$, we define the distributions $\mathcal{D}_{B_i}^{\perp}$, the hypothesis $f_i^{\perp}(\mathbf{x}^{\perp}) = \mathrm{sign}(\mathbf{x}^{\perp} \cdot (\mathbf{w}^*)^{\perp \mathbf{w}} - b_i)$ and the noise functions $\eta_i^{\perp}(\mathbf{x}^{\perp})$. We remind that from the assumptions, we have that for any $t' \in \mathbb{R}$, it holds $\mathbf{E}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}[\beta(\mathbf{x})\mathbb{1}\{f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t')\}] \geq \epsilon$. Choose $t' = t/\sin\theta$, where $\theta = \theta(\mathbf{w}, \mathbf{w}^*)$ and an application of Claim 2.4.8 to the set $\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t')\}$, gives us that there exists an index $i'$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}[\mathbb{1}\{f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t')\}\mathbb{1}\{\mathbf{x} \in B_{i'}\}\beta(\mathbf{x})] \gtrsim \frac{\epsilon\rho}{\sqrt{\log(1/\epsilon)}} \, .$$

Moreover, note that for the distribution $\mathcal{D}_{B_{i'}}$, that is the distribution $\mathcal{D}$ conditioned on $B$, it holds $\mathbf{E}_{\mathbf{x} \sim (\mathcal{D}_{B_{i'}})_{\mathbf{x}}}[\beta(\mathbf{x})\mathbb{1}\{f(\mathbf{x}) \neq \mathrm{sign}(\mathbf{w} \cdot \mathbf{x} + t')\}] \gtrsim \frac{\epsilon}{\sqrt{\log(1/\epsilon)}}$, where we used the Gaussian concentration. We have $f_{i'}^{\perp}(\mathbf{x})$ agrees almost everywhere with $f(\mathbf{x})$ with respect the distribution $\mathcal{D}_{B_{i'}}$, i.e., we have that $\mathbf{E}_{(\mathbf{x}_{\mathbf{w}}, \mathbf{x}^{\perp}) \sim (\mathcal{D}_{B_{i'}})_i}[\mathbb{1}\{f((\mathbf{x}_{\mathbf{w}}, \mathbf{x}^{\perp})) \neq f^{\perp}(\mathbf{x}^{\perp})\}] \lesssim \rho/\epsilon$. Thus using the triangle inequality, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x}_{\mathbf{w}}, \mathbf{x}^{\perp}) \sim (\mathcal{D}_{B_{i'}})_{\mathbf{x}}} \left[(1 - 2\eta^{\perp}(\mathbf{x}^{\perp}))\mathbb{1}\{\mathrm{sign}(\mathbf{x}_{\mathbf{w}} + t') \neq f^{\perp}(\mathbf{x}^{\perp})\}\right] \gtrsim \frac{\epsilon}{\sqrt{\log(1/\epsilon)}} - \rho/\epsilon \, .$$

The proof concludes by noting that in each $B_{i'}$, it holds $\mathrm{sign}(\mathbf{x}_{\mathbf{w}} + t') = \mathrm{sign}(-b_i)$ from construction of $f^{\perp}(\mathbf{x})$. This completes the proof of Lemma A.4.3.

□

We are now ready to prove the Theorem A.4.1. Note that this is the same proof as Theorem 2.4.1 the only difference is that we need to test the different thresholds.

*Proof of Theorem A.4.1.* First, let $\widehat{\mathcal{D}}_N$ be the empirical distribution of $\mathcal{D}$ using $N \in \mathbb{Z}_+$ samples. Let $\mathbf{w}^{(i)}$ be the current guess. If $\min_{t \in \mathbb{R}} \mathbf{Pr}_{(\mathbf{x},y) \sim \mathcal{D}}[\mathrm{sign}(\mathbf{w}^{(i)} \cdot \mathbf{x} + t) \neq y] \geq \mathrm{OPT} + \epsilon$ and $\theta(\mathbf{w}^{(i)}, \mathbf{w}^a st) \leq \pi - \epsilon$, then from Proposition A.4.2 with $N_1 = d^{O(\log(1/\epsilon))} \log(1/\delta_1)$ samples from $\widehat{\mathcal{D}}$ and $\mathrm{poly}(N_1, d)$ time, we compute a subspace $V$ such that $\|\mathrm{proj}_V(\mathbf{w}^*)\|_2 \geq \mathrm{poly}(\epsilon)$ and from Fact 2.4.15, we get a random a unit vector $\mathbf{v} \in V$ such that $\mathbf{v} \cdot \mathbf{w}^* = \mathrm{poly}(\epsilon)$ and $\mathbf{v} \cdot \mathbf{w}^{(i)} = 0$, with probability $(1 - \delta_1)/3$. We call this event $\mathcal{E}_i$. By conditioning on the event $\mathcal{E}_i$; from Lemma 2.4.14, after we update our current hypothesis vector $\mathbf{w}^{(i)}$ with $\mathbf{v}$, we get a unit vector $\mathbf{w}^{(i+1)}$ such that $\mathbf{w}^{(i+1)} \cdot \mathbf{w}^* \geq \mathbf{w}^{(i+1)} \cdot \mathbf{w}^* + \mathrm{poly}(\epsilon)$.

After running the update step $k$ times and conditioning on the events $\mathcal{E}_1, \ldots, \mathcal{E}_k$, then $\mathbf{w}^{(k)} \cdot \mathbf{w}^* \geq \mathbf{w}^{(1)} \cdot \mathbf{w}^* + k\,\mathrm{poly}(\epsilon)$; therefore, for $k = \mathrm{poly}(1/\epsilon)$, we get that the vector $\mathbf{w}^{(k)}$ that is competitive with the optimal hypothesis, i.e., $\min_{t\in\mathbb{R}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}^{(k)}\cdot\mathbf{x}+t) \neq y] \leq \mathrm{OPT}+\epsilon$. The probability that all the events $\mathcal{E}_1, \ldots, \mathcal{E}_k$ hold simultaneously is at least $(1-k\delta_1)+(1/3)^k$, and thus by choosing $\delta_1 \leq 1/(3k)$, the probability of success is at least $\delta_2 = (1/3)^k$. By running the algorithm above $M = \log(1/\delta)/\delta_2$ times and along with an application of Hoeffding's inequality, we get a list $L$ of $M$ vectors such that contains all the unit vectors $\mathbf{w}$ the algorithm calculated along with $-\mathbf{w}$, therefore $L$ contains a vector such $\mathbf{w}$ that such that $\min_{t\in\mathbb{R}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t) \neq y] \leq \mathrm{OPT}+\epsilon$, with probability $1-\delta/2$. Moreover, from Fact 2.2.12 we get that if we set $\mathcal{T} = \{\pm\epsilon^2, \pm 2\epsilon^2, \ldots, \pm 4\sqrt{\log(1/\epsilon)}\}$ we have that $\min_{t\in\mathcal{T}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t) \neq y] \leq \min_{t\in\mathbb{R}} \mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}+t) \neq y]+\epsilon^2$. Therefore, we construct the list $\mathcal{H} = \{(\mathbf{w}, t) : \mathbf{w} \in L, t \in \mathcal{T}\}$. Finally, to evaluate all the vectors from the list, we need a few samples, from the distribution $\mathcal{D}$ to obtain the best among them, i.e., the one that minimizes the zero-one loss.

The size of the list of candidates is at most $M \leq 2^{\mathrm{poly}(1/\epsilon)} \log(1/\delta)$. Therefore, from Hoeffding's inequality, it follows that $O(\mathrm{poly}(1/\epsilon)\log(1/\delta))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most $\epsilon$ with probability at least $1-\delta/2$. Thus, with $N = d^{\log(1/\epsilon)}\log(1/\delta)$ samples and $\mathrm{poly}(N, d, 2^{\mathrm{poly}(1/\epsilon)})$ runtime, we get a hypothesis $(\hat{\mathbf{w}}, \hat{t})$ such that $\mathbf{Pr}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{sign}(\hat{\mathbf{w}}\cdot\mathbf{x}+\hat{t}) \neq y] \leq \mathrm{OPT}+\epsilon$ with probability $1-\delta$. This completes the proof. □

# Appendix B

# Appendix to Chapter 3

**Organization** The structure of this appendix is as follows: In Section B.1, we provide a polynomial time cutting-planes based algorithm with sample complexity $\widetilde{O}(1/(\epsilon^2\gamma^4))$.

## B.1 Learning Margin Massart Halfspaces via Cutting Planes

In this section, we show how to use the cutting-planes method along with Lemma 3.2.2 to efficiently learning margin Massart Halfspaces using $\widetilde{O}(1/(\gamma^4\epsilon^2))$ samples.

Specifically, we establish the following result:

**Theorem B.1.1** (Learning Margin Massart Halfspaces with Cutting Planes)**.** *Let $D$ be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ which satisfies the $\eta$-Massart noise condition with respect to the $\gamma$-margin halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Given $N = \Theta(\log(1/(\gamma\delta)/(\gamma^4\epsilon^2))$ i.i.d. samples from $\mathcal{D}$, there is a $\mathrm{poly}(d, N)$ time algorithm that returns a vector $\hat{\mathbf{w}}$ such that $\mathrm{err}_{\mathcal{D}}(\hat{\mathbf{w}}) \leq \eta + \epsilon$ with probability at least $1 - \delta$.*

**Remark B.1.2.** We can always assume that $d = \widetilde{O}(1/\gamma^2)$. This holds since we can efficiently preprocess the data, using the Johnson-Lindenstrauss transform [JL84]. Similar dimension-reduction steps have been use in prior work, e.g., [CKMY20, DDK+23a].

Given the above remark, it suffices to establish the following:

**Theorem B.1.3.** *Let $D$ be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ which satisfies the $\eta$-Massart noise condition with respect to the $\gamma$-margin halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Given $N = \Theta(d \log(1/(\gamma\delta)/(\gamma^2\epsilon^2))$ i.i.d. samples from $\mathcal{D}$, there is a $\mathrm{poly}(d, N)$ time algorithm that returns a vector $\hat{\mathbf{w}}$ such that $\mathrm{err}_{\mathcal{D}}(\hat{\mathbf{w}}) \leq \eta + \epsilon$ with probability at least $1 - \delta$.*

The idea of using the cutting plane method is slightly adapted from [CKMY20]. Given access to a separation oracle for a convex set $\mathcal{K}$, we can find a point inside the set $\mathcal{K}$ by querying the separation oracle $O(d \log d)$ times. The difference with [CKMY20] is that we are using a more sophisticated (and sample efficient) separation oracle. This allows us to use $O(1/\epsilon^2)$ samples, instead of $O(1/\epsilon^3)$ samples, and leads to the optimal sample complexity as a function of $\epsilon$ (but not $\gamma$).

**Fact B.1.4.** *Suppose that $\mathcal{K}$ is an (unknown) convex body in $\mathbb{R}^d$ which contains a Euclidean ball of radius $r > 0$ and contained in a Euclidean ball centered at the origin of radius $R > 0$. There exists an algorithm which, given access to a separation oracle for $\mathcal{K}$, finds a point $\mathbf{x}^* \in \mathcal{K}$, runs in time $\mathrm{poly}(\log(R/r), d)$, and makes $O(d \log(Rd/r))$ calls to the separation oracle.*

We first show that if we get enough samples, we can efficiently approximate the gradients $\mathbf{G}(\mathbf{w}, \mathbf{w})$. Formally, we have:

**Proposition B.1.5** (Separation Oracle)**.** *Let $\epsilon, \delta \in (0,1)$ and let $D$ be a distribution on $\mathbb{S}^{d-1} \times \{\pm 1\}$ satisfying the $\eta$-Massart noise condition with respect to the halfspace $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x})$. Fix $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 \le 1$. Let $N \gtrsim \log(1/(\gamma\delta))/(\epsilon^2\gamma^2))$ and $\widehat{D}_N$ be the corresponding empirical distribution. Then, with probability at least $1-\delta$, it holds that*

$$\mathbf{G}_{\widehat{D}_N}(\mathbf{w}, \mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \ge 2(\mathrm{err}_D(\mathbf{w}) - \eta) - \epsilon \, .$$

*Proof.* By construction, $\mathbf{G}_{\widehat{D}_N}(\mathbf{w}, \mathbf{w}) = \mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) + \mathbf{G}^2_{\widehat{D}_N}(\mathbf{w})$ and by Lemma 3.2.2 we have that $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \ge 2(\mathrm{err}_{\widehat{D}_N}(\mathbf{w}) - \eta)$. By definition, we have

$$\mathop{\mathbf{E}}_{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)}) \sim \mathcal{D}} [\mathbf{G}^2_{\widehat{D}_N}(\mathbf{w})] = 0,$$

where the expectation is taken with respect to the sample set. Note that the norm of $\mathbf{g}^1(\mathbf{w}, \mathbf{x}), \mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)$, i.e., $\|\mathbf{g}^1(\mathbf{w}, \mathbf{x})\|_2, \|\mathbf{g}^2(\mathbf{w}, \mathbf{x}, y)\|_2$, is bounded pointwise from above by $4/\gamma$ for all $\mathbf{w} \in \mathbb{R}^d$. This can be seen as $\|\mathbf{x}\|_2 \le 1$, $W(\cdot, \gamma/2) \le 2/\gamma$, and $(1-2\eta), (1-2\eta(\mathbf{x})) \le 1$.

We use the following concentration inequality to show that our sample size is enough to guarantee that the estimated gradient is close to its population version.

**Fact B.1.6** ([SZ07], Lemma 1)**.** *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \in \mathbb{R}^d$ be random vectors such that for each $i \in [n]$ it holds $\|\mathbf{Z}_i\|_2 \le M < \infty$ almost surely and let $\sigma^2 = \sum_{i=1}^n \mathbf{E}[\|\mathbf{Z}_i\|_2^2]$. Then, we have that for any $\epsilon > 0$,*

$$\mathbf{Pr}\left[\left\| \frac{1}{n} \sum_{i=1}^n (\mathbf{Z}_i - \mathbf{E}[\mathbf{Z}_i]) \right\|_2 \ge \epsilon \right] \le 2 \exp\left( -\frac{n\epsilon}{2M} \log\left(1 + \frac{nM\epsilon}{\sigma^2}\right)\right) \, .$$

Using Fact B.1.6, along with the inequality $\log(1+z) \geq z/2$, for $z \in (0,1)$, we get that if $N \geq \Theta(\frac{\log(1/\delta)}{(\epsilon\gamma)^2})$, with probability at least $1-\delta$, we have

$$\left\| \mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}^1(\mathbf{w},\mathbf{x})] \right\|_2 \leq \epsilon \ , \tag{B.1}$$

and

$$\left\| \mathbf{G}^2_{\widehat{D}_N}(\mathbf{w}) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}^2(\mathbf{w},\mathbf{x},y)] \right\|_2 \leq \epsilon \ . \tag{B.2}$$

To complete the proof, recall that by Lemma 3.2.2 it holds $\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) \cdot (\mathbf{w}-\mathbf{w}^*) \geq 2(\mathrm{err}_{\widehat{D}_N}(\mathbf{w}) - \eta) - \epsilon$. Therefore, by taking the expectation over $D_{\mathbf{x}}$, we get that

$$\mathbf{G}^1_D(\mathbf{w}) \cdot (\mathbf{w}-\mathbf{w}^*) \geq 2(\mathrm{err}_D(\mathbf{w}) - \eta) \ .$$

The proof is completed by recalling that $\|\mathbf{G}^1_{\widehat{D}_N}(\mathbf{w}) - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}^1(\mathbf{w},\mathbf{x})]\|_2 \leq \epsilon$ from Inequality (B.1) and that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{g}^2(\mathbf{w},\mathbf{x},y)] = 0$. □

Equipped with Proposition B.1.5, we are ready to prove a weaker version of Theorem 3.2.1 using separation oracles and the cutting plane algorithm. Formally, we show that

*Proof of Theorem B.1.3.* Our convex set $\mathcal{K}$ is a Euclidean ball of radius $\gamma/2$ centered at $\mathbf{w}^*$. To see that, note that for any $\mathbf{v}$ such that $\|\mathbf{w}^* - \mathbf{v}\|_2 \leq \gamma/2$, we have that $|(\mathbf{w}^* - \mathbf{v}) \cdot \mathbf{x}| \leq \gamma/2$ for any $\mathbf{x}$ with $\|\mathbf{x}\|_2 = 1$. This implies that $\gamma/2 + \mathbf{w}^* \cdot \mathbf{x} \geq \mathbf{v} \cdot \mathbf{x} \geq \mathbf{w}^* \cdot \mathbf{x} - \gamma/2$. Moreover, by definition we have that $\mathbf{w}^* \cdot \mathbf{x} \geq \gamma$. Hence, if $\mathbf{w}^* \cdot \mathbf{x} \geq 0$, we have that $\mathbf{v} \cdot \mathbf{x} \geq \gamma/2$; and if $\mathbf{w}^* \cdot \mathbf{x} \leq 0$, we have that $\mathbf{v} \cdot \mathbf{x} \leq -\gamma/2$. Therefore, this ball contains all the vectors $\mathbf{w}$ with margin $\gamma/2$ and separates the points in the same way as $\mathbf{w}^*$.

Therefore, as long as we are not in the set $\mathcal{K}$ or the 0-1 error is more than $\eta+\epsilon$, we can use Proposition B.1.5 to construct a new separation oracle. By Fact B.1.4, the maximum number of calls to the separation oracle is $T = O(d\log(d/\gamma))$. By Proposition B.1.5, in each round we need $n = O(\log(T/\delta))/(\epsilon^2\gamma^2)$ samples from $\mathcal{D}$ to construct a separation oracle. Therefore, the maximum number of samples is $O(nT) = O(d\log(T/\delta))/(\epsilon^2\gamma^2)$. This completes the proof. □

# Appendix C

# Appendix to Chapter 5

## C.1 Learning Halfspaces using the Ramp Activation

In this section we show that by optimizing as a surrogate function the ramp activation, we can find a $\hat{\mathbf{w}} \in \mathbb{R}^d$, that gets error $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{sign}(\hat{\mathbf{w}}\cdot\mathbf{x}) - y)^2] = O(\epsilon)$. Specifically, we show the following corollary of Corollary 5.4.2.

**Corollary C.1.1.** *Let $\mathcal{D}$ be an $(L,R)$-well-behaved distribution on $\mathbb{R}^d \times \{\pm 1\}$. Define $F(\mathbf{w}) = (1/2)\,\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) - y)^2]$ and $\mathrm{OPT} = \inf_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w})$. Fix $\epsilon, \delta \in (0,1)$ and set $\kappa = \mathrm{poly}(LR)$. There exists an algorithm that given $N = \widetilde{\Theta}(d/\epsilon \log(1/\delta))\ \mathrm{poly}(1/\kappa)$ samples, runs in time $T = \mathrm{poly}(1/(\epsilon\kappa))$ and returns a vector $\hat{\mathbf{w}} \in \mathbb{R}^d$ that, with probability $1-\delta$, satisfies*

$$F(\hat{\mathbf{w}}) \leq \mathrm{poly}(1/\kappa)\ \mathrm{OPT} + \epsilon\,.$$

*Proof.* We are going to show that by optimizing the ramp activation instead of the $\mathrm{sign}(\cdot)$, we can find a good candidate solution. First note that $r(t)$ is an $(1, e, 1)$-sigmoidal activation. The proof relies on the following fact:

**Fact C.1.2.** *Let $\mathbf{w} \in \mathbb{R}^d$ be a unit vector. Then, for any $\mathbf{x} \in \mathbb{R}^d$, it holds that*

$$\lim_{z\to\infty} r(z\mathbf{w}\cdot\mathbf{x}) = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\,.$$

Therefore, for any unit vector $\mathbf{w} \in \mathbb{R}^d$, from Lemma 5.2.5 there exists a $\mathbf{w}' \in \mathbb{R}^d$ with $\|\mathbf{w}'\|_2 = O(1/\epsilon)$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(r(\mathbf{w}'\cdot\mathbf{x}) - \mathrm{sign}(\mathbf{w}\cdot\mathbf{x}))^2] \leq \epsilon$. From Corollary 5.4.2, we get that there exists an algorithm that with $N = \widetilde{\Theta}(d/\epsilon\log(1/\delta))\ \mathrm{poly}(1/\kappa)$ samples, and runtime $T = \mathrm{poly}(1/(\epsilon\kappa))$ returns $\hat{\mathbf{v}} \in \mathbb{R}^d$ with $\|\hat{\mathbf{v}}\|_2 = O(1/\epsilon)$, such that with

probability $1-\delta$, it holds

$$(1/2)\,\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(r(\hat{\mathbf{v}}\cdot\mathbf{x})-y)^2] \le \mathrm{poly}(1/(LR))\mathrm{OPT}+\epsilon\,.$$

The proof follows by noting that because $r(\hat{\mathbf{v}})\in(-1,1)$ and $y\in\{\pm 1\}$, it holds that

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(r(\hat{\mathbf{v}}\cdot\mathbf{x})-y)^2] \ge (1/2)\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{sign}(r(\hat{\mathbf{v}}\cdot\mathbf{x}))-y)^2] = (1/2)\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathrm{sign}(\hat{\mathbf{v}}\cdot\mathbf{x})-y)^2]\,.$$

Hence, $F(\hat{\mathbf{v}})\le \mathrm{poly}(1/(LR))\mathrm{OPT}+2\epsilon$. □

Next, we prove the following sample complexity lower bound for optimizing sigmoidal activations.

**Lemma C.1.3** (Sample Complexity Lower Bound for Sigmoidal Activations)**.** *Fix any absolute constant $C\ge 1$ and let $\sigma(t)=1/(1+e^{-t})$ be the logistic activation. Any $C$-approximate algorithm that learns (with success probability at least $2/3$) the logistic activation under any $\epsilon$-corrupted $\mathcal{D}$ with standard normal $\mathbf{x}$-marginal requires $\Omega(d/\epsilon)$ samples from $\mathcal{D}$.*

*Proof.* Assume that $y=\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\ge 0\}$ for some unit vector $\mathbf{w}^*$, i.e., $y$ is a noiseless halfspace. It is easy to see that the corresponding distribution $D$ with Gaussian $\mathbf{x}$-marginal and $y$ given by $\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\ge 0\}$ is $\epsilon$-corrupted for any $\epsilon>0$. Since $\lim_{z\to\infty}\sigma(z\mathbf{w}^*\cdot\mathbf{x})=\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\ge 0\}$, by the dominated convergence theorem we have that $\lim_{z\to\infty}\mathbf{E}_{(\mathbf{x},y)\sim D}\,F^{D,\sigma}(z\mathbf{w}^*)=0$. Therefore, $\inf_{\mathbf{w}\in\mathbb{R}^d}F^{D,\sigma}(\mathbf{w})=0$. Let $\mathcal{A}$ be any algorithm that given a sigmoidal activation $\sigma$ and $N>0$ samples from $\mathcal{D}$, returns a vector $\mathbf{w}$ such that $F^{D,\sigma}(\mathbf{w})\le\epsilon$. We will show that $\mathbf{w}$ corresponds to a classifier with at most $2C\epsilon$ error. By noting that $\sigma(\mathbf{w})\in(0,1)$ and $y\in\{0,1\}$, it holds that

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2] &\ge (1/2)\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathbb{1}\{\sigma(\mathbf{w}\cdot\mathbf{x})\ge 1/2\})-y)^2]\\ &= (1/2)\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}\{\sigma(\mathbf{w}\cdot\mathbf{x})\ge 1/2\})\ne y]\,.\end{aligned}$$

Thus, we have constructed a binary classifier that achieves disagreement at most $2C\epsilon$ with $y$. It is well-known that any algorithm that finds a vector $\mathbf{w}\in\mathbb{R}^d$ such that $\mathbf{Pr}[\mathrm{sign}(\mathbf{w}\cdot\mathbf{x})\ne y]\le O(\epsilon)$, needs $\Omega(d/\epsilon)$ samples (see, e.g., [Lon95]). Hence, $\mathcal{A}$ needs at least $N=\Omega(d/\epsilon)$ samples. □

# Appendix D

# Appendix to Chapter 6

## D.1 Remarks about the Distribution Class and the Activation Class

In this section, we show that without the loss of generality we can assume that the parameters $L, R$ in the distributional assumptions (Definition 5.1.1) can be taken less than 1, while the parameters $a, b$ of the activations functions (see Definition 1.3.20) can be taken as $a, 1/b \leq 1$.

**Remark D.1.1** (Distribution/Activation Parameters, (Definitions 1.3.20 and 5.1.1))**.** *We observe that if a distribution $\mathcal{D}_{\mathbf{x}}$ is $(L, R)$-well-behaved, then it is also $(L', R')$-well-behaved for any $0 < L' \leq L, 0 < R' \leq R$. Hence, it is without loss of generality to assume that $L, R \in (0, 1]$. Similarly, if an activation is $(a, b)$-unbounded, it is also an $(a, b')$-unbounded activation with $b' \geq b$. Thus, we assume that $b \geq 1$. We can similarly assume $a \leq 1$.*

In addition, we remark that the $(L, R)$-well behaved distributions are sub-exponential.

**Remark D.1.2** (Sub-exponential Tails of Well-Behaved Distributions, Definition 5.1.1)**.** *Definition 5.1.1 might seem abstract, but to put it plain it implies that the random variable $\mathbf{x}$ has a $(1/L)$-sub-exponential tail, and that the pdf of the projected random variable $\mathbf{x}_V$ onto the space $V$ is lower bounded by $L$. To see the first statement, given any unit vector $\mathbf{p}$, let $\mathbf{x}_{\mathbf{p}}$ be the projection of $\mathbf{x}$ onto the one-dimensional linear space $V_{\mathbf{p}} = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{z} = t\mathbf{p}, t \in \mathbb{R}\}$, i.e., $\mathbf{x}_{\mathbf{p}} = \mathbf{p} \cdot \mathbf{x} \in V_{\mathbf{p}}$. Then, by the anti-concentration and concentration property, we have*

$$\mathbf{Pr}[|\mathbf{p} \cdot \mathbf{x}| \geq r] = \mathbf{Pr}[|\mathbf{x}_{\mathbf{p}}| \geq r] \leq \int_{|x| \geq r} \gamma(x)\, \mathrm{d}x \leq 2 \int_r^{\infty} \frac{1}{L} \exp(-Lx)\, \mathrm{d}x = \frac{2}{L^2} \exp(-Lr),$$

*which implies that $\mathbf{x}$ possesses a sub-exponential tail.*

## D.2 Efficiently Computing the Optimal Empirical Activation

In this section, we show that the optimization problem (P) can be solved efficiently, following the framework from [LH22] with minor modifications. We show that, for any $\epsilon > 0$, there is an efficient algorithm that runs in $\tilde{O}(m^2 \log(1/\epsilon))$ time and outputs a solution $\hat{v}^t(z)$ such that $\|\hat{v}^t(z) - \hat{u}^t(z)\|_\infty \leq \epsilon$. We then argue that using such approximate solutions to the optimization problem (P) does not negatively impact our error guarantee, sample complexity, or runtime (up to constant factors).

**Proposition D.2.1** (Approximating the Optimal Empirical Activation)**.** *Let $\epsilon > 0$, and $\mathcal{D}_{\mathbf{x}}$ be $(L, R)$-well behaved. Let $\hat{u}^t \in \mathcal{U}_{(a,b)}$ be the optimal solution of the optimization problem (P) given a sample set $S$ of size $m$ drawn from $\mathcal{D}$ and a parameter $\mathbf{w}^t \in \mathbb{B}(W)$. There exists an algorithm that produces an activation $\hat{v}^t \in \mathcal{U}_{(a,b)}$ such that $\|\hat{v}^t - \hat{u}^t\|_\infty \leq \epsilon$, with runtime $\tilde{O}(m^2 \log(bW/(L\epsilon)))$.*

To prove Proposition D.2.1, we leverage the following result:

**Lemma D.2.2** (Section 5 [LH22])**.** *Let $f_i(y)$ and $h_i(y)$ be any convex lower semi-continuous functions for $i = 1, \ldots, m$. Consider the following convex optimization problem*

$$(\hat{y}_1, \ldots, \hat{y}_m) = \operatorname*{argmin}_{y_1,\ldots,y_m} \sum_{i=1}^{m} f_i(y_i) + \sum_{i=1}^{m-1} h_i(y_i - y_{i+1}), \tag{D.1}$$

*where $y_i \in [-U, U]$ for all $i = 1, \ldots, m$ for some positive constant $U$. Then, for any $\epsilon > 0$, there exists an algorithm (the cc-algorithm [LH22]) that outputs an $\epsilon$-close solution $\{y_1, \ldots, y_m\}$ such that $|y_i - \hat{y}_i| \leq \epsilon$ for all $i \in [m]$ with runtime $O(m^2 \log(U/\epsilon))$.*

*Proof of Proposition D.2.1.* We first reformulate problem (P) as a quadratic optimization problem with linear constraints. To guarantee that $\hat{u}^t$ is an element in $\mathcal{U}_{(a,b)}$ that satisfies $\hat{u}^t(0) = 0$, we add a zero point $(\mathbf{x}^{(0)}, y^{(0)}) = (\mathbf{0}, 0)$ to the data set $S$ if $S$ does not contain $(\mathbf{0}, 0)$ in the first place. We can thus assume without loss of generality that the data set contains $(\mathbf{0}, 0)$. Denote $z_i = \mathbf{w} \cdot \mathbf{x}^{(i)}$ such that $z_1 \leq z_2 \leq \cdots \leq z_m$ after rearranging the order of $(\mathbf{x}^{(i)}, y^{(i)})$'s, and suppose $z_k = \mathbf{w} \cdot \mathbf{x}_0 = 0$ for a $k \in [m]$. Then (P) is equivalent to the following optimization problem:

$$\begin{aligned}
(\hat{y}^{(1)}, \cdots, \hat{y}^{(m)}) = \operatorname*{argmin}_{\tilde{y}^{(i)}, i\in[m]} \ & \sum_{i=1}^{m} (\tilde{y}^{(i)} - y^{(i)})^2 \\
\text{s.t.} \ \ & 0 \leq \tilde{y}^{(i+1)} - \tilde{y}^{(i)}, && 1 \leq i \leq k-1, \\
& a(z_{i+1} - z_i) \leq \tilde{y}^{(i+1)} - \tilde{y}^{(i)}, && 1 \leq i \leq k-1, \\
& \tilde{y}^{(i+1)} - \tilde{y}^{(i)} \leq b(z_{i+1} - z_i), && 1 \leq i \leq m-1, \\
& \tilde{y}^{(k)} = 0.
\end{aligned}$$

Define $h_i(y) = \mathcal{I}_{[-b(z_{i+1}-z_i),0]}(y)$ for $i = 1\ldots,k-1$, $h_i(y) = \mathcal{I}_{[-b(z_{i+1}-z_i),-a(z_{i+1}-z_i)]}(y)$ for $i = k\ldots,m-1$, where $\mathcal{I}_{\mathcal{Y}}(y)$ is the indicator function of a convex set $\mathcal{Y}$, i.e., $\mathcal{I}_{\mathcal{Y}}(y) = 0$ if $y \in \mathcal{Y}$ and $\mathcal{I}_{\mathcal{Y}}(y) = +\infty$ otherwise. It is known that $h_i$'s are convex and sub-differentiable on their domain $\mathcal{Y}_i$. In addition, let $f_i(y) = \frac{1}{2}(y - y^{(i)})^2$ for $i \neq k$ and $f_k(y) = \mathcal{I}_{\{0\}}(y)$. Then, we have the following formulation for problem (P):

$$(\hat{y}^{(1)}, \cdots, \hat{y}^{(m)}) = \operatorname*{argmin}_{\tilde{y}^{(i)}, i=1,\ldots,m} \sum_{i=1}^{m} f_i(\tilde{y}^{(i)}) + \sum_{i=1}^{m-1} h_i(\tilde{y}^{(i)} - \tilde{y}^{(i+1)}) \tag{P1}$$

Note that the functions $f_i$ and $h_i$ we defined above satisfy the conditions of Lemma D.2.2. Thus, it only remains to find the bounds on the variables $\tilde{y}^{(i)}$. This is easy to achieve as all $\tilde{y}^{(i)}$ must satisfy $|\tilde{y}^{(i)}| \leq b|z_i| = b|\mathbf{w} \cdot \mathbf{x}^{(i)}|$ and we know that $\mathbf{x}^{(i)}$ are sub-exponential random variables. Therefore, following the same idea from the proof of Lemma D.3.2, we know that for $U = \frac{2W}{L}\log(m/(L\delta))$, it holds that with probability at least $1-\delta$, $|\tilde{y}^{(i)}| \leq b|\mathbf{w} \cdot \mathbf{x}^{(i)}| \leq bU$ for all $i \in [m]$. Hence, applying Lemma D.2.2 to problem (P1), we get that it can be solved within $\epsilon$-error in runtime $\tilde{O}(m^2 \log(bW/(L\epsilon)))$. □

**The effect of approximation error in (P)** Since the solution $\hat{v}^t$ is $\epsilon$-close to $\hat{u}^t$, this approximated solution will only result in an $\epsilon$-additive error in the sharpness result Proposition 6.2.1 and the gradient norm concentration Lemma 6.3.3. In more detail, for the result of Proposition 6.2.1, we have

$$\begin{aligned}
&\left|(\nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{v}^t) - \nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t; \hat{u}^t)) \cdot (\mathbf{w}^t - \mathbf{w}^*)\right| \\
&= \left|\frac{1}{m}\sum_{i=1}^{m}(\hat{v}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - \hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}^{(i)}))(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)}\right| \\
&\leq \frac{\epsilon}{m}\sum_{i=1}^{m}\left|(\mathbf{w}^t - \mathbf{w}^*) \cdot \mathbf{x}^{(i)}\right| \leq 2\epsilon U,
\end{aligned}$$

since $|\mathbf{w}^t \cdot \mathbf{x}^{(i)}| \leq U$ and $|\mathbf{w}^* \cdot \mathbf{x}^{(i)}| \leq U$ with probability at least $1-\delta$. Therefore, choosing $\epsilon' = \epsilon/U$ we have that Proposition 6.2.1 holds for approximate activations $\hat{v}^t$ with an additional $\epsilon$ error. Observe that this does not affect the approximation factor in our final $O(\text{OPT}) + \epsilon$ result, while the value of $\epsilon$ only needs to be rescaled by a constant factor, effectively increasing the sample size and the runtime by constant factors.

Let us denote the unit ball by $\mathbb{B}$. For the gradient norm concentration lemma

Lemma 6.3.3, note that at any iteration $t$, it always holds that

$$\|\nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t;\hat{v}^t)-\nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t;\hat{u}^t)\|_2=\max_{\mathbf{v}\in\mathbb{B}}\frac{1}{m}\sum_{i=1}^m(\hat{v}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)})-\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}^{(i)}))\mathbf{x}^{(i)}\cdot\mathbf{v}$$
$$\leq\max_{\mathbf{v}\in\mathbb{B}}\frac{\epsilon}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}|.$$

Since $\mathbf{E}[\mathbf{x}\mathbf{x}^\intercal]\preccurlyeq\mathbf{I}$ and $\mathbf{v}\in\mathbb{B}$, we have $\mathbf{E}[|\mathbf{x}\cdot\mathbf{v}|]\leq\sqrt{\mathbf{E}[(\mathbf{x}\cdot\mathbf{v})^2]}\leq 1$. Now since $|\mathbf{x}^{(i)}\cdot\mathbf{v}|$ are independent $1/L$-sub-exponential random variables, applying Bernstein's inequality it holds that for any $\mathbf{v}\in\mathbb{B}$ and an absolute constant $c$,

$$\mathbf{Pr}[|\frac{1}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}|-\mathbf{E}[|\mathbf{x}\cdot\mathbf{v}|]|\geq s]\leq 2\exp(-c\min\{\frac{m^2s^2}{m/L^2},\frac{ms}{1/L}\})=2\exp(-cmL^2s^2).$$

Let $N(\mathbb{B},\epsilon;\ell_2)$ be the $\epsilon$-net of the unit ball $\mathbb{B}$. Note that the cover number of these $\mathbf{v}\in\mathbb{B}$ is of order $(1/\epsilon)^{O(d)}$, therefore, applying a union bound on $N(\mathbb{B},\epsilon;\ell_2)$ and for all $t_0JT=O(\log(1/\epsilon)/\sqrt{\epsilon})$ iterations, and setting $s=1$, it holds

$$\mathbf{Pr}\left[\forall\mathbf{v}\in N(\mathbb{B},\epsilon;\ell_2),\ \left|\frac{1}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}|-\mathbf{E}[|\mathbf{x}\cdot\mathbf{v}|]\right|\geq 1\right]\leq 2\exp(-cmL^2+c'd\log(1/\epsilon))\leq\delta,$$

where the last inequality comes from the fact that we have $m\gtrsim\frac{W^{9/2}b^{14}d\log(1/\delta)\log^4(d/\epsilon)}{(L^4\mu^9\delta\epsilon^{3/2})}$ as the batch size. Let $\mathbf{v}^*=\text{argmax}_{\mathbf{v}\in\mathbb{B}}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}|$. Then there exists a $\mathbf{v}'\in N(\mathbb{B},\epsilon;\ell_2)$ such that $\|\mathbf{v}'-\mathbf{v}^*\|_2\leq\epsilon$ and hence,

$$\frac{1}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}^*|\leq\frac{1}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot(\mathbf{v}^*-\mathbf{v}')|+\frac{1}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}'|$$
$$=\frac{\epsilon}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\frac{\mathbf{v}^*-\mathbf{v}'}{\epsilon}|+\frac{1}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}'|$$
$$\leq\frac{\epsilon}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}^*|+\frac{1}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}'|,$$

where the last inequality comes from the observation that as $(\mathbf{v}^*-\mathbf{v}')/\epsilon\leq\mathbb{B}$, it holds $\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot((\mathbf{v}^*-\mathbf{v}')/\epsilon)|\leq\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}^*|$, by the definition of $\mathbf{v}^*$. Therefore, with probability at least $1-\delta$ we have

$$\frac{1}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}^*|\leq\frac{1}{1-\epsilon}\frac{1}{m}\sum_{i=1}^m|\mathbf{x}^{(i)}\cdot\mathbf{v}'|\leq 2(1+\mathbf{E}[|\mathbf{v}\cdot\mathbf{x}|])\leq 4.$$

This implies that $\|\nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t;\hat{v}^t)\|_2 \leq \|\nabla\widehat{\mathcal{L}}_{\text{sur}}(\mathbf{w}^t;\hat{u}^t)\|_2 + 4\epsilon$ for all iterations with probability at least $1-\delta$. Therefore, Lemma 6.3.3 continues to hold for the $\epsilon$-approximate activation $\hat{v}^t$.

Thus, we have that the inequalities (6.39) and (6.41) in the proof of Theorem 6.3.2 remain valid for $\epsilon$-approximate $\hat{v}^t$, and hence the results in Theorem 6.3.2 are unchanged.

## D.3 Uniform Convergence of Activations

In this section, we review and provide standard uniform convergence results showing that the sample-optimal activations concentrate around their population-optimal counterparts. We first bound the $L_2^2$ distance between the sample-optimal and population-optimal activations under $\mathbf{w}^t$. To do so, we build on Lemma 8 in [KKSK11]. Note that Lemma 8 from [KKSK11] only works for bounded 1-Lipschitz activations $u:\mathbb{R}\mapsto[0,1]$, hence it is not directly applicable to our case. Fortunately, since $\mathcal{D}_{\mathbf{x}}$ has a sub-exponential tail (see Definition 5.1.1), we are able to bound the range of $u(\mathbf{w}\cdot\mathbf{x})$ for $u\in\mathcal{U}_{(a,b)}$ and $\mathbf{w}\in\mathbb{B}(W)$ with high probability. Concretely, we prove the following lemma. Note that in the lemma statement, $\hat{u}^{*t}$ is a random variable defined w.r.t. the (random) dataset $S^*$, and thus the probabilistic statement is for this random variable.

We make use of the following fact from [KKSK11]:

**Fact D.3.1** (Lemma 8 [KKSK11])**.** *Let $\mathcal{V}$ be the set of non-deceasing 1-Lipschitz functions such that $v:\mathbb{R}\to[0,1]$, $\forall v\in\mathcal{V}$. Given $S_m=\{(\mathbf{x}^{(i)},y^{(i)})\}_{i=1}^m$, where $(\mathbf{x}^{(i)},y^{(i)})$ are sampled i.i.d. from some distribution $\mathcal{D}'$, let*

$$\hat{v}_{\mathbf{w}}\in\operatorname*{argmin}_{v\in\mathcal{V}}\frac{1}{m}\sum_{i=1}^m(v(\mathbf{w}\cdot\mathbf{x}^{(i)})-y^{(i)})^2.$$

*Then, with probability at least $1-\delta$ over the random dataset $S_m$, for any $\mathbf{w}\in\mathbb{B}(W)$ it holds uniformly that*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}[(\hat{v}_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x})-y)^2]-\inf_{v\in\mathcal{V}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}'}[(v(\mathbf{w}\cdot\mathbf{x})-y)^2]=O\left(W\left(\frac{d\log(Wm/\delta)}{m}\right)^{2/3}\right).$$

The first lemma states that with sufficient many of samples, the idealized sample-optimal activation $\hat{u}^{*t}$ defined as the optimal solution of (P*) is close to its population counterpart $u^{*t}$, the optimal solution of (EP*).

**Lemma D.3.2** (Approximating Population-Optimal Noiseless Activation by Sample-Optimal)**.** *Let $\mathcal{D}_{\mathbf{x}}$ be $(L,R)$-well behaved and let $\mathbf{w}^t\in\mathbb{B}(W)$. Provided a dataset $S^*=\{(\mathbf{x}^{(i)},y^{*(i)})\}$, where $\mathbf{x}^{(i)}$ are i.i.d. samples from $\mathcal{D}_{\mathbf{x}}$ and $y^{*(i)}=u^*(\mathbf{w}^*\cdot\mathbf{x}^{(i)})$, let $\hat{u}^{*t}$ be the sample-optimal activation on $S^*$ as defined in (P*). In addition, let $u^{*t}$ be the*

*corresponding population-optimal activation, following the definition in* (EP*). *Then, for any* $\epsilon, \delta > 0$, *if the size* $m$ *of the dataset* $S^*$ *is sufficiently large*

$$m \gtrsim d \log^4(d/(\epsilon\delta)) \left( \frac{b^2 W^3}{L^2 \epsilon} \right)^{3/2},$$

*we have that with probability at least* $1 - \delta$, *for any* $\mathbf{w}^t \in \mathbb{B}(W)$:

$$\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] \leq \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}_\mathbf{x}} [(u^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2] + \epsilon \,,$$

*and, furthermore,*

$$\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \leq \epsilon.$$

*Proof.* Our goal is to show that with high probability, the sample-optimal activation $\hat{u}^{*t} \in \mathcal{U}_{(a,b)}$ and the population optimal activation $u^{*t} \in \mathcal{U}_{(a,b)}$ can be scaled to 1-Lipschitz functions mapping $\mathbb{R}$ to $[0, 1]$, then, Fact D.3.1 can be applied.

Since $\mathbf{x}$ possesses a sub-exponential tail, for any $\mathbf{w} \in \mathbb{B}(W)$ we have $\mathbf{Pr}[|\mathbf{w} \cdot \mathbf{x}| \geq \|\mathbf{w}\|_2 r] \leq \frac{2}{L^2} \exp(-Lr)$. Therefore, with probability at least $1 - (\delta_1/m)^2$ it holds $|\mathbf{w} \cdot \mathbf{x}| \leq \frac{2W}{L} \log(m/(L\delta_1))$. Since we have $m$ samples, a union bound on these $m$ samples yields that with probability at least $1 - \delta_1^2/m$ it holds $|\mathbf{w} \cdot \mathbf{x}^{(i)}| \leq \frac{2W}{L} \log(m/(L\delta_1))$, for any given $\mathbf{w} \in \mathbb{B}(W)$. Let $r = \frac{2W}{L} \log(m/(L\delta_1))$. In the remainder of the proof, we assume that $\mathbf{w}^t \cdot \mathbf{x}^{(i)} \leq r$ holds for every $\mathbf{x}^{(i)}$ in the dataset $S^*$, which happens with probability at least $1 - \delta_1^2/m \geq 1 - \delta_1$.

Let $\mathcal{V}$ be the set of non-decreasing 1-Lipschitz functions $v : \mathbb{R} \to [0, 1]$ such that $v(0) = 1/2$, and $v(z_1) - v(z_2) \geq (a/(2br))(z_1 - z_2)$ for all $z_1 \geq z_2 \geq 0$. We observe that restricted on the interval $|z| \leq r$, $(\hat{u}^{*t}(z)/(2br) + 1/2)|_{|z| \leq r}$ is 1-Lipschitz, non-decreasing and bounded in the interval $[0, 1]$. Thus, $(\hat{u}^{*t}(z)/(2br) + 1/2)|_{|z| \leq r} = \hat{v}^{*t}(z)|_{|z| \leq r}$, for some $\hat{v}^{*t} \in \mathcal{V}$. Furthermore, under the condition that $|\mathbf{w}^t \cdot \mathbf{x}^{(i)}| \leq r$, since $(\hat{u}^{*t}(z)/(2br) + 1/2)|_{|z| \leq r} = \hat{v}^{*t}(z)|_{|z| \leq r}$, we observe that $v^{*t}(z)$ is the optimal activation in the function space $\mathcal{V}$, given the dataset $S^*$ and parameter $\mathbf{w}^t$, i.e.,

$$\hat{v}^{*t} \in \operatorname*{argmin}_{v \in \mathcal{V}} \frac{1}{m} \sum_{i=1}^{m} (v(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - (u^*(\mathbf{w}^* \cdot \mathbf{x}^{(i)})/(2br) + 1/2))^2.$$

In other words, $\hat{u}^{*t}(z)/(2br) + 1/2$ is the sample-optimal activation in the function class $\mathcal{V}$ when restricted to the interval $|z| \leq r$. Consider $\mathbf{x} \sim \mathcal{D}_\mathbf{x}$. Then $\mathbf{Pr}[|\mathbf{w}^t \cdot \mathbf{x}| \geq r] \leq (\delta_1/m)^2$ and for any $\mathbf{w}^t \in \mathbb{B}(W)$, the expectation $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^*(\mathbf{w}^* \cdot \mathbf{x}))^2]$ can

be decomposed into the following terms

$$\begin{aligned}\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] &= \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\le r\}]\\ &\quad+\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|> r\}]\\ &\le \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\le r\}]\\ &\quad+2\,\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2+(u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|> r\}]\end{aligned} \tag{D.2}$$

Since both $\hat{u}^{*t}$ and $u^*$ are $(a,b)$-unbounded functions such that $\hat{u}^{*t}(0)=u^*(0)=0$, we have $(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2\le b^2W^2((\mathbf{w}^t/\|\mathbf{w}^t\|_2)\cdot\mathbf{x})^2$ and similarly, $(u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\le b^2W^2((\mathbf{w}^*/\|\mathbf{w}^*\|_2)\cdot\mathbf{x})^2$. Furthermore, since for any unit vector $\mathbf{a}$, the random variable $\mathbf{a}\cdot\mathbf{x}$ follows a $(1/L)$-sub-exponential distribution as $\mathcal{D}_\mathbf{x}$ is $(L,R)$-well behaved, thus, it holds that $\mathbf{E}[(\mathbf{a}\cdot\mathbf{x})^4]\le c/L^4$ for some absolute constant $c$. Therefore, after applying Cauchy-Schwarz inequality to $\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\ge r\}]$, we get

$$\begin{aligned}\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\ge r\}] &\le b^2W^2\sqrt{\mathbf{E}[((\mathbf{w}^t/\|\mathbf{w}^t\|_2)\cdot\mathbf{x})^4]\,\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}|\ge r]}\\ &\le cb^2W^2\delta_1/(L^2m),\end{aligned} \tag{D.3}$$

and similarly, $\mathbf{E}[(u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\ge r\}]\le cb^2W^2\delta_1/(L^2m)$. Thus, plugging these inequalities back into (D.2), we get

$$\begin{aligned}\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] &\le \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\le r\}]\\ &\quad+2cb^2W^2\delta_1/(L^2m).\end{aligned}$$

We are now ready to apply Fact D.3.1 (note that $\mathcal{V}$ is a smaller function class compared to the class of 1-Lipschitz functions described Fact D.3.1, hence Fact D.3.1 applies). Denote $A=\{\mathbf{x}:|\mathbf{w}^t\cdot\mathbf{x}|\le r\}$. Let $y'=y^*/(2br)+1/2$, $y^*=u^*(\mathbf{w}^*\cdot\mathbf{x})$. Since conditioning on $A$, $\hat{u}^{*t}(z)/(2br)+1/2$ is the sample-optimal activation, applying Fact D.3.1 we get that with probability at least $1-\delta_2$:

$$\begin{aligned}&\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x})/(2br)+1/2-(u^*(\mathbf{w}^*\cdot\mathbf{x})/(2br)+1/2))^2|A]\\ &=\mathbf{E}[(\hat{v}^{*t}(\mathbf{w}^t\cdot\mathbf{x})-y')^2|A]\\ &\le\inf_{v\in\mathcal{V}}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(v(\mathbf{w}^t\cdot\mathbf{x})-y')^2|A]+\tilde{O}(W(d\log(m/\delta_2)/m)^{2/3}).\end{aligned}$$

Let $\mathcal{V}|_{|z|\le r}$ and $\mathcal{U}_{(a,b)}|_{|z|\le r}$ be the functions from $\mathcal{V}$ and $\mathcal{U}_{(a,b)}$ restricted on the interval $|z|\le r$, respectively. It is not hard to see that by the definition of $\mathcal{U}_{(a,b)}$ and $\mathcal{V}$,

$(\mathcal{U}_{(a,b)}|_{|z|\le r})/(2br) + 1/2 \subset \mathcal{V}|_{|z|\le r}$. Therefore,

$$\begin{aligned}\inf_{v\in\mathcal{V}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(v(\mathbf{w}^t\cdot\mathbf{x}) - y')^2|A] &\le \inf_{u\in\mathcal{U}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t\cdot\mathbf{x})/(2br) + 1/2 - y')^2|A] \\ &\le \frac{1}{4b^2r^2}\inf_{u\in\mathcal{U}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t\cdot\mathbf{x}) - y^*)^2|A].\end{aligned}$$

Hence, with probability at least $1-\delta_2$,

$$\begin{aligned}&\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{A\}] \\ &= 4b^2r^2\,\mathbf{E}[(\hat{v}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - y')^2|A]\,\mathbf{Pr}[A] \\ &\le 4b^2r^2\inf_{v\in\mathcal{V}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(v(\mathbf{w}^t\cdot\mathbf{x}) - y')^2|A]\,\mathbf{Pr}[A] + \tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3})\,\mathbf{Pr}[A] \\ &\le \inf_{u\in\mathcal{U}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbb{1}\{A\}] + \tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3}) \\ &\le \inf_{u\in\mathcal{U}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] + \tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3})\ .\end{aligned}$$

Setting $\delta_1 = \delta_2 = \delta/2$ and plugging everything back into (D.4), we finally get that with probability at least $1-\delta$,

$$\begin{aligned}&\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \\ &\le \inf_{u\in\mathcal{U}} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] + O\left(\frac{b^2W^3}{L^2}\log^2\left(\frac{m}{L\delta}\right)\left(\frac{d\log(m/\delta)}{m}\right)^{2/3}\right).\end{aligned}$$

To complete the first part of the claim, it remains to choose $m$ as the following value

$$m = \Theta\left(d\log^4(d/(\epsilon\delta))\left(\frac{b^2W^3}{L^2\epsilon}\right)^{3/2}\right).$$

For the second part of the claim, note that $\mathcal{U}_{(a,b)}$ is a closed convex set of functions, and that the infimum $\inf_{u\in\mathcal{U}} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(u(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2]$ is attained by $u^{*t}(z)$. Observe that we have shown that with the sample size $m$ specified above, with probability at least $1-\delta$, it holds

$$\begin{aligned}&\epsilon \ge \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2 - (u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))^2] \\ &\quad= \mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) + u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - 2u^*(\mathbf{w}^*\cdot\mathbf{x}))] = \\ &\quad\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))^2] + 2\,\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - u^*(\mathbf{w}^*\cdot\mathbf{x}))].\end{aligned}$$

Since $\hat{u}^{*t}(z) \in \mathcal{U}_{(a,b)}$, applying the second part of Claim 6.2.4 with $v' = \hat{u}^{*t}$ we get

$$\mathbf{E}[(u^{*t}(\mathbf{w}^t\cdot\mathbf{x}) - \hat{u}^{*t}(\mathbf{w}^t\cdot\mathbf{x}))(u^*(\mathbf{w}^*\cdot\mathbf{x}) - u^{*t}(\mathbf{w}^t\cdot\mathbf{x}))] \ge 0.$$

Thus, we have:

$$\mathbf{E}[(\hat{u}^{*t}(\mathbf{w}^t \cdot \mathbf{x}) - u^{*t}(\mathbf{w}^t \cdot \mathbf{x}))^2] \leq \epsilon.$$

This completes the proof of Lemma D.3.2. □

To prove a similar uniform convergence result for the attainable activations $\hat{u}^t$, we make use of the following fact from prior literature, which shows that we can without loss of generality take the noisy labels to be bounded by $M = O(\frac{bW}{L}\log(bW/\epsilon))$, due to $\mathcal{D}_{\mathbf{x}}$ being $(L, R)$-well behaved.

**Fact D.3.3** (Lemma D.8 [WZDD23])**.** *Let $y' = \text{sign}(y)\min(|y|, M)$ for $M = \frac{bW}{L}\log(\frac{16b^4W^4}{\epsilon^2})$. Then:*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^*(\mathbf{w}^* \cdot \mathbf{x}) - y')^2] = \text{OPT} + \epsilon.$$

In other words, we can assume $|y| \leq M$ without loss of generality by truncating labels that are larger than $M$. Under this assumption, as stated in Lemma D.3.4 below, we bound the $L_2^2$ distance between $\hat{u}^t$ and $u^t$ using similar arguments as in Lemma D.3.2.

**Lemma D.3.4** (Approximating Population-Optimal Activation by Sample-Optimal)**.** *Let $\mathbf{w}^t \in \mathbb{B}(W)$. Given a distribution $\mathcal{D}$ whose marginal $\mathcal{D}_{\mathbf{x}}$ is $(L, R)$-well behaved, let $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$, where $(\mathbf{x}^{(i)}, y^{(i)})$ for $i \in [m]$ are i.i.d. samples from $\mathcal{D}$. Let $\hat{u}^t$ be a sample-optimal activation for the dataset $S$ and parameter vector $\mathbf{w}^t$, as defined in (P). In addition, let $u^t$ be the corresponding population-optimal activation, as defined in (EP). Then, for any $\epsilon, \delta > 0$, choosing a sufficiently large*

$$m \gtrsim d\log^4(d/(\epsilon\delta))\left(\frac{b^2W^3}{L^2\epsilon}\right)^{3/2},$$

*we have that for any $\mathbf{w}^t \in \mathbb{B}(W)$, with probability at least $1-\delta$ over the dataset $S$:*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2] \leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t \cdot \mathbf{x}) - y)^2] + \epsilon\,,$$

*and, furthermore,*

$$\mathbf{E}[(\hat{u}^t(\mathbf{w}^t \cdot \mathbf{x}) - u^t(\mathbf{w}^t \cdot \mathbf{x}))^2] \leq \epsilon.$$

*Proof.* As in the proof of Lemma D.3.2, we choose $r = \frac{2cW}{L}\log(m/(L\delta_1))$ so that $|\mathbf{w}^t \cdot \mathbf{x}^{(i)}| \leq r$ for all $\mathbf{x}^{(i)}$'s from the dataset with probability at least $1 - \delta_1^2/m \geq 1 - \delta_1$. We now condition on the event that $|\mathbf{w}^t \cdot \mathbf{x}^{(i)}| \leq r$ for all $i = 1, \ldots, m$. Let $\mathcal{V}$ be the set of non-decreasing 1-Lipschitz functions such that $\forall v \in \mathcal{V}$, $v(0) = 1/2$, and $v(z_1) - v(z_2) \geq (a/(2br))(z_1 - z_2)$ for all $z_1 \geq z_2 \geq 0$. Then, conditioned on this event, we similarly have that $(\hat{u}^t(z)/(2br) + 1/2)|_{|z|\leq r} = \hat{v}^t(z) \in \mathcal{V}$, and $\hat{v}^t(z)$ satisfies:

$$\hat{v}^t(z) \in \mathop{\text{argmin}}_{v\in\mathcal{V}} \frac{1}{m}\sum_{i=1}^m (v(\mathbf{w}^t \cdot \mathbf{x}^{(i)}) - y^{(i)})^2.$$

Again, studying the $L_2^2$ distance between $\hat{u}^t(z)$ and $u^t(z)$, we have:

$$\begin{aligned}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2] = &\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|\le r\}]\\ &+\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|> r\}].\end{aligned}$$

The probability of $|\mathbf{w}^t\cdot\mathbf{x}|>r$ is small due to the fact that $\mathcal{D}_\mathbf{x}$ possesses sub-exponential tail: $\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}|>r]\le(\delta_1/m)^2$. Now note that $|y|\le M$ and $\mathbf{E}[((\mathbf{w}^t/\|\mathbf{w}^t\|_2)\cdot\mathbf{x})^4]\le c/L^4$ by the sub-exponential property of $\mathcal{D}_\mathbf{x}$, we thus have:

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|>r\}]\\ &\le 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[((\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x}))^2+y^2)\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|>r\}]\\ &\le 2\,\mathbf{E}[b^2W^2((\mathbf{w}^t/\|\mathbf{w}^t\|_2)\cdot\mathbf{x})^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|>r\}]+2M^2\,\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}|>r]\\ &\le 2b^2W^2\sqrt{\mathbf{E}[((\mathbf{w}^t/\|\mathbf{w}^t\|_2)\cdot\mathbf{x})^4]\,\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}|>r]}+2M^2\,\mathbf{Pr}[|\mathbf{w}^t\cdot\mathbf{x}|>r]\\ &\le 2cb^2W^2\delta_1/(L^2m)+2M^2(\delta_1/m)^2,\end{aligned}$$

where in the second inequality we used the fact that $\hat{u}^t$ is $b$-Lipschitz and $\mathbf{w}^t\in\mathbb{B}(W)$, and in the third inequality we applied Cauchy-Schwarz. Since $M=\frac{bW}{L}\log(\frac{16b^4W^4}{\epsilon^2})$, we have $M^2(\delta_1/m)\lesssim cb^2W^2/L^2$ for $m\gtrsim\log(bW/\epsilon)$, thus, we get

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbb{1}\{|\mathbf{w}^t\cdot\mathbf{x}|>r\}]\le 4c(bW/L)^2\delta_1/m, \tag{D.4}$$

for some absolute constant $c$.

The rest remains the same as in the proof of Lemma D.3.2. Let $A=\{\mathbf{x}:|\mathbf{w}^t\cdot\mathbf{x}|\le r\}$. Let $y'=y/(2br)+1/2$. As $\hat{v}^t(z)=\hat{u}^t(z)/(2br)+1/2$ is the sample-optimal activation in $\mathcal{V}$ given $\mathbf{w}^t$ (conditioned on $A$), applying Fact D.3.1 we have that with probability at least $1-\delta$:

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[((\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})/(2br)+1/2)-y')^2|A]=\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{v}^t(\mathbf{w}^t\cdot\mathbf{x})-y')^2|A]\\ &\le\inf_{v\in\mathcal{V}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(v(\mathbf{w}^t\cdot\mathbf{x})-y')^2|A]+\tilde{O}(W(d\log(m/\delta_2)/m)^{2/3}).\end{aligned}$$

Since $\mathcal{U}_{(a,b)}|_{|z|\le r}/(2br)+1/2\subset\mathcal{V}|_{|z|\le r}$, we further have

$$\begin{aligned}\inf_{v\in\mathcal{V}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(v(\mathbf{w}^t\cdot\mathbf{x})-y')^2|A]&\le\inf_{u\in\mathcal{U}_{(a,b)}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}^t\cdot\mathbf{x})/(2br)+1/2-y')^2|A]\\ &\le\frac{1}{4b^2r^2}\inf_{u\in\mathcal{U}_{(a,b)}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}^t\cdot\mathbf{x})-y)^2|A].\end{aligned}$$

Therefore, $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbb{1}\{A\}]$ can be bounded from above by

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbb{1}\{A\}]\\
&= 4b^2r^2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{v}^t(\mathbf{w}^t\cdot\mathbf{x})-y')^2|A]\,\mathbf{Pr}[A]\\
&\le 4b^2r^2\inf_{v\in\mathcal{V}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(v(\mathbf{w}^t\cdot\mathbf{x})-y')^2|A]\,\mathbf{Pr}[A]+\tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3})\\
&\le \inf_{u\in\mathcal{U}_{(a,b)}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}^t\cdot\mathbf{x})-y)^2\mathbb{1}\{A\}]+\tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3})\\
&\le \inf_{u\in\mathcal{U}_{(a,b)}}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u(\mathbf{w}^t\cdot\mathbf{x})-y)^2]+\tilde{O}(b^2r^2W(d\log(m/\delta_2)/m)^{2/3}).
\end{aligned}$$

Thus, combining with (D.4), we get that with probability at least $1-\delta_1-\delta_2$,

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2]\le\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2]+\tilde{O}(Wb^2r^2(\frac{d\log(m/\delta_2)}{m})^{2/3})+(\frac{bW}{L})^2\frac{\delta_1}{m}.$$

Choosing the size of the sample set to be:

$$m=\Theta\left(d\log^4(d/(\epsilon\delta))\left(\frac{b^2W^3}{L^2\epsilon}\right)^{3/2}\right),$$

and recalling that $r=\frac{2cW}{L}\log(m/(L\delta_1))$, we finally have

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2]\le\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[(u^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2]+\epsilon,$$

with probability at least $1-\delta$, after choosing $\delta_1=\delta_2=\delta/2$.

To prove the final claim of the lemma, we follow the same argument as in the proof of Lemma D.3.2. Since we have just shown that with probability at least $1-\delta$, it holds

$$\begin{aligned}
\epsilon &\ge \mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2-(u^t(\mathbf{w}^t\cdot\mathbf{x})-y)^2]\\
&= \mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))^2]+2\,\mathbf{E}[(\hat{u}^t(\mathbf{w}^t\cdot\mathbf{x})-u^t(\mathbf{w}^t\cdot\mathbf{x}))(u^t(\mathbf{w}^t\cdot\mathbf{x})-y)],
\end{aligned}$$

applying the first statement in Claim 6.2.4 completes the proof. □