Learning From Imperfect Data: Noisy Labels, Truncation, and Coarsening

by

Vasilis Kontonis

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2023

Date of final oral examination: 12/07/2023

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

Ilias Diakonikolas, Professor, University of Wisconsin-Madison, Computer Sciences
Yingyu Liang, Assistant Professor, University of Wisconsin-Madison, Computer Sciences
Dimitrios Papailiopoulos, Associate Professor University of Wisconsin-Madison, Electrical and Computer Engineering
Christos Tzamos, Assistant Professor, University of Wisconsin-Madison and National Kapodestrian University of Athens, Computer Sciences

ACKNOWLEDGMENTS

I was very fortunate to be adviced by Prof. Christos Tzamos. His exceptional mentorship and unwavering support were instrumental in making this thesis possible. Christos dedicated countless hours to help me with my research and I learned really a lot from him. Our collaboration was perfect; doing research with Christos is always exciting and fun and I look forward to continuing our collaboration in the future.

I was also very lucky to collaborate closely with Prof. Ilias Diakonikolas during my Ph.D. I want to thank Ilias for all his help and support during my Ph.D.; a big part of this thesis would not be possible without him and in many ways he was like a second advisor to me. Ilias has contributed a lot in both shaping my research directions and my research methodology. I learned a lot from him and I hope that we continue our collaboration in the future.

I extend my thanks to Prof. Yingyu Liang and Prof. Dimitris Papailiopoulos for serving in my committee and for providing feedback and support throughout my Ph.D. Along with other Professors of UW-Madison such as Shuchi Chawla, Rob Nowak, Jelena Diakonikolas, and Sebastien Roch, they contributed a great deal in making UW-Madison an ideal place to learn and do research.

Furthermore, I wish to acknowledge all my collaborators throughout this journey: Manolis, Fotis, Costis, Daniel, Dimitris, Nikos, Piotr, Alkis, Sihan, Constantinos, Ali, Christos, Ilias, Cenk, Stratis, Khoa, Erik, Paul, Gaurav, Mingchen. I learned a lot from collaborating with each one of them and hope to continue working together. A special mention goes to Manolis Zampetakis who helped and promoted me a lot when I started my Ph.D. journey and always has a great piece of advice about research and other aspects of academia (also about roadtrips and national parks). Another special mention goes to my friends and fellow Ph.D. students Nikos and Alkis: we collaborated extensively – also had a great time doing so – and helped each other overcome the challenges of being a Ph.D. student.

I would also like to thank the Distillation team at Google research where I interned: Erik, Gaurav, Fotis, Cenk, and Khoa. I learned a lot from them, worked

on exciting new projects, and had a great time during my internship at Mountain View. A special mention goes to Fotis Iliopoulos who was my co-host during my internship. Fotis was always there to help, and was the main reason that my internship was so productive and fun. Apart from doing research together, we had long discussions that really changed my perspective about many things related to life in academia and industry.

Next, I want to thank my friends and fellow students in UW-Madison with whom we discussed research and often worked together but also shared some great moments during my stay in Madison: Nikos, Konstantinos, Thanasis, Angeliki, Takis, Argiris, Andreas, Elena, Evangelia, Ankit, Sushrut, Eleni, Giorgos, Jasper, Efi, Yuxin, Lisheng, Shashank.

Moving to friends and colleagues from Greece, I want to thank all my friends and professors from Corelab at National Technical University of Athens: Dimitris, Aris, Stathis, Stratis, Orestis, Alkis, Thanasis, Eleni. Corelab is a great community and interacting with its members was the main inspiration for me to pursue a Ph.D. in theoretical computer science.

I am deeply grateful to my parents Andreas and Katerina, my siblings, Asimina and Ilias, my partner Georgia, and my friends, Vangelis, Anastasia, Stavros, Mina, Angeliki, Giannis, Giorgos, Ioanna, Melina, Korina, Takis, Marios, Nikos, Mitsos, Vasiliki, Nikos, Giannis, Constantia, Nikos, Vasilis. Their love and support is a constant motivation for me.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LEARNING FROM IMPERFECT DATA: NOISY LABELS, TRUNCATION, AND COARSENING

Vasilis Kontonis

Under the supervision of Professor Christos Tzamos
At the University of Wisconsin-Madison

The datasets used in machine learning and statistics are *huge* and often *imperfect*, *e.g.*, they contain corrupted data, examples with wrong labels, or hidden biases. Most existing approaches (i) produce unreliable results when the datasets are corrupted, (ii) are computationally inefficient, or (iii) come without any theoretical/provable performance guarantees. In this thesis, we *design learning algorithms* that are **computationally efficient** and at the same time **provably reliable**, even when used on imperfect datasets.

We first focus on supervised learning settings with noisy labels. We present efficient and optimal learners under the semi-random noise models of Massart and Tsybakov – where the true label of each example is flipped with probability at most 50% – and an efficient approximate learner under adversarial label noise – where a small but arbitrary fraction of labels is flipped – under structured feature distributions. Apart from classification, we extend our results to noisy label-ranking.

In truncated statistics, the learner does not observe a representative set of samples from the whole population, but only truncated samples, *i.e.*, samples from a potentially small subset of the support of the population distribution. We give the first efficient algorithms for learning Gaussian distributions with unknown truncation sets and initiate the study of non-parametric truncated statistics. Closely related to truncation is *data coarsening*, where instead of observing the class of an example, the learner receives a set of potential classes, one of which is guaranteed to be the correct class. We initiate the theoretical study of the problem, and present the first efficient learning algorithms for learning from coarse data.

Christos Tzamos

1 OVERVIEW

Here we summarize the results of this thesis and provide an overview of its structure. We first briefly present the contents of each chapter and then provide a more detailed summary of the novel contributions included in each one.

Chapter 2 We present the first polynomial time algorithm for learning Massart halfspaces with respect to a broad class of structured distributions. Based on Diakonikolas et al. (2020c) that was published in the Conference on Learning Theory (COLT).

Chapter 3 We present a quasi-polynomial time algorithm for learning Tsybakov halfspaces under structured distributions. This work introduced a novel and general new learning framework based on certifying non-optimality of hypotheses. Based on Diakonikolas et al. (2021b) that was published in the Symposium on Theory of Computing (STOC) (merged with Diakonikolas et al. (2021a)).

Chapter 4 We improve upon the quasi-polynomial algorithm of Chapter 3 and provide a polynomial-time algorithm for Tsybakov halfspaces under structured marginals. Based on Diakonikolas et al. (2021a) that was published in the Symposium on Theory of Computing (STOC) (merged with Diakonikolas et al. (2021b)).

Chapter 5 We provide the first efficient algorithms for noisy linear label ranking. Based on Fotakis et al. (2022) that was published in the Conference on Learning Theory (COLT).

Chapter 6 We provide the first efficient algorithm for learning Gaussian distributions with unknown truncation sets. Our method only requires that the (Gaussian) surface area of the uknown truncation set is bounded. Based on Kontonis et al. (2019) that was published in Foundations of Computer Science (FOCS).

Chapter 7 We initiate the study of non-parametric truncated statistics and provide the first efficient algorithms for recovering truncated smooth densities. Based on Daskalakis et al. (2021) that was published in the Conference on Learning Theory (COLT).

Chapter 8 We initiate the theoretical study of inference from coarse labels and establish a connection with Statistical Query learning algorithms. Based on Fotakis et al. (2021a) that was published in the Conference on Learning Theory (COLT).

1.1 Learning Halfspaces with Massart Noise

The algorithmic problem of learning an unknown halfspace from random labeled examples has been extensively investigated since the 1950s — starting with Rosenblatt's Perceptron algorithm (Rosenblatt, 1958) — and has arguably been one of the most influential problems in the field of machine learning. In the realizable case, i.e., when all the labels are consistent with the target halfspace, this learning problem amounts to linear programming, hence can be solved in polynomial time (see, e.g., Maass and Turan (1994); Shawe-Taylor and Cristianini (2000)). The problem turns out to be much more challenging algorithmically in the presence of noisy labels, and its computational complexity crucially depends on the noise model.

We study the problem of distribution-specific PAC learning of halfspaces in the presence of Massart noise (Massart and Nedelec, 2006). In the Massart noise model, an adversary can flip each label independently with probability *at most* $\eta < 1/2$, and the goal of the learner is to reconstruct the target halfspace to arbitrarily high accuracy. More formally, we have:

Definition 1.1 (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, $0 \leq \eta < 1/2$, 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 $(\boldsymbol{x}, y)$, such that: (a) $\boldsymbol{x} \sim D_{\boldsymbol{x}}$,* *where $D_{\boldsymbol{x}}$ is a fixed distribution in $\mathcal{F}$, and (b) $y = f(\boldsymbol{x})$ with probability $1 - \eta(\boldsymbol{x})$ and $y = -f(\boldsymbol{x})$ with probability $\eta(\boldsymbol{x})$, for an* unknown *parameter $\eta(\boldsymbol{x}) \leq \eta$. Let D denote the joint distribution on $(\boldsymbol{x}, y)$ generated by the above oracle. A learning algorithm is given i.i.d. samples from D and its goal is to output a hypothesis h such that with high probability h is ϵ-close to f, i.e., it holds $\mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})] \leq \epsilon$.*

Massart noise is a realistic model of random noise that has attracted significant attention in recent years (see Section 1.1 for a summary of prior work). This noise model goes back to the 80s, when it was studied by Rivest and Sloan (Sloan, 1988; Rivest and Sloan, 1994a) under the name "malicious misclassification noise", and a very similar asymmetric noise model was considered even earlier by Vapnik (Vapnik, 1982). The Massart noise condition lies in between the Random Classification Noise (RCN) (Angluin and Laird, 1988) – where each label is flipped independently with probability *exactly* $\eta < 1/2$ – and the agnostic model (Haussler, 1992; Kearns et al., 1994a) – where an adversary can flip any small constant fraction of the labels.

The sample complexity of PAC learning with Massart noise is well-understood. Specifically, if $\mathcal{C}$ is the class of d-dimensional halfspaces, it is well-known (Massart and Nedelec, 2006) that $O(d/(\epsilon \cdot (1-2\eta)^2))$ samples information-theoretically suffice to determine a hypothesis h that is ϵ-close to the target halfspace f with high probability (and this sample upper bound is best possible). The question is whether a computationally efficient algorithm exists.

The algorithmic question of efficiently computing an accurate hypothesis in the distribution-specific PAC setting with Massart noise was initiated in Awasthi et al. (2015), and subsequently studied in a sequence of works (Awasthi et al., 2016a; Zhang et al., 2017b; Yan and Zhang, 2017a; Mangoubi and Vishnoi, 2019a). This line of work has given polynomial-time algorithms for learning halfspaces with Massart noise, when the underlying marginal distribution $D_{\boldsymbol{x}}$ is the uniform distribution on the unit sphere (i.e., the family $\mathcal{F}$ in Definition 1.1 is a singleton).

The question of designing a computationally efficient learning algorithm for this problem that succeeds under more general distributional assumptions remained open, and has been posed as an open problem in a number of places (Awasthi et al., 2016a; Awasthi, 2018; Balcan and Haghtalab, 2020a). Specifically, Awasthi et al. (2016a) asked whether there exists a polynomial-time algorithm for all log-concave distributions, and the same question was more recently highlighted in Balcan and Haghtalab (2020a). In more detail, Awasthi et al. (2016a) gave an algorithm that succeeds under any log-concave distribution, but has sample complexity and running time $d^{2^{\mathrm{poly}(1/(1-2\eta))}}/\mathrm{poly}(\epsilon)$, i.e., doubly exponential in $1/(1-2\eta)$. Balcan and Haghtalab (2020a) asked whether a $\mathrm{poly}(d, 1/\epsilon, 1/(1-2\eta))$ time algorithm exists for log-concave marginals. As a corollary of our main algorithmic result (Theorem 2.3), we answer this question in the affirmative. Perhaps surprisingly, our algorithm is extremely simple (performing SGD on a natural non-convex surrogate) and succeeds for a broader family of structured distributions, satisfying certain (anti)-anti-concentration and tail bound properties.

Main Results and Techniques

Our main result is the first polynomial-time algorithm for learning halfspaces with Massart noise with respect to a broad class of well-behaved distributions. Before we formally state our algorithmic result, we define the family of distributions $\mathcal{F}$ for which our algorithm succeeds:

Definition 1.2 (Well-Behaved Distributions)**.** *For $L, R, U, \beta > 0$ a distribution $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ is called (L, R, U, β)-well-behaved if for any projection $(D_{\boldsymbol{x}})_V$ of $D_{\boldsymbol{x}}$ on a subspace V of $\mathbb{R}^d$ of dimension at most 3, the corresponding pdf γ_V on V satisfies the following properties:*

1. *$\gamma_V(\boldsymbol{x}) \leq U$ for all $\boldsymbol{x} \in V$ (anti-concentration).*

2. *$\gamma_V(\boldsymbol{x}) \geq L$, for all $\boldsymbol{x} \in V$ with $\|\boldsymbol{x}\|_2 \leq R$ (anti-anti-concentration).*

3. *$\mathbf{Pr}_{\boldsymbol{x}\sim(D_x)_V}[\|\boldsymbol{x}\|_2 \geq t] \leq \exp(1 - t/\beta)$ (sub-exponential concentration).*

When the parameters L, R, U, β are all universal constants (independent from any other parameter of the problem) we will simply say that the distribution is well-behaved without specifying the constants explicitly.

We remark that the above class contains many distributions previously considered in the literature, such as the standard normal and isotropic log-concave distributions. In particular, the corresponding constants in Definition 1.2 for the standard normal and isotropic log-concave distributions the definition is satisfied with the corresponding parameters L, R, U, β being universal constants. Our algorithm for Massart Noise succeeds under more general assumptions (namely it does not require the sub-exponential concentration, see Definition 2.2) for a more precise definition.

Our main result for learning with Massart noise is the following.

Theorem 1.3 (Informal – Learning Halfspaces with Massart Noise)**.** *There is a computationally efficient algorithm that learns halfspaces in the presence of Massart noise with respect to any well-behaved distribution on $\mathbb{R}^d$. Specifically, the algorithm draws $m = \mathrm{poly}\,(1/(1-2\eta)) \cdot O(d/\epsilon^4)$ samples from a noisy example oracle at rate $\eta < 1/2$, runs in sample-polynomial time, and outputs a hypothesis halfspace h that is ϵ-close to the target with probability at least $9/10$.*

See Theorem 2.3 for a more detailed statement. Theorem 1.3 provides the first polynomial-time algorithm for learning halfspaces with Massart noise under a fairly broad family of well-behaved distributions. Specifically, our algorithm runs in $\mathrm{poly}(d, 1/\epsilon, 1/(1-2\eta))$ time, as long as the parameters R, U are bounded above by some $\mathrm{poly}(d)$, and the function $t(\epsilon)$ is bounded above by some $\mathrm{poly}(d/\epsilon)$. These conditions do not require a specific parametric or nonparametric form for the underlying density and are satisfied for several reasonable continuous distribution families.

As we mentioned earlier, is not hard to show that the class of isotropic log-concave distributions is (U, R, L, β)-bounded, for $U, L, R, \beta = \Theta(1)$ (see Fact A.5). Similar implications hold for a broader class of distributions, known as s-concave distributions. (See Appendix A.1.) Therefore, we immediately obtain the following corollary:

Corollary 1.4 (Learning Halfspaces with Massart Noise Under Log-concave Distributions)**.** *There exists a polynomial-time algorithm that learns halfspaces with Massart* *noise under any isotropic log-concave distribution. The algorithm has sample complexity* $m = \widetilde{O}(d/\epsilon^4) \cdot \mathrm{poly}(1/(1-2\eta))$ *and runs in sample-polynomial time.*

Corollary 1.4 gives the first polynomial-time algorithm for this problem, answering an open question of Awasthi et al. (2016a); Awasthi (2018); Balcan and Haghtalab (2020a). We obtain similar implications for s-concave distributions. (See Appendix A.1 for more details.)

While the preceding discussion focused on polynomial learnability, our algorithm establishing Theorem 1.3 is extremely simple and can potentially be practical. Specifically, our algorithm simply performs SGD (with projection on the unit ball) on a natural *non-convex* surrogate loss, namely an appropriately smoothed version of the misclassification error function, $\mathrm{err}_{0-1}^{\mathcal{D}}(\boldsymbol{w}) = \mathbf{Pr}_{(\boldsymbol{x},y)\sim D}[\mathrm{sign}(\boldsymbol{x} \cdot \boldsymbol{w}) \neq y]$. We also note that the sample complexity of our algorithm for log-concave marginals is optimal as a function of the dimension d, within constant factors.

Our approach for establishing Theorem 1.3 is fairly robust and immediately extends to a slightly stronger noise model, considered in Zhang et al. (2017b), which we term *strong Massart noise*. In this model, the flipping probability can be arbitrarily close to $1/2$ for points that are very close to the true separating hyperplane. These implications are stated and proved in Section 2.5.

Prior and Related Work

We start with a summary of prior work on distribution-specific PAC learning of halfspaces with Massart noise. The study of this learning problem was initiated in Awasthi et al. (2015). That work gave the first polynomial-time algorithm for the problem that succeeds under the uniform distribution on the unit sphere, assuming the upper bound on the noise rate η is smaller than a sufficiently small constant ($\approx 10^{-6}$). Subsequently, Awasthi et al. (2016a) gave a learning algorithm with sample and computational complexity $d^{2^{\mathrm{poly}(1/(1-2\eta))}}/\mathrm{poly}(\epsilon)$ that succeeds for any noise rate $\eta < 1/2$ under any log-concave distribution.

The approach in Awasthi et al. (2015, 2016a) uses an iterative localization-based method. These algorithms operate in a sequence of phases and it is shown that they make progress in each phase. To achieve this, Awasthi et al. (2015, 2016a) leverage a distribution-specific agnostic learner for halfspaces (Kalai et al., 2008) and develop sophisticated tools to control the trajectory of their algorithm.

Inspired by the localization approach, Yan and Zhang (2017a) gave a perceptron-like algorithm (with sample complexity linear in d) for learning halfspaces with Massart noise under the uniform distribution on the sphere. Their algorithm again proceeds in phases and crucially exploits the symmetry of the uniform distribution to show that the angle between the current hypothesis $\widehat{\boldsymbol{w}}^{(i)}$ and the target halfspace $\boldsymbol{w}^*$ decreases in every phase. Zhang et al. (2017b) also gave a polynomial-time algorithm for learning halfspaces with Massart noise under the uniform distribution on the unit sphere. Their algorithm works in the strong Massart noise model and is based on the Stochastic Gradient Langevin Dynamics (SGLD) algorithm applied to a smoothed version of the empirical $0-1$ loss. Their method leads to sample complexity $\Omega_\eta(d^4/\epsilon^4)$ and its running time involves $\Omega_\eta(d^{13.5}/\epsilon^{16})$ inner product evaluations. More recently, Mangoubi and Vishnoi (2019a) improved these bounds to $\Omega_\eta(d^{8.2}/\epsilon^{11.4})$ inner product evaluations via a similar approach. Our method is much simpler in comparison, running SGD directly on the population loss and using one sample per iteration with a significantly improved sample complexity and running time.

Furthermore, in contrast to the aforementioned approaches, we study a more general setting (in the sense that our method works for a broad family of distributions), and our approach is not tied to the iterations of any particular algorithm. Our structural lemma (Lemma 2.6) shows that *any* approximate stationary point of our non-convex surrogate loss suffices. As a consequence, one can apply any first-order method that converges to stationarity (and in particular vanilla SGD with projection on the unit sphere works). The upshot is that we do not need to establish guarantees for the trajectory of the method used to reach such a stationary point. The only thing that matters is the endpoint of the algorithm. Intriguingly, for a generic distribution in the class we consider, it is unclear if it is possible to establish a monotonicity property for a first-order method reaching a stationary point.

We note that the d-dependence in the sample complexity of our algorithm is information-theoretically optimal, even under Gaussian marginals. The ϵ-dependence seems tight for our approach, given recent lower bounds for the convergence of SGD (Drori and Shamir, 2019), or any stochastic first-order method (Arjevani et al., 2019), to stationary points of smooth non-convex functions.

Finally, we comment on the relation to a recent work on distribution-independent PAC learning of halfspaces with Massart noise (Diakonikolas et al., 2019a). Diakonikolas et al. (2019a) gave a distribution-independent PAC learner for halfspaces with Massart noise that approximates the target halfspace within misclassification error $\approx \eta$, i.e., it does not yield an arbitrarily close approximation to the true function. In contrast, the aforementioned distribution-specific algorithms achieve information-theoretically optimal misclassification error, which implies that the output hypothesis can be arbitrarily close to the true target halfspace. As a result, the results of this work are not subsumed by Diakonikolas et al. (2019a).

Comparison to RCN and Agnostic Settings It is instructive to compare the complexity of learning halfspaces in the Massart model with two related noise models. In the RCN model, a polynomial-time algorithm is known in the distribution-independent PAC model (Blum et al., 1996, 1997). In sharp contrast, even weak agnostic learning is hard in the distribution-independent setting (Guruswami and Raghavendra, 2006; Feldman et al., 2006a; Daniely, 2016a). Moreover, obtaining information-theoretically optimal error guarantees remains computationally hard in the agnostic model, even when the marginal distribution is the standard Gaussian (Klivans and Kothari, 2014) (assuming the hardness of noisy parity). On the other hand, recent work (Awasthi et al., 2017; Diakonikolas et al., 2018b) has given efficient algorithms (for Gaussian and log-concave marginals) with error $O(\mathrm{opt}) + \epsilon$, where opt is the misclassification error of the optimal halfspace.

1.2 Learning with Tsybakov Noise

We study the algorithmic problem of learning halfspaces under a well-known generalization of the Massart Noise model: the Tsybakov noise condition Tsybakov (2004). The Tsybakov noise model is a challenging noise model that has been extensively studied in the statistics and machine learning communities. While the information-theoretic aspects of learning with Tsybakov noise have been largely characterized, prior to this work, the computational aspects of this broad problem had remained wide open.

The Tsybakov noise condition prescribes that the label of each example is independently flipped with some probability which is controlled by an adversary. Importantly, this noise condition allows the flipping probabilities to be *arbitrarily close to* $1/2$ for a fraction of the examples. More formally, we have the following definition:

Definition 1.5 (PAC Learning with Tsybakov Noise)**.** *Let $\mathcal{C}$ be a concept class of Boolean-valued functions over $X = \mathbb{R}^d$, $\mathcal{F}$ be a family of distributions on X, $0 < \epsilon < 1$ be the error parameter, and $0 \leq \alpha < 1$, $A > 0$ be parameters of the noise model. Let f be an unknown target function in $\mathcal{C}$. A* Tsybakov example oracle, $\mathrm{EX}^{\mathrm{Tsyb}}(f, \mathcal{F})$, *works as follows: Each time $\mathrm{EX}^{\mathrm{Tsyb}}(f, \mathcal{F})$ is invoked, it returns a labeled example $(\boldsymbol{x}, y)$, such that: (a) $\boldsymbol{x} \sim D_{\boldsymbol{x}}$, where $D_{\boldsymbol{x}}$ is a fixed distribution in $\mathcal{F}$, and (b) $y = f(\boldsymbol{x})$ with probability $1 - \eta(\boldsymbol{x})$ and $y = -f(\boldsymbol{x})$ with probability $\eta(\boldsymbol{x})$. Here $\eta(\boldsymbol{x})$ is an* unknown *function that satisfies the Tsybakov noise condition with parameters (α, A). That is, for any $0 < t \leq 1/2$, $\eta(\boldsymbol{x})$ satisfies the condition $\mathbf{Pr}_{\boldsymbol{x} \sim D_{\boldsymbol{x}}}[\eta(\boldsymbol{x}) \geq 1/2 - t] \leq A\, t^{\frac{\alpha}{1-\alpha}}$.*

Let D denote the joint distribution on $(\boldsymbol{x}, y)$ generated by the above oracle. A learning algorithm is given i.i.d. samples from D and its goal is to output a hypothesis function $h : X \to \{\pm 1\}$ such that with high probability h is ϵ-close to f, i.e., it holds $\mathbf{Pr}_{\boldsymbol{x} \sim D_{\boldsymbol{x}}}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})] \leq \epsilon$.

Motivation for Tsybakov Noise Model The bounded (Massart) noise assumption, i.e., that the probability that labels are flipped is *globally* bounded away from

$1/2$, fails to accurately capture a number of practically relevant noise sources, including the *human annotator noise* Klebanov and Beigman (2010, 2009); Beigman and Klebanov (2009); Chhikara and McKeon (1984). In particular, the humans responsible for labeling the training data are much more prone to incorrectly classify points closer to the decision boundary (where “cats" and “dogs" look almost the same), than points far from the boundary. For example, it was empirically shown in Klebanov and Beigman (2010) that when non-expert annotators (Amazon Mechanical Turk) were used to annotate the RTE-1 dataset Dagan et al. (2005), roughly 20% of the datapoints were classified almost at random, i.e., had $\eta(\boldsymbol{x}) \approx 1/2$. More broadly, a long line of research (both applied and theoretical) Castro and Nowak (2008); Frénay and Verleysen (2013); Zhang et al. (2017b); Menon et al. (2018); Hopkins et al. (2020); Diakonikolas et al. (2020c) focuses on noise models that do not restrict the flipping probability globally, but allow it to be arbitrarily close to $1/2$ near the decision boundary. On the other hand, since datapoints from low-density regions are also likely to be classified almost randomly (see, e.g., Frénay and Verleysen (2013) and references therein), assuming that high noise rates occur only close to the decision boundary does not sufficiently capture these situations.

The Tsybakov noise model Mammen and Tsybakov (1999) provides a unified framework that significantly extends the Massart noise condition to capture the above scenarios: it prescribes that the label of each example is independently flipped with some probability which is controlled by an adversary, but is not uniformly bounded by a constant less than $1/2$. In particular, it allows the flipping probabilities to be *arbitrarily close to* $1/2$ for a fraction of the examples. Importantly, it makes *no geometric assumptions* about the noise, e.g., that it is only potentially large close to the decision boundary.

The noise model of Definition 1.5 was first proposed in Mammen and Tsybakov (1999) and subsequently refined in Tsybakov (2004). Since these initial works, a long line of research in statistics and learning theory has focused on understanding a range of statistical aspects of the model in various settings (see, e.g., Tsybakov (2004); Boucheron et al. (2005); Bartlett et al. (2006); Balcan et al. (2007); Hanneke

(2011); Hanneke and Yang (2015) and references therein). Ignoring computational considerations, it is known that the class of halfspaces is learnable in this model with $\mathrm{poly}(d, 1/\epsilon^{1/\alpha})$ samples, where d is the dimension and ϵ is the error to the target halfspace.

On the other hand, the algorithmic question has remained poorly understood. Roughly speaking, the only known algorithms in this noise model (for any non-trivial concept class in high dimension) are the ones that follow via the naive reduction to agnostic learning.

Main Results and Techniques

As explained in the above discussion, obtaining computationally efficient learning algorithms in the presence of Tsybakov noise in *any* non-trivial setting — that is, for any natural concept class and under any distributional assumptions — has been a long-standing open problem in learning theory. *Our main algorithmic result resolves the complexity of learning halfspaces in this model.*

We first present our result for learning under the well-behaved class of distributions defined in Definition 1.2.

Theorem 1.6 (Informal – Learning Tsybakov Halfspaces under Well-Behaved Distributions)**.** *Let D be a well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$. There exists an algorithm that draws $N = O_{A,\alpha}(d/\epsilon)^{O(1/\alpha)}$ samples from D, runs in $\mathrm{poly}(N, d)$ time, and computes a vector $\widehat{\boldsymbol{w}}$ such that, with high probability we have that $\mathrm{err}_{0-1}^{D_x}(h_{\widehat{\boldsymbol{w}}}, f) \leq \epsilon$.*

See Theorem 4.39 for a more detailed statement. We remark that for the special case of isotropic log-concave densities we are able to obtain a more efficient algorithm with sample complexity and runtime $\mathrm{poly}(d)\, O(A/\epsilon)^{O(1/\alpha^2)}$, see Theorem 4.40. Since the sample complexity of the problem is $\mathrm{poly}(d, 1/\epsilon^{1/\alpha})$, the algorithm of Theorem 4.3 is qualitatively close to best possible.

The Certificate Framework The main technical novelty of our works for learning with Tsybakov noise is the development of the certificate framework. The certificate framework was introduced in our work Diakonikolas et al. (2021b) where a quasipolynomial time algorithm for learning with Tsybakov noise was given. In the follow-up work Diakonikolas et al. (2021a), we improved the runtime of the certificate algorithm to polynomial which yielded Theorem 1.6.

At a high-level, this framework allows us to efficiently reduce the problem of *finding* a near-optimal halfspace to the (easier) problem of *certifying* whether a candidate halfspace $h_{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x})$ is "far" from the optimal halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$. The idea is to use a certificate algorithm (as a black-box) and combine it with an online convex optimization routine. Roughly speaking, starting from an initial guess $\boldsymbol{w}_0$ for $\boldsymbol{w}^*$, a judicious combination of these two ingredients allows us to efficiently compute a near-optimal halfspace $\widehat{\boldsymbol{w}}$, i.e., one that the certifying algorithm cannot reject. We note that a similar approach has been used in Chen et al. (2020a) for converting non-proper learners to proper learners in the Massart noise model.

The key idea to design a certificate in the Tsybakov noise model is the following simple but crucial observation: If $\boldsymbol{w}^*$ is the normal vector to true halfspace, then for any non-negative function $T(\boldsymbol{x})$, it holds that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x}) y\, \boldsymbol{w}^* \cdot \boldsymbol{x}] \geq 0$. On the other hand, for any $\boldsymbol{w} \neq \boldsymbol{w}^*$ there exists a non-negative function $T(\boldsymbol{x})$ such that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\, y\, \boldsymbol{w} \cdot \boldsymbol{x}] < 0$. In other words, there exists a *reweighting of the space* that makes the expectation of $y\boldsymbol{w} \cdot \boldsymbol{x}$ negative (Fact 3.3). Note that we can always use as $T(\boldsymbol{x})$ the indicator of the disagreement region between the candidate halfspace $h_{\boldsymbol{w}}(\boldsymbol{x})$ and the optimal halfspace $f(\boldsymbol{x}) = h_{\boldsymbol{w}^*}(\boldsymbol{x})$.

Of course, since optimizing over the space of non-negative functions is intractable, we need to restrict our search space to a "simple" parametric family of functions.

Certificates via Low-Degree Polynomials Diakonikolas et al. (2021b) We start by showing that given a candidate hypothesis $\boldsymbol{w}$ that is "far" from being optimal, that is the angle $\theta(\boldsymbol{w}, \boldsymbol{w}^*)$ is bounded away from zero, we can construct a *low* *complexity* certificate F that will satisfy $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[F(\boldsymbol{x})\boldsymbol{w}\cdot\boldsymbol{x}y] < 0$. In particular, we construct a certificate that is the product of a square of a low degree non-negative polynomial and an indicator function that depends on the hypothesis $\boldsymbol{w}$.

Proposition 1.7 (Informal – Low Degree Certificate). *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the Tsybakov noise condition with parameters (α, A) and the marginal $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ well-behaved. Let $\boldsymbol{w}^* \in \mathbb{S}^{d-1}$ be the normal vector to the optimal halfspace and $\widehat{\boldsymbol{w}} \in \mathbb{S}^{d-1}$ be such that the dissagreement probability of its corresponding halfspace hypothesis $h_{\widehat{\boldsymbol{w}}}$ with the optimal halfpsace f is at least ϵ: $\mathrm{err}_{0-1}^{D_{\boldsymbol{x}}}(h_{\widehat{\boldsymbol{w}}}, f) \geq \epsilon$. There exists polynomial $p : \mathbb{R}^d \mapsto \mathbb{R}$ of degree $k = O_{\alpha,A}(\log^2(1/\epsilon))$ satisfying $\|p\|_2^2 \leq d^{O(k)}$ such that*

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}\left[p(\boldsymbol{x})^2\, \mathbb{1}\{0 \leq \boldsymbol{w}\cdot\boldsymbol{x} \leq \Theta(\epsilon)\}\, y\boldsymbol{w}\cdot\boldsymbol{x}\right] \leq -\Omega(\epsilon)\ .$$

We show that we can efficiently compute our polynomial certificate given labeled examples from the target distribution. The crucial property that enables efficient computation of our polynomial certificate is that it is a square of some low-degree polynomial. Therefore, we can solve the following relaxation using Sum-of-Squares (SoS) optimization (for which efficient algorithms based on Semidefinite programming exist, see, e.g., Fleming et al. (2019)):

$$\min_{p\in\mathrm{SoS},\deg(p)\leq k}\ \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}\left[p(\boldsymbol{x})\ \mathbb{1}_B(\boldsymbol{x})\boldsymbol{w}\cdot\boldsymbol{x}y\right]$$

For more details, we refer to Section 3.4.

Certificates via Intersections of Halfspaces Diakonikolas et al. (2021a) The previous approach relying on low-degree polynomials is inherently limited to quasi-polynomial time (see also Section 4.3) and new ideas are needed to obtain a polynomial time algorithm.

In the work Diakonikolas et al. (2021a), we considered certifying functions of the form:

$$T(\boldsymbol{x}; \boldsymbol{v}, \sigma_1, \sigma_2, t_1, t_2) = \\ \frac{1}{\boldsymbol{w} \cdot \boldsymbol{x}} \mathbb{1} \left\{ \sigma_1 \le \boldsymbol{w} \cdot \boldsymbol{x} \le \sigma_2 \,, -t_1 \le \boldsymbol{v} \cdot \mathrm{proj}_{\boldsymbol{w}^\perp} \frac{\boldsymbol{x}}{\boldsymbol{w} \cdot \boldsymbol{x}} \le -t_2 \right\}$$

that are parameterized by a vector $\boldsymbol{v}$ and scalar thresholds $\sigma_1, \sigma_2, t_1, t_2 > 0$. Here $\mathrm{proj}_{\boldsymbol{w}^\perp}$ denotes the orthogonal projection on the subspace orthogonal to $\boldsymbol{w}$.

We observe that, in contrast with the $d^{\mathrm{poly}(\log(1/\epsilon))}$ parameters needed to specify a polynomial of degree $\mathrm{poly}(\log(1/\epsilon))$ in d dimensions, the above function class can be specified by $O(d)$ parameters. Of course, it may not be a priori clear why functions of this form can be used as certifying functions in our setting. The intuition behind choosing functions of this simple form is given in Section 4.3. In particular, in Claim 4.6, we show that for any incorrect guess $\boldsymbol{w}$ there *exists* a *certifying vector* $\boldsymbol{v}$ that makes the expectation $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x}) \, y \, \boldsymbol{w} \cdot \boldsymbol{x}]$ negative. In fact, the vector $\boldsymbol{v} = \mathrm{proj}_{\boldsymbol{w}^\perp} \boldsymbol{w}^* / \left\| \mathrm{proj}_{\boldsymbol{w}^\perp} \boldsymbol{w}^* \right\|_2 := (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ suffices for this purpose.

The key challenge is in finding such a certifying vector $\boldsymbol{v}$ algorithmically. We note that our algorithm in general does not find $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. But it does find a vector $\boldsymbol{v}$ with similar behavior, in the sense of making the $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x}) \, y \, \boldsymbol{w} \cdot \boldsymbol{x}]$ sufficiently negative. To achieve this goal, we take a two-step approach: The first step involves computing an initialization vector $\boldsymbol{v}_0$ that has non-trivial correlation with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. In our second step, we give a perceptron-like update rule that iteratively improves the initial guess until it converges to a certifying vector $\boldsymbol{v}$. While this algorithm is relatively simple, its correctness relies on a win-win analysis (Lemma 4.14) whose proof is quite elaborate. In more detail, we show that for any *non-certifying* vector $\boldsymbol{v}$ that is sufficiently correlated with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$, we can efficiently compute a direction that improves its correlation to $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. We then argue (Lemma 4.19) that by choosing an appropriate step size this iteration converges to a certifying vector within a small number of steps. We show the next result:

Theorem 1.8 (Informal – Efficiently Certifying Non-Optimality)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the Tsybakov noise condition with parameters (α, A)*

and the marginal $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ well-behaved. Let f be the optimal halfspace and $\widehat{\boldsymbol{w}} \in \mathbb{S}^{d-1}$ be such that the dissagreement probability of its corresponding halfspace hypothesis $h_{\widehat{\boldsymbol{w}}}$ with the optimal halfpsace f is at least ϵ: $\mathrm{err}_{0-1}^{D_x}(h_{\widehat{\boldsymbol{w}}}, f) \geq \epsilon$. There is an algorithm that, given $N = \Big((A) \cdot (d/\epsilon)\Big)^{O(1/\alpha)}$ samples from D, runs in sample polynomial time, and with high probability returns parameters $\boldsymbol{v} \in \mathbb{R}^d$, $\sigma_1, \sigma_2, t_1, t_2$ such that

$$\mathop{\mathbf{E}}_{(x,y)\sim D}[T(\boldsymbol{x}; \boldsymbol{v}, \sigma_1, \sigma_2, t_1, t_2)\; y\boldsymbol{w} \cdot \boldsymbol{x}] \leq -\left(\frac{\epsilon}{A\,d}\right)^{O(1/\alpha)} . \tag{1.1}$$

For more details, we refer to Chapter 4.

1.3 Noisy Label Ranking

Label Ranking (LR) is the problem of learning a hypothesis that maps features to rankings over a finite set of labels. Given a feature vector $\boldsymbol{x} \in \mathbb{R}^d$, a sorting function $\sigma(\cdot)$ maps it to a ranking of k alternatives, i.e., $\sigma(\boldsymbol{x})$ is an element of the symmetric group with k elements, $\mathbb{S}_k$. Assuming access to a training dataset of features labeled with their corresponding rankings, i.e., pairs of the form $(\boldsymbol{x}, \pi) \in \mathbb{R}^d \times \mathbb{S}_k$, the goal of the learner is to find a sorting function $h(\boldsymbol{x})$ that generalizes well over a fresh sample. LR has received significant attention over the years Dekel et al. (2003); Shalev-Shwartz (2007); Hüllermeier et al. (2008); Cheng and Hüllermeier (2008); Fürnkranz et al. (2008) due to the large number of applications. For example, ad targeting Djuric et al. (2014) is an LR instance where for each user we want to use their feature vector to predict a ranking over ad categories and present them with the most relevant. The practical significance of LR has lead to the development of many techniques based on probabilistic models and instance-based methods Cheng and Hüllermeier (2008); Cheng et al. (2010), Grbovic et al. (2012); Zhou et al. (2014a), decision trees Cheng et al. (2009), entropy-based ranking trees Rebelo de Sá et al. (2015), bagging Aledo et al. (2017), and random forests de Sá et al. (2017); Zhou and Qiu (2018). However, almost all of these works come without provable guarantees and/or fail to learn in the presence of noise in the observed rankings.

Linear Sorting Functions (LSFs). In this work, we focus on the fundamental concept class of Linear Sorting functions Har-Peled et al. (2003). A linear sorting function parameterized by a matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ with k rows $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_k$ takes a feature $\boldsymbol{x} \in \mathbb{R}^d$, maps it to $\boldsymbol{W}\boldsymbol{x} = (\boldsymbol{W}_1 \cdot \boldsymbol{x}, \ldots, \boldsymbol{W}_k \cdot \boldsymbol{x}) \in \mathbb{R}^k$ and then outputs an ordering $(i_1, \ldots, i_k)$ of the k alternatives such that $\boldsymbol{W}_{i_1} \cdot \boldsymbol{x} \geq \boldsymbol{W}_{i_2} \cdot \boldsymbol{x} \geq \ldots \geq \boldsymbol{W}_{i_k} \cdot \boldsymbol{x}$. In other words, a linear sorting function ranks the k alternatives (corresponding to rows of $\boldsymbol{W}$) with respect to how well they correlate with the feature $\boldsymbol{x}$. We denote a linear sorting function with parameter $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ by $\sigma_{\boldsymbol{W}}(\boldsymbol{x}) \triangleq \mathrm{argsort}(\boldsymbol{W}\boldsymbol{x})$ where $\mathrm{argsort} : \mathbb{R}^k \to \mathbb{S}_k$ takes as input a vector $(v_1, \ldots, v_k) \in \mathbb{R}^k$, sorts it in decreasing order to obtain $v_{i_1} \geq v_{i_2} \geq \ldots \geq v_{i_k}$ and returns the ordering $(i_1, \ldots, i_k)$.

Noisy Ranking Distributions Learning LSFs in the noiseless setting can be done efficiently by using linear programming. However, the common assumption both in theoretical and in applied works is that the observed rankings are noisy in the sense that they do not always correspond to the ground-truth ranking. We assume that the probability that the order of two elements i, j in the observed ranking π is different than their order in the ground-truth ranking σ^* is at most $\eta < 1/2$.

Definition 1.9 (Noisy Ranking Distribution)**.** *Fix $\eta \in [0, 1/2)$. An η-**noisy ranking distribution** $\mathcal{M}(\sigma^\star)$ with ground-truth ranking $\sigma^* \in \mathbb{S}_k$ is a probability measure over $\mathbb{S}_k$ that, for any $i, j \in [k]$, with $i \neq j$, satisfies $\mathbf{Pr}_{\pi \sim \mathcal{M}(\sigma^*)}[i \prec_\pi j \mid i \succ_{\sigma^*} j] \leq \eta$.* [1]

Note that, when $\eta = 0$, we always observe the ground-truth permutation and, in the case of $\eta = 1/2$, we may observe a uniformly random permutation. We remark that most natural ranking distributions satisfy this bounded noise property, e.g., (i) the Mallows model, which is probably the most fundamental ranking distribution (see, e.g., Braverman and Mossel (2009); Lu and Boutilier (2011); Caragiannis et al. (2013); Awasthi et al. (2014); Busa-Fekete et al. (2019); Fotakis et al. (2021c); De et al. (2018); Liu and Moitra (2018); Mao and Wu (2020); Liu and Moitra (2021) for a small sample of this line of research) and (ii) the Bradley-Terry-Mallows

[1] We use $i \succ_\pi j$ (resp. $i \prec_\pi j$) to denote that the element i is ranked higher (resp. lower) than j according to the ranking π.

model Mallows (1957), which corresponds to the ranking distribution analogue of the Bradley-Terry-Luce model Bradley and Terry (1952); Luce (2012) (the most studied pairwise comparisons model; see, e.g., Hunter (2004); Negahban et al. (2017); Agarwal et al. (2018) and the references therein). For more details, see Appendix D.5. Appendix D.5.

We consider the fundamental setting where the feature vector $\boldsymbol{x} \in \mathbb{R}^d$ is generated by a standard normal distribution and the ground-truth ranking for each sample $\boldsymbol{x}$ is given by the LSF $\sigma_{\boldsymbol{W}^*}(\boldsymbol{x})$ for some unknown parameter matrix $\boldsymbol{W}^* \in \mathbb{R}^{k \times d}$. For a fixed $\boldsymbol{x}$, the ranking that we observe comes from an η-noisy ranking distribution with ground-truth ranking $\sigma_{\boldsymbol{W}^*}(\boldsymbol{x})$.

Definition 1.10 (Noisy Linear Label Ranking Distribution)**.** *Fix $\eta \in [0, 1/2)$ and some ground-truth parameter matrix $\boldsymbol{W}^* \in \mathbb{R}^{k \times d}$. We assume that the **$\eta$-noisy linear label ranking distribution** $\mathcal{D}$ over $\mathbb{R}^d \times \mathbb{S}_k$ satisfies the following:*

1. *The $\boldsymbol{x}$-marginal of $\mathcal{D}$ is the d-dimensional standard normal distribution.*
2. *For any $(\boldsymbol{x}, \pi) \sim \mathcal{D}$, the distribution of π conditional on $\boldsymbol{x}$ is an η-noisy ranking distribution with ground-truth ranking $\sigma_{\boldsymbol{W}^*}(\boldsymbol{x})$.*

At first sight, the assumption that the underlying $\boldsymbol{x}$-marginal is the standard normal may look too strong. However, for $k = 2$, Definition 1.10 captures the problem of learning linear threshold functions with Massart noise. Without assumptions for the $\boldsymbol{x}$-marginal, it is known Chen et al. (2020b); Diakonikolas and Kane (2020); Nasser and Tiegel (2022) that optimal learning of halfspaces under Massart noise requires super-polynomial time (in the Statistical Query model of Kearns (1998)). On the other hand, a lot of recent works Balcan and Zhang (2017b); Mangoubi and Vishnoi (2019b); Diakonikolas et al. (2020e); Zhang et al. (2020b); Zhang and Li (2021) have obtained efficient algorithms for learning Massart halfspaces under Gaussian marginals. The goal of this work is to provide efficient algorithms for the more general problem of learning LSFs with bounded noise under Gaussian marginals.

Main Results and Techniques

Our main contributions are the first efficient algorithms for learning LSFs with bounded noise with respect to Kendall's Tau distance and top-r disagreement loss.

Learning in Kendall's Tau Distance. The most standard metric in rankings Shalev-Shwartz and Ben-David (2014b) is Kendall's Tau (KT) distance which, for two rankings $\pi, \tau \in \mathbb{S}_k$, measures the fraction of pairs (i, j) on which they disagree. That is, $\Delta_{\mathrm{KT}}(\pi, \tau) = \sum_{i \prec_\pi j} \mathbb{1}\{i \succ_\tau j\} / \binom{k}{2}$. Our first result is an efficient learning algorithm that, given samples from an η-noisy linear label ranking distribution $\mathcal{D}$, computes a parameter matrix $\boldsymbol{W}$ that ranks the alternatives almost optimally with respect to the KT distance from the ground-truth ranking $\sigma_{\boldsymbol{W}^*}(\cdot)$.

Theorem 1.11 (Informal – Learning LSFs in KT Distance)**.** *Fix $\eta \in [0, 1/2)$ and $\epsilon \in (0, 1]$. Let $\mathcal{D}$ be an η-noisy linear label ranking distribution satisfying the assumptions of Definition 1.10 with ground-truth LSF $\sigma_{\boldsymbol{W}^*}(\cdot)$. There exists an algorithm that draws $N = \mathrm{poly}(1/(1-2\eta))\, \widetilde{O}(d \log k/\epsilon)$ samples from $\mathcal{D}$, runs in sample-polynomial time, and computes a matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ such that, with high probability,*

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d} [\Delta_{\mathrm{KT}}(\sigma_{\boldsymbol{W}}(\boldsymbol{x}), \sigma_{\boldsymbol{W}^*}(\boldsymbol{x}))] \le \epsilon \,.$$

Theorem 5.1 gives the first efficient algorithm with provable guarantees for the supervised problem of learning noisy linear rankings. We remark that the sample complexity of our learning algorithm is qualitatively optimal (up to logarithmic factors) since, for $k = 2$, our problem subsumes learning a linear classifier with Massart noise [2] for which $\Omega(d/\epsilon)$ are known to be information theoretically necessary Massart and Nédélec (2006). Moreover, our learning algorithm is *proper* in the sense that it computes a linear sorting function $\sigma_{\boldsymbol{W}}(\cdot)$. As opposed to improper learners (see also Section 5.2), a proper learning algorithm gives us a compact representation (storing $\boldsymbol{W}$ requires $O(kd)$ memory) of the sorting function that allows

[2] Notice that in this case Kendall's Tau distance is simply the standard 0-1 binary loss.

us to efficiently compute (with runtime $O(kd + k \log k)$) the ranking corresponding to a fresh datapoint $\boldsymbol{x} \in \mathbb{R}^d$.

Learning in top-r Disagreement We next present our learning algorithm for the top-r metric formally defined as $\Delta_{\mathrm{top}-r}(\pi, \tau) = \mathbb{1}\{\pi_{1..r} \neq \tau_{1..r}\}$, where by $\pi_{1..r}$ we denote the ordering on the first r elements of the permutation π. The top-r metric is a disagreement metric in the sense that it takes binary values and for $r = 1$ captures the standard (multiclass) top-1 classification loss. We remark that, in contrast with the top-r classification loss, which only requires the predicted label to be in the top-r predictions of the model, the top-r ranking metric that we consider here requires that the model puts *the same elements in the same order* as the ground truth in the top-r positions. The top-r ranking is well-motivated as, for example, in ad targeting (discussed in Section 5.1) we want to be accurate on the top-r ad categories for a user so that we can diversify the content that they receive.

Theorem 1.12 (Informal – Learning LSFs in top-r Disagreement)**.** *Fix $\eta \in [0, 1/2)$, $r \in [k]$ and ϵ. Let $\mathcal{D}$ be an η-noisy linear label ranking distribution satisfying the assumptions of Definition 1.10 with ground-truth LSF $\sigma_{\boldsymbol{W}^\star}(\cdot)$. There exists an algorithm that draws $N = \mathrm{poly}(1/(1-2\eta))\, \widetilde{O}(dkr/\epsilon)$ samples from $\mathcal{D}$, runs in sample-polynomial time and computes a matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ such that, with high probability,*

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\Delta_{\mathrm{top}-r}(\sigma_{\boldsymbol{W}}(\boldsymbol{x}), \sigma_{\boldsymbol{W}^\star}(\boldsymbol{x}))] \leq \epsilon\,.$$

As a direct corollary of our result, we obtain a proper algorithm for learning the top-1 element with respect to the standard 0-1 loss that uses $\widetilde{O}(kd)$ samples. In fact, for small values of r, i.e., $r = O(1)$, our sample complexity is essentially tight. It is known that $\Theta(kd)$ samples are information theoretically necessary Natarajan (1989) for top-1 classification. [3] For the case $r = k$, i.e., when we want to learn

[3] Strictly speaking, those lower bounds do not directly apply in our setting because our labels are whole rankings instead of just the top classes but, in the Appendix D.4, we show that we can adapt the lower bound technique of Daniely et al. (2011) to obtain the same sample complexity lower bound for our ranking setting.

the whole ranking with respect to the 0-1 loss, our sample complexity is $O(k^2 d)$. However, using arguments similar to Daniely et al. (2011), one can show that in fact $O(dk)$ ranking samples are sufficient in order to learn the whole ranking with respect to the 0-1 loss. In this case, it is unclear whether a better sample complexity can be achieved with an efficient algorithm and we leave this as an interesting open question for future work.

1.4 Truncated Statistics with Unknown Truncation

A classical challenge in Statistics is estimation from truncated samples. Truncation occurs when samples falling outside of some subset S of the support of the distribution are not observed. Truncation of samples has myriad manifestations in business, economics, engineering, social sciences, and all areas of the physical sciences.

Statistical estimation under truncated samples has had a long history in Statistics, going back to at least the work of Galton Galton (1897) who analyzed truncated samples corresponding to speeds of American trotting horses. Following Galton's work, Pearson and Lee Pearson (1902); Pearson and Lee (1908); Lee (1914) used the method of moments in order to estimate the mean and standard deviation of a truncated univariate normal distribution and later Fisher Fisher (1931) used the maximum likelihood method for the same estimation problem. Since then, there has been a large volume of research devoted to estimating the truncated normal distribution; see e.g. Schneider (1986); Cohen (2016); Balakrishnan and Cramer (2014). Nevertheless, the first algorithm that is provably computationally and statistically efficient was only recently developed by Daskalakis et al. Daskalakis et al. (2018), under the assumption that the truncation set S is known.

In virtually all these works the question of estimation under unknown truncation set is raised. Our work resolves this question by providing tight sample complexity guarantees and an efficient algorithm for recovering the underlying Gaussian distribution. Although this estimation problem has clear and important practical and theoretical motivation too little was known prior to our work even in the asymptotic regime. In the early work of Shah and Jaiswal Shah and Jaiswal (1966) it was proven that the method of moments can be used to estimate a single dimensional Gaussian distribution when the truncation set is unknown but it is assumed to be an interval. In the other extreme where the set is allowed to be arbitrarily complex, Daskalakis et al. Daskalakis et al. (2018) showed that it is information theoretically impossible to recover the parameters. We provide the first complete analysis of the number of samples needed for recovery taking into account the complexity of the underlying set.

Main Results and Techniques

Our work studies the estimation task when the truncation set belongs in a family $\mathcal{C}$ of "low complexity". We use two different notions for quantifying the complexity of sets: the VC-dimension and the Gaussian Surface Area.

Our first result is that for any set family with VC-dimension $\mathrm{VC}(\mathcal{C})$, the mean and covariance of the true d-dimensional Gaussian Distribution can be recovered up to accuracy ϵ using only $\tilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{\epsilon} + \frac{d^2}{\epsilon^2}\right)$ truncated samples.

Theorem 1.13 (Informal – Identifiaility via VC-Dimension)**.** *Let $\mathcal{C}$ be a class of sets with VC-dimension $\mathrm{VC}(\mathcal{C})$ and let $N = \tilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{\epsilon} + \frac{d^2}{\epsilon^2}\right)$. Given N samples from a d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma)$ with unknown mean μ and covariance $\boldsymbol{r}\Sigma$, truncated on a set $S \in \mathcal{C}$ with mass at least α, it is possible to find an estimate $(\hat{\boldsymbol{\mu}}, \boldsymbol{r}\hat{\Sigma})$ such that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma), \mathcal{N}(\hat{\boldsymbol{\mu}}, \boldsymbol{r}\hat{\Sigma})) \leq \epsilon$.*

The estimation method computes the set of smallest mass that maximizes the likelihood of the data observed and learns the truncated distribution within error $O(\epsilon)$ in total variation distance. To translate this error in total variation to parameter distance, we prove a general result showing that it is impossible to create a set (no matter the complexity) so that two Gaussians whose parameters are far have similar truncated distributions (see Lemma 6.10).

A simple but not successful approach would be to first try to learn an approximation of the truncation set with symmetric difference roughly ϵ^2/d^2 with the true set and then run the algorithm of Daskalakis et al. (2018) using the approximate oracle. This approach would lead to a $\mathrm{VC}(\mathcal{S})d^2/\epsilon^2$ sample complexity that is worse than what we get. More importantly, doing empirical risk minimization[4] using truncated samples does not guarantee that we will find a set of small *symmetric* difference with the true and it is not clear how one could achieve that.

Our result bounds the sample complexity of identifying the underlying Gaussian distribution in terms of the VC-dimension of the set but does not yield a computationally efficient method for recovery. Obtaining a computationally efficient algorithm seems unlikely, unless one restricts attention to simple specific set families, such as axis aligned rectangles. One would hope that exploiting the fact that samples are drawn from a "tame" distribution, such as a Gaussian, can lead to general computationally efficient algorithms and even improved sample complexity.

Indeed, our main result is an algorithm that is both computationally and statistically efficient for estimating the parameters of a spherical Gaussian and uses only $d^{O(\Gamma^2(\mathcal{C}))}$ samples, where $\Gamma(\mathcal{C})$ is the *Gaussian Surface Area* of the class $\mathcal{C}$, an alternative complexity measure introduced by Klivans et al. Klivans et al. (2008):

Theorem 1.14 (Informal – Efficient Estimation of Truncated Gaussians)**.** *Let $\mathcal{C}$ be a class of sets with Gaussian surface area at most $\Gamma(\mathcal{C})$ and let $k = \mathrm{poly}(1/\alpha, 1/\epsilon)\Gamma(\mathcal{C})^2$. Given $N = d^k$ samples from a spherical d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{r} I)$, truncated on a set $S \in \mathcal{C}$ with mass at least α,in time* poly(m)*, we can find an estimate $\hat{\mu}, \hat{\sigma}^2$ such that*

$$d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{r} I), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2 \boldsymbol{r} I)) \leq \epsilon.$$

The notion of Gaussian surface area can lead to better sample complexity bounds even when the VC dimension is infinite. An example of such a case is when $\mathcal{C}$ is the class of all convex sets. Table 1.1 summarizes the known bounds for the Gaussian surface area of different concept classes and the implied sample complexity in our setting when combined with our main theorem.

[4] That is finding a set of the family that contains all the observed samples.

Concept Class	**Gaussian Surface Area**
Polynomial threshold functions of degree k	$O(k)$ Kane (2011)
Intersections of k halfspaces	$O(\sqrt{\log k})$ Klivans et al. (2008)
General convex sets	$O(d^{1/4})$ Ball (1993)

Table 1.1: Summary of known results for Gaussian Surface Area. The last column gives the sample complexity we obtain for our setting.

Beyond spherical Gaussians, our main result extends to Gaussians with arbitrary diagonal covariance matrices. In addition, we provide an information theoretic result showing that the case with general covariance matrices can also be estimated using the same sample complexity bound by finding a Gaussian and a set that matches the moments of the true distribution. We remark our main algorithmic result Theorem 1.15 uses Gaussian Surface Area whereas our sample complexity result Theorem 1.14 uses VC-dimension. We discuss the differences of the two approaches in Section 6.6.

Theorem 1.15 (Informal – Identifiability via GSA)**.** *Let $\mathcal{C}$ be a class of sets with Gaussian surface area at most $\Gamma(\mathcal{C})$ and let $k = \mathrm{poly}(1/\alpha, 1/\epsilon)\Gamma(\mathcal{C})^2$. Any truncated Gaussian with $\mathcal{N}(\hat{\boldsymbol{\mu}}, \boldsymbol{r}\hat{\Sigma}, \hat{S})$ with $\hat{S} \in \mathcal{C}$ that approximately matches the moments up to degree k of a truncated d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma, S)$ with $S \in \mathcal{C}$, satisfies $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma), \mathcal{N}(\hat{\boldsymbol{\mu}}, \boldsymbol{r}\hat{\Sigma})) \leq \epsilon$. The number of samples to estimate the moments within the required accuracy is at most $d^{O(k)}$.*

This shows that the first few moments are sufficient to identify the parameters. Analyzing the guarantees of moment matching methods is notoriously challenging as it involves bounding the error of a system of many polynomial equations. Even for a single-dimensional Gaussian with truncation in an interval, where closed form solutions of the moments exist, it is highly non-trivial to bound these errors Shah and Jaiswal (1966). In contrast, our analysis using Hermite polynomials allows us to easily obtain bounds for arbitrary truncation sets in high dimensions, even though no closed form expression for the moments exists.

We conclude by showing that the dependence of our sample complexity bounds both on the VC-dimension and the Gaussian Surface Area is tight *up to polynomial factors*. In particular, we construct a family in d dimensions with VC dimension 2^d and Gaussian surface area $O(d)$ for which it is not possible to learn the mean of the underlying Gaussian within 1 standard deviation using $o(2^{d/2})$ samples.

Theorem 1.16 (Informal)**.** *There exists a family of sets $\mathcal{S}$ with $\Gamma(\mathcal{S}) = O(d)$ and VC-dimension 2^d such that any algorithm that draws N samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}I, S)$ and computes an estimate $\widetilde{\boldsymbol{\mu}}$ with $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq 1$ must have $N = \Omega(2^{d/2})$.*

Our techniques and relation to prior work The work of Klivans et al. Klivans et al. (2008) provides a computationally and sample efficient algorithm for learning geometric concepts from labeled examples drawn from a Gaussian distribution. On the other hand, the recent work of Daskalakis et al. Daskalakis et al. (2018) provides efficient estimators for truncated statistics with *known* sets. One could hope to combine these two approaches for our setting, by first learning the set and then using the algorithm of Daskalakis et al. (2018) to learn the parameters of the Gaussian. This approach, however, fails for two reasons. First, the results of Klivans et al. Klivans et al. (2008) apply in the supervised learning setting where one has access to both positive and negative samples, while our problem can be thought of as observing only positive examples (those falling inside the set). In addition, any direct approach that extends their result to work with positive only examples requires that the underlying Gaussian distribution is known in advance.

One of our key technical contributions is to extend the techniques of Klivans et al. Klivans et al. (2008) to work with *positive only examples* from an *unknown* Gaussian distribution, which is the major case of interest in truncated statistics. To perform the set estimation Klivans et al. Klivans et al. (2008), rely on a family of orthogonal polynomials with respect to the Gaussian distribution, namely the Hermite polynomials and show that the indicator function of the set is well approximated by its low degree Hermite expansion. While we cannot learn this function directly in our setting, we are able to recover an alternative function, that contains “entangled” information of both the true Gaussian parameters and the underlying set. After learning the function, we formulate an optimization problem whose solution enables us to decouple these two quantities and retrieve both the Gaussian parameters and the underlying set. We describe our estimation method in more detail in Section 6.3. As a corollary of our approach, we obtain the first efficient algorithm for learning geometric concepts from positive examples drawn from an unknown spherical Gaussian distribution.

1.5 Non-Parametric Truncated Statistics: Learning Smooth-Densities

Non-parametric density estimation is a well-developed field in Statistics and Machine Learning Wasserman (2006); Tsybakov (2008) with applications to many scientific ares including economics Ahamada and Flachaire (2010); Li and Racine (2007), and survival analysis Woodroofe et al. (1985). A central challenge in this field is estimating a probability density function $D(\boldsymbol{x})$ from samples, without making strong parametric assumptions about the density. Of course, this is quite challenging as D may exhibit very rich behavior which might be difficult or information theoretically impossible to discern given a finite number of samples. Thus, to make the task feasible at all, some constraints are placed on D, typically in the form of smoothness, which allows estimators to *interpolate* among the observed samples. Indeed, a prominent method for non-parametric density estimation is based on kernels Wand and Jones (1994); Botev et al. (2010); Simonoff (2012); Scott (2015), whose usual interpolating estimate takes the form $\hat{D}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} k(\boldsymbol{x}_i; \boldsymbol{x})$, for some kernel function $k(\cdot;\cdot)$, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are the observations from D. In some settings is also preferable to use kernels to estimate the log-density function Canu and Smola (2006). Even with smoothness assumptions, the problem is challenging enough information theoretically, that the achievable error takes the form $n^{-O(r/(r+d))}$, under various norms, where d is the dimension and r is the assumed order of smoothness of D McDonald (2017); Li and Racine (2007). Similar results can be obtained using histograms density estimation Barron et al. (1992).

Despite the fact that both kernel based and histogram based estimators achieve the information theoretic optimal consistency rates, the resulting estimator does not have a form that is appealing for other statistical uses after estimation. For example, if the estimation is then used to solve some hypothesis testing problems then it would be helpful if the estimated distribution is represented as a member of an exponential family Neyman (1937); Good (1963). A parallel line of research has hence devoted in *exponential series estimators* of non-parametric densities, starting with the celebrated work of Barron and Sheu Barron and Sheu (1991) which was later extended to multidimensional settings as well Wu (2010). Our work follows this line of research and the estimators that we compute are always members of some exponential family distributions.

Our goal is to extend this literature from the traditional *interpolating* regime to the much more challenging *extrapolating* regime. In particular, we consider settings wherein we are constrained to observe samples of D in a *subset* of its support, yet we want to procure estimates $\hat{D}$ that approximate D over its *entire* support. This question problem is motivated by truncated statistics, another well-developed field in Statistics and Econometrics Cohen (1991); Heckman (1976); Maddala (1987); Börsch-Supan and Hajivassiliou (1993), which targets statistical estimation in settings where the samples are truncated depending on their membership in some set. Truncation may occur for several reasons, ranging from measurement device saturation effects to data collection practices, bad experimental design, ethical or privacy considerations that disallow the use of some data, etc.

Non-parametric density estimation from truncated samples is well-studied problem in statistics with many applications in economics and survival analysis Padgett and McNichols (1984); Woodroofe et al. (1985); Lai and Ying (1991); Stute et al. (1993); Gajek et al. (1988); Lai and Ying (1991). However, due to the very challenging nature of this problem, all the previous works on this topic consider only a soft truncation model that does not completely hide some part of the support but only decreases the probability of observing something that lies in the truncation set. In particular, each sample $\boldsymbol{x}_i$ from D also samples a truncation set S_i which then determines whether this sample is truncated or not. As a result, samples from the entire support are ultimately collected, thus the unknown density can be interpolated, with some appropriate re-weighting, from those samples covering the entire support. Additionally the existing work only targets single-dimensional densities despite the importance of non-parametric estimation in multiple dimensions as we discussed above.

In this work we target in solving the seemingly impossible problem of estimating a non-parametric density, even in parts of the support that we don't observe any sample! More precisely, we consider the more standard, in truncated statistics, hard truncation model, wherein there is a fixed set S that determines whether each and every sample from D is truncated. We solve this problem under slightly stronger but similar assumptions to the ones used in the vanilla non-parametric density estimation problems. At the same time, we extend the non-parametric density estimation from truncated samples to the multi-dimensional settings, which is a significant generalization of the existing work.

Our main theorems, summarized below, can be interpreted as a statistical version of Taylor's theorem, which allows us to use *truncated samples* from some sufficiently smooth density D and *extrapolate* from these samples an estimate $\hat{D}$ which approximates D on its *entire support*. The statistical rates achieved by our theorems are slightly worse but comparable to those known in non-parametric density estimation under *untruncated* samples, i.e., in the *interpolating* regime. It is an interesting open problem whether we can improve the novel *extrapolation* rates that we provide in this work, to match exactly the *interpolation* rates of the vanilla non-parametric density estimation.

From a technical point of view, a central challenge that we face is to bound the extrapolation error of multivariate polynomial approximation, which is a challenging problem that is a subject of active area of research. Our main technical contribution is to show a novel way to prove strong bounds on the extrapolation error of our algorithms invoking only well-studied *anti-concentration* theorems, which is of independent interest and we believe that it will have applications beyond truncated statistics. More precisely, one of our main technical results is a "Distortion of Conditioning" lemma (Lemma 7.18), providing a tight relationship between the ℓ_1 distance between two exponential families as computed under conditioning on different subsets of the support. As we said, this lemma is proven using probabilistic techniques, and provides a viable route to prove our statistical Taylor result in high dimensions, where polynomial approximation theory techniques do not appear sufficient.

Main Results and Techniques

As we already mentioned, in this work we provide provable extrapolation of non-parametric density functions from samples, i.e., given samples from the conditional density on some subset S of the support we want to recover the shape of the density function *outside* of S. We consider densities proportional to $e^{f(x)}$, where f is a sufficiently smooth function. Our observation consists of samples from a density proportional to $\mathbb{1}_S(x)e^{f(x)}$, where S is a known (via a membership oracle) subset of the support. For this problem to even be well-posed we need further assumptions on the density function. Even if we are given the exact conditional density $\mathbb{1}_S(x)e^{f(x)}$, it is easy to see that, if $f \notin C_\infty$, i.e., if f is not infinitely times differentiable everywhere in the whole support, there is no hope to extrapolate its curve outside of S; for a simple example, if we observe a density proportional to $e^{|x|}$ truncated in $(-\infty, 0]$ we cannot extrapolate this density to $(0, +\infty)$, because we cannot distinguish whether we are observing truncated samples from e^{-x} or $e^{|x|}$. On the other hand, if the log-density f is analytic and sufficiently smooth, then the value of f at every x can be determined only from local information, namely its derivatives at a single point. This well known property of analytic functions is quantified by Taylor's remainder theorem. In this work we build upon this intuition and show that even given *samples* from a sufficiently smooth density and even if these samples are *conditioned in a small subset of the support* we can still determine the function in the entire support and most importantly this can be done in a statistically and computationally efficient way.

In the light of the above observation, we restrict our attention to functions f that satisfy specific smoothness assumptions. In particular, we assume that the k-th order derivatives of f do not increase faster than exponentially in k, i.e., $|f^{(k)}(\boldsymbol{x})| \leq M^k$ for some $M \in \mathbb{R}_+$ and all $\boldsymbol{x}$ in the support (see Definition 7.4). Notice that similar assumptions are standard in non-parametric density estimation problems, even when no extrapolation is needed, see, for example, Barron and Sheu (1991); Wu (2010).

We start our exposition with the single-dimensional version of our extrapolation problem in Section 7.2. We make this choice for several reasons: (1) it is easier to compare with the existing line of work on non-parametric density estimation both in the vanilla non-truncated and in the truncated setting, (2) in the single-dimensional setting we are able to show a slightly stronger information theoretic result, and (3) the single dimensional setting serves as a nice example where the difference between interpolation and extrapolation. In this single dimensional setting we assume that there exists some unknown log-density function f, a known set S, and we observe samples from the distribution $D(f, S)$, which has density proportional to $\mathbb{1}_S e^{f(x)}$. Our goal is to estimate the whole distribution $D(f)$ which for simplicity we assume that f is supported on $[0, 1]$ and hence $S \subseteq [0, 1]$. Our first step is to consider the *semi-parametric* class of densities p that consists of polynomial series that can approximate the unknown non-parametric log-density f. Then we truncate this polynomial series and we only consider densities of the form $e^{p(x)}$, where p is a degree k polynomial, with large enough k; observe that these densities belong to an exponential family.

Our first result shows that the polynomial which maximizes the likelihood with respect to the *conditional* distribution $D(f, S)$ (let us call this polynomial the "MLE polynomial") approximates the density $e^{f(x)}$ *everywhere* on $[0, 1]$, i.e. the MLE polynomial has small extrapolation error. Observe, that this result cannot follow just from the fact that for example the Taylor polynomial extrapolates, because the MLE polynomial and the Taylor polynomial are in principle very different. While it is conceptually clear that the MLE polynomial of sufficiently large degree will have small interpolation error and hence will approximate well the density inside S, our result is the first to show that the same polynomial has small extrapolation error and hence approximates the density on the entire interval $[0, 1]$. For the formal version, we refer to Theorem 7.7

Theorem 1.17 (Informal – MLE Extrapolation Error)**.** *Let $D(f, [0,1])$ be a probability distribution with sufficiently smooth log-density f and let $D(f, S)$ be its conditional distribution on $S \subset [0,1]$. The MLE w.r.t $D(f, S)$ polynomial p^* of degree $O(\log(1/\epsilon))$ satisfies $d_{\mathrm{TV}}(D(f, [0,1]), D(p^*, [0,1])) \leq \epsilon$.*

Extending the previous result to multivariate densities is significantly more challenging. The reason is that multivariate polynomial interpolation is much more intricate and is a subject of active research, see for example the survey Gasca and Sauer (2000). Instead of trying to characterize the properties of the exact MLE polynomial we give an alternative method for obtaining multivariate extrapolation guarantees that does not rely on multivariate polynomial interpolation. Our approach uses the additional assumption that the set S from which we observe samples has non-trivial volume, that is $\mathrm{vol}(S) \geq a$ for some $\alpha > 0$. Under this natural assumption we obtain the following theorem (for the formal version see Therorem 7.8).

Theorem 1.18 (Informal – Multivariate MLE Extrapolation Error)**.** *Let $D(f, [0,1]^d)$ be a probability distribution with sufficiently smooth log-density f and let $D(f, S)$ be its conditional distribution on $S \subset [0,1]^d$ with $\mathrm{vol}(S) \geq \alpha$. The MLE w.r.t $D(f, S)$ polynomial p^* of degree $O(d^3/\alpha^2 + \log(1/\epsilon))$ satisfies $d_{\mathrm{TV}}(D(f, [0,1]^d), D(p^*, [0,1]^d)) \leq \epsilon$.*

Our approach for proving Theorem 1.18 is more general and of independent interest. In particular, we use a structural result that quantifies the distortion of the metric space of exponential families under conditioning. Given a polynomial p with corresponding density $D(p, [0,1]^d)$ we consider the conditioning map that maps $D(p, [0,1]^d)$ to the distribution $D(p, S)$. We show this conditioning map distorts the total variation distance metric by a factor of order at most $(d/\alpha)^{O(k)}$. In other words, distributions that are close in the image space of the conditioning map are also close in the domain space and vice versa (for the formal version, see Lemma 7.18)

Lemma 1.19 (Informal – Distortion of Conditioning). *Let p, q be polynomials of degree at most k. For every $S \subseteq [0,1]^d$ with $\mathrm{vol}(S) \geq \alpha$ it holds*

$$(d/\alpha)^{-O(k)} \leq \frac{d_{\mathrm{TV}}(\, D(p,[0,1]^d),\ D(q,[0,1]^d)\,)}{d_{\mathrm{TV}}(D(p,S),\ D(q,S))} \leq (d/\alpha)^{O(k)}\,.$$

Using the above theorem our strategy for showing Theorem 1.18 is illustrated in Figure 1.1 and is as follows. Our first step is to use Taylor's remainder theorem to prove that there exists a polynomial p, associated with f, such that both $d_{\mathrm{TV}}(D(p,S), D(f,S))$ and $d_{\mathrm{TV}}(D(p,[0,1]^d), D(f,[0,1]^d))$ are both very small when p has sufficiently large degree. Next, we show that optimizing the likelihood function on S over the space of degree k polynomials we obtain the MLE polynomial q which achieves very small total variation distance to f on S, i.e. $d_{\mathrm{TV}}(D(q,S), D(f,S))$ is also small. Hence, from the triangle inequality we have that $d_{\mathrm{TV}}(D(q,S), D(p,S))$ is also very small. The next step, which is the crucial one, is that we can now apply our novel Theorem 1.19 to obtain that $d_{\mathrm{TV}}(D(q,[0,1]^d), D(p,[0,1]^d))$ blows up at most by a factor of $(d/\alpha)^{O(k)}$. This argument leads to an upper bound on the extrapolation error (y in Figure 1.1). The last key observation is that the quantity $d_{\mathrm{TV}}(D(p,S), D(f,S))$ decreases faster than $(d/\alpha)^{-O(k)}$ as the degree k increases and hence we can make the extrapolation error arbitrarily small by choosing sufficiently high degree.

So far we have argued about the extrapolation error of the population MLE polynomial, i.e., we assume that we have access to the population distribution $D(f,S)$ and that we can maximize the population MLE with no error. Therefore, our next step is to show how we can incorporate the statistical error from the access to only finitely many samples from $D(f,S)$ and to provide an efficient algorithm that computes the MLE polynomial with small enough approximation loss (for the formal version see Theorem 7.9).

Theorem 1.20 (Informal – Extrapolation Algorithm). *Let $D(f,[0,1]^d)$ be sufficiently*

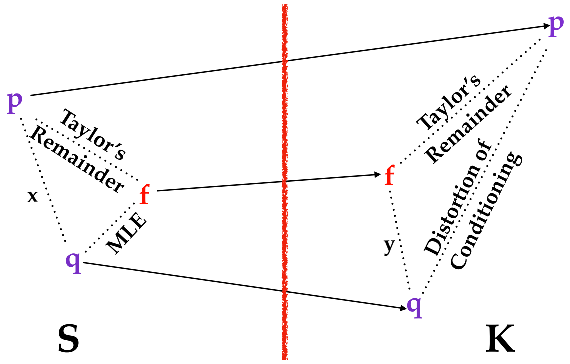


Figure 1.1: Using Theorem 1.19 to show the extrapolation guarantees of MLE. $K = [0,1]^d$. p is the Taylor Polynomial of f: from Taylor's remainder theorem we know that, in both S and K, p is very close to f. q is the MLE polynomial on S: it is very close to f in S. The distance x is bounded by triangle inequality. The distance of p and q in K is upper bounded by x $(d/\alpha)^{O(k)}$ by Theorem 1.19. Finally, y is the extrapolation error of the MLE polynomial q on K and is bounded by another triangle inequality. Overall, $y \le d_{\mathrm{TV}}(D(f,K), D(p,K)) + (d/\alpha)^{O(k)}x \le d_{\mathrm{TV}}(D(f,K), D(p,K)) + (d/\alpha)^{O(k)}(d_{\mathrm{TV}}(D(f,S), D(p,S)) + d_{\mathrm{TV}}(D(f,S), D(q,S)))$.

smooth. Let $S \subseteq [0,1]^d$ be such that $\mathrm{vol}(S) \ge \alpha$*. There exists an algorithm that draws*

$$N = 2^{\widetilde{O}(d^4/\alpha^2)} \cdot (1/\epsilon)^{O(d+\log(1/\alpha))}$$

samples from $D(f,S)$, runs in time polynomial in the number of samples, and with probability at least 99% *outputs a polynomial q of degree $\widetilde{O}(d^3/\alpha^2) + O(\log(1/\epsilon))$ such that $d_{\mathrm{TV}}(D(f,K), D(q,K)) \le \epsilon$.*

It is well known that non-parametric density estimation (in the interpolation regime, i.e. from untruncated samples) under smoothness assumptions requires samples that depend exponentially in the dimension, i.e. the typical rate is $(1/\epsilon)^{\Theta(d)}$, see for example Tsybakov (2008); McDonald (2017); Li and Racine (2007). The usual assumption is that the density has bounded derivatives, i.e. it belongs to a Sobolev or a Besov space. Our problem of extrapolating the density function is a strict generalization of non-parametric density estimation and therefore our sample complexity naturally scales as $(1/\epsilon)^{O(d+\log(1/\alpha))}$, where the $\log(1/\alpha)$ reflects the impact of conditioning on a small volume set S. Our estimation algorithm suffers from an additional $2^{\widetilde{O}(d^4/\alpha^2)}$ which does not depend on the accuracy parameter ϵ. For sets of constant volume, in constant dimensions, we obtain a almost the same asymptotic sample complexity with the interpolation setting and in particular depends polynomially in the accuracy parameter ϵ. For higher dimensions it is an interesting open problem whether this additional factor is necessary or not.

For many learning problems one may not have access to fine grained label information; e.g., an image can be labeled as husky, dog, or even animal depending on the expertise of the annotator. In this work, we formalize these settings and study the problem of learning from such coarse data. Instead of observing the actual labels from a set $\mathcal{Z}$, we observe coarse labels corresponding to a partition of $\mathcal{Z}$ (or a mixture of partitions).

Our main algorithmic result is that essentially any problem learnable from fine grained labels can also be learned efficiently when the coarse data are sufficiently informative. We obtain our result through a generic reduction for answering Statistical Queries (SQ) over fine grained labels given only coarse labels. The number of coarse labels required depends polynomially on the information distortion due to coarsening and the number of fine labels $|\mathcal{Z}|$.

We also investigate the case of (infinitely many) real valued labels focusing on a central problem in censored and truncated statistics: Gaussian mean estimation from coarse data. We provide an efficient algorithm when the sets in the partition are convex and establish that the problem is NP-hard even for very simple non-convex sets.

1.6 Learning from Coarse Data

Supervised learning from labeled examples is a classical problem in machine learning and statistics: given labeled examples, the goal is to train some model to achieve low classification error. In most modern applications, where we train complicated models such as neural nets, large amounts of labeled examples are required. Large datasets such as Imagenet, Russakovsky et al. (2015), often contain thousands of different categories such as animals, vehicles, etc., each one of those containing many *fine grained* subcategories: animals may contain dogs and cats and dogs may be further split into different breeds etc. In the last few years, there have been many works that focus on fine grained recognition, Guo et al. (2018); Chen et al. (2018); Touvron et al. (2020); Qin et al. (2020); Lei et al. (2017); Jiao et al. (2019, 2020); Bukchin et al. (2020); Taherkhani et al. (2019). Collecting a sufficient amount of accurately labeled training examples is a hard and expensive task that often requires hiring experts to annotate the examples. This has motivated the problem of learning from *coarsely* labeled datasets, where a dataset is not fully annotated with fine grained labels but a combination of fine, e.g., cat, and coarse labels, e.g., animal, is given, Deng et al. (2013); Ristin et al. (2015).

Inference from coarse data naturally arises also in unsupervised, i.e., distribution learning settings: instead of directly observing samples from the target distribution, we observe “representative" points that correspond to larger sets of samples. For example, instead of observing samples from a real valued random variable, we round them to the closest integer. An important unsupervised problem that fits in the coarse data framework is censored statistics, Cohen (2016); Wolynetz (1979); Breen et al. (1996); Schneider (1986). Interval censoring, that arises in insurance adjustment applications, corresponds to observing points in some interval and point masses at the endpoints of the interval instead of observing fine grained data from the whole real line. Moreover, the problem of learning the distribution of the output of neural networks with non-smooth activations (e.g., ReLU networks, Wu et al. (2019)) also fits in our model of distribution learning with coarse data, see Figure 8.2(c).

Even though the problem of learning from coarsely labeled data has attracted significant attention from the applied community, from a theoretical perspective little is known. In this work, we provide efficient algorithms that work in both the supervised and the unsupervised coarse data settings.

Our Model and Results

We start by describing the generative model of coarsely labeled data in the supervised setting. We model coarse labels as subsets of the domain of all possible fine labels. For example, assume that we hire an expert on dog breeds and an expert on cat breeds to annotate a dataset containing images of dogs and cats. With probability 1/2, we get samples labeled by the dog expert, i.e., labeled according to the partition

$$\{\text{cat} = \{\text{persian cat}, \text{bengal cat}, \ldots\}, \{\text{maltese dog}\}, \{\text{husky dog}\}, \ldots\}\,.$$

On the other hand, the cat expert will provide a fine grained partition over cat breeds and will group together all dog breeds. Our coarse data model captures exactly this mixture of different label partitions.

Definition 1.21 (Generative Process of Coarse Data with Context)**.** *Let $\mathcal{X}$ be an arbitrary domain, and let $\mathcal{Z} = \{1, \ldots, k\}$ be the discrete domain of all possible fine labels. We generate coarsely labeled examples as follows:*

1. *Draw a finely labeled example (x, z) from a distribution D on $\mathcal{X} \times \mathcal{Z}$.*
2. *Draw a coarsening partition $\mathcal{S}$ (of $\mathcal{Z}$) from a distribution π.*
3. *Find the unique set $S \in \mathcal{S}$ that contains the fine label z.*
4. *Observe the coarsely labeled example (x, S).*

We denote D_π the distribution of the coarsely labeled example (x, S).

In the supervised setting, our main focus is to answer the following question.

Question 1.22. *Can we train a model, using coarsely labeled examples $(x, S) \sim D_\pi$, that classifies finely labeled examples $(x, z) \sim D$ with accuracy comparable to that of a classifier that was trained on examples with fine grained labels?*

Definition 8.1 does not impose any restrictions on the distribution over partitions π. It is clear that if partitions are very rough, e.g., we split $\mathcal{Z}$ into two large disjoint subsets, we lose information about the fine labels and we cannot hope to train a classifier that performs well over finely labeled examples. In order for Question 8.2 to be information theoretically possible, we need to assume that the partition distribution π preserves fine-label information. The following definition quantifies this by stating that reasonable partition distributions π are those that preserve the total variation distance between different distributions supported on the domain of the fine labels $\mathcal{Z}$. We remark that the following definition does not require $\mathcal{D}$ to be supported on pairs (x, z) but is a general statement for the unsupervised version of the problem, see also Definition 8.10.

Definition 1.23 (Information Preserving Partition Distribution)**.** *Let $\mathcal{Z}$ be any domain and let $\alpha \in (0, 1]$. We say that π is an α-information preserving partition distribution if for every two distributions D^1, D^2 supported on $\mathcal{Z}$, it holds that $d_{\mathrm{TV}}(D^1_\pi, D^2_\pi) \geq \alpha \cdot d_{\mathrm{TV}}(D^1, D^2)$, where $d_{\mathrm{TV}}(D^1, D^2)$ is the total variation distance of D^1 and D^2.*

For example, the partition distribution defined in the dog/cat dataset scenario, discussed before Definition 8.1, is 1/2-information preserving, since we observe fine labels with probability 1/2. In this case, it is easy, at the expense of losing the statistical power of the coarse labels, to combine the finely labeled examples from both experts in order to obtain a dataset consisting only of fine labels. However, our model allows the partitions to have arbitrarily complex combinatorial structure that makes the process of “inverting" the partition transformation computationally challenging. For example, specific fine labels may be complicated functions of coarse labels: “medium sized" and “pointy ears" and “blue eyes" may be mapped to the “husky dog" fine label.

Our first result is a positive answer to Question 8.2 in essentially full generality: we show that concept classes that are efficiently learnable in the Statistical Query (SQ) model, Kearns (1998), are also learnable from coarsely labeled examples. Our result is similar in spirit with the result of Kearns (1998), where it is proved that SQ learnability implies learnability under random classification noise.

Theorem 1.24 (Informal – SQ Learnability implies Learnability from Coarse Examples)**.** *Any concept class $\mathcal{C}$ that is efficiently learnable with M statistical queries from finely labeled examples $(x, z) \sim D$, can be efficiently learned from $O(\mathrm{poly}(k/\alpha)) \cdot M$ coarsely labeled examples $(x, S) \sim D_\pi$ under any α-information preserving partition distribution π.*

Statistical Queries are queries of the form $\mathbf{E}_{(x,z)\sim D}[q(x, z)]$ for some query function $q(x, z)$. It is known that almost all known machine learning algorithms Aslam and Decatur (1998); Blum et al. (1998, 2005); Dunagan and Vempala (2008); Balcan and Feldman (2015); Feldman et al. (2017) can be implemented in the SQ model. In particular, in Feldman et al. (2015a), it is shown that (Stochastic) Gradient Descent can be simulated by statistical queries. This implies that our result can be applied, even in cases where it is not possible to obtain formal optimality guarantees, e.g., training deep neural nets. We can train such models using coarsely labeled data and guarantee the same performance as if we had direct access to fine labels (see also Appendix G.1). [5] As another application, we consider the problem of multiclass logistic regression with coarse labels. It is known, see e.g., Friedman et al. (2001), that given finely labeled examples $(x, z) \sim D$, the likelihood objective for multiclass logistic regression is concave with respect to the weight matrix. Even though the likelihood objective is no-longer concave when we consider coarsely labeled examples $(x, S) \sim D_\pi$, our theorem bypasses this difficulty and allows us to efficiently perform multiclass logistic regression with coarse labels.

Formally, we design an algorithm (Algorithm 12) that, given coarsely labeled examples (x, S), efficiently simulates statistical queries over finely labeled examples (x, z). Surprisingly, the runtime and sample complexity of our algorithm do not depend on the combinatorial structure of the partitions, but only on the

[5] Given any objective of the form $L(\boldsymbol{v}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(\boldsymbol{v}; \boldsymbol{x}, y)]$, its gradients correspond to $\nabla_{\boldsymbol{v}} L(\boldsymbol{v}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\nabla_{\boldsymbol{v}} \ell(\boldsymbol{v}; \boldsymbol{x}, y)]$. Having Statistical Query access to the distribution of (x, y), we can directly obtain estimates of the above gradients using the query functions $q_i(\boldsymbol{x}, y) = (\nabla_{\boldsymbol{v}} \ell(\boldsymbol{v}; \boldsymbol{x}, y))_i$. In Feldman et al. (2015a), the precise accuracy required for specific SQ implementations of first order methods depends on the complexity of the underlying distribution and the particular objective function $\ell(\cdot)$.

number of fine labels k and the information preserving parameter α of the partition distribution π.

Theorem 1.25 (Informal – SQ from Coarsely Labeled Examples)**.** *Consider a distribution D_π over coarsely labeled examples in $\mathbb{R}^d \times [k]$, (see Definition 8.1) with α-information preserving partition distribution π. Let $q : \mathbb{R}^d \times [k] \to [-1, 1]$ be a query function, that can be evaluated on any input in time T, and $\tau \in (0, 1)$. There exists an algorithm, that draws $N = \mathrm{poly}(k/(\tau\alpha))$ coarsely labeled examples from D_π and, in $\mathrm{poly}(N, T)$ time, computes an estimate $\hat{r}$ such that, with high probability, it holds $\left| \mathbf{E}_{(x,z)\sim D}[q(x, z)] - \hat{r} \right| \le \tau$.*

Learning Parametric Distributions from Coarse Samples. In many important applications, instead of a discrete distribution over fine labels, a continuous parametric model is used. A popular example is when the domain $\mathcal{Z}$ of Definition 8.1 is the entire Euclidean space $\mathbb{R}^d$, and the distribution of finely labeled examples is a Gaussian distribution whose parameters possibly depend on the context x. Such censored regression settings are known as Tobit models Tobin (1958); Maddala (1986); Gourieroux (2000). Lately, significant progress has been made from a computational point of view in such censored/truncated settings in the distribution specific setting, e.g., when the underlying distribution is Gaussian Daskalakis et al. (2018); Kontonis et al. (2019), mixtures of Gaussians Nagarajan and Panageas (2019), linear regression Daskalakis et al. (2019); Ilyas et al. (2020); Daskalakis et al. (2020). In this distribution specific setting, we consider the most fundamental problem of learning the mean of a Gaussian distribution given coarse data.

Definition 1.26 (Coarse Gaussian Data)**.** *Consider the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^\star)$, with mean $\boldsymbol{\mu}^\star \in \mathbb{R}^d$ and identity covariance matrix. We generate a sample as follows:*

1. *Draw $\boldsymbol{z}$ from $\mathcal{N}(\boldsymbol{\mu}^\star)$.*

2. *Draw a partition $\boldsymbol{\Sigma}$ (of $\mathbb{R}^d$) from π.*

3. *Observe the set $S \in \boldsymbol{\Sigma}$ that contains $\boldsymbol{z}$.*

We denote the distribution of S as $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$.

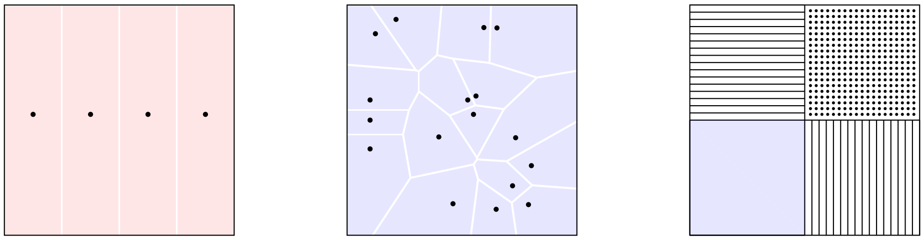
Remark 1.27. *We remark that we only require membership oracle access to the subsets of the partition $\mathcal{S}$. A set $S \subseteq \mathbb{R}^d$ corresponds to a membership oracle $\mathcal{O}_S : \mathbb{R}^d \to \{0, 1\}$ that given $\boldsymbol{x} \in \mathbb{R}^d$ outputs whether the point lies inside the set S or not.*

We first study the above problem, from a computational viewpoint. For the corresponding problems in censored and truncated statistics no geometric assumptions are required for the sets: in Daskalakis et al. (2018) it was shown that an efficient algorithm exists for arbitrarily complex truncation sets. In contrast in our more general model of coarse data we show that having sets with geometric structure is necessary. In particular we require that every set of the partition is convex, see Figure 8.2(b,c). We show that when the convexity assumption is dropped, learning from coarse data is a computationally hard problem even under a mixture of very simple sets.

Theorem 1.28 (Informal – Hardness of Matching the Observed Distribution with General Partitions)**.** *Let π be a general partition distribution. Unless RP = NP, no algorithm with sample access to $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$, can compute, in* $\mathrm{poly}(d)$ *time, a $\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^d$ such that $d_{\mathrm{TV}}(\mathcal{N}_\pi(\widetilde{\boldsymbol{\mu}}), \mathcal{N}_\pi(\boldsymbol{\mu}^\star)) < 1/d^c$ for some absolute constant $c > 1$.*

We prove our hardness result using a reduction from the well known MAX-CUT problem, which is known to be NP-hard, even to approximate Håstad (2001). In our reduction, we use partitions that consist of simple sets: fat hyperplanes, ellipsoids and their complements: the computational hardness of this problem is rather inherent and not due to overly complicated sets.

On the positive side, we identify a geometric property that enables us to design a computationally efficient algorithm for this problem: namely we require all the sets of the partitions to be *convex*, e.g., Figure 8.2(b,c). We remark that having finite or countable subsets, is not a requirement of our model. For example, we can handle convex partitions of the form (c) that correspond to the output distribution of a ReLU neural network, see Wu et al. (2019). We continue with our theorem for learning Gaussians from coarse data.



(a) Non-Identifiable Case (b) Convex Partition Case (c) ReLU Case

Figure 1.2: Convex Partitions of $\mathbb{R}^2$.

Figure 1.3: (a) is a very rough partition, that makes learning the mean impossible: Gaussians $\mathcal{N}((0,z))$ centered along the same vertical line $(0,z)$ assign exactly the same probability to all cells of the partitions and therefore, $d_{TV}(\mathcal{N}_\pi((0,z_1)),\mathcal{N}_\pi((0,z_2)))=0$: it is impossible to learn the second coordinate of the mean. (b) is a convex partition of $\mathbb{R}^2$, that makes recovering the Gaussian possible. (c) is the convex partition corresponding to the output distribution of one layer ReLU networks. When both coordinates are positive, we observe a fine sample (black points correspond to singleton sets). When exactly one coordinate (say $\boldsymbol{x}_1$) is positive, we observe the line $L_z = \{\boldsymbol{x} : \boldsymbol{x}_2 < 0, \boldsymbol{x}_1 = z > 0\}$ that corresponds to the ReLU output $(\boldsymbol{x}_1, 0)$. If both coordinates are negative, we observe the set $\{\boldsymbol{x} : \boldsymbol{x}_1 < 0, \boldsymbol{x}_2 < 0\}$, that corresponds to the point $(0,0)$.

Theorem 1.29 ((Informal) - Gaussian Mean Estimation with Convex Partitions)**.** *Let $\epsilon \in (0,1)$. Consider the generative process of coarse d-dimensional Gaussian data $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$. Assume that the partition distribution π is α-information preserving and is supported on convex partitions of $\mathbb{R}^d$. Then, the empirical log-likelihood objective*

$$\mathcal{L}_N(\boldsymbol{\mu}) = \frac{1}{N}\sum_{i=1}^{N} \log \mathcal{N}(\boldsymbol{\mu}; S_i)$$

is concave with respect to $\boldsymbol{\mu}$ for $S_i \sim \mathcal{N}_\pi(\boldsymbol{\mu}^\star)$. Moreover, it suffices to draw $N = \widetilde{O}(d/(\epsilon^2\alpha^2))$ samples from $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$ so that the maximizer $\widetilde{\boldsymbol{\mu}}$ of the empirical log-likelihood satisfies

$$d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}}), \mathcal{N}(\boldsymbol{\mu}^\star)) \le \epsilon\,,$$

with probability at least 99%.

Our algorithm for mean estimation of a Gaussian distribution relies on the log-likelihood being concave when the partitions are convex. We remark that, similar to our approach, one can use the concavity of likelihood to get efficient algorithms for regression settings, e.g., Tobit models, where the mean of the Gaussian is given by a linear function of the context $\boldsymbol{A}\boldsymbol{x}$ for some unknown matrix $\boldsymbol{A}$.

Part I

Learning From Noisy Labels

2 LEARNING WITH MASSART NOISE

2.1 Formal Statement of Results

Preliminaries

Let $\boldsymbol{e}_i$ be the i-th standard basis vector in $\mathbb{R}^d$. For $d \in \mathbb{N}$, let $\mathbb{S}^{d-1} :- \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$. Let $\mathrm{proj}_U(\boldsymbol{x})$ be the projection of $\boldsymbol{x}$ to subspace $U \subset \mathbb{R}^d$ and $U^\perp$ be its orthogonal complement.

We consider the binary classification setting where labeled examples $(\boldsymbol{x}, y)$ are drawn i.i.d. from a distribution D on $\mathbb{R}^d \times \{\pm 1\}$. We denote by $D_{\boldsymbol{x}}$ the marginal of D on $\boldsymbol{x}$. The misclassification error of a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ (with respect to D) is $\mathrm{err}_{0-1}^{D}(h) :- \mathbf{Pr}_{(\boldsymbol{x},y)\sim D}[h(\boldsymbol{x}) \neq y]$. The zero-one error between two functions f, h (with respect to $D_{\boldsymbol{x}}$) is $\mathrm{err}_{0-1}^{D_{\boldsymbol{x}}}(f, h) :- \mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[f(\boldsymbol{x}) \neq h(\boldsymbol{x})]$.

We will use the following simple claim relating the zero-one loss between two halfspaces (with respect to a bounded distribution) and the angle between their normal vectors (see Appendix A.1 for the proof).

Claim 2.1. *Let $D_{\boldsymbol{x}}$ be a (U, R)-bounded distribution on $\mathbb{R}^d$. For any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ we have that $R^2/U \cdot \theta(\boldsymbol{u}, \boldsymbol{v}) \leq \mathrm{err}_{0-1}^{D_{\boldsymbol{x}}}(h_{\boldsymbol{u}}, h_{\boldsymbol{v}})$. Moreover, if $D_{\boldsymbol{x}}$ is $(U, R, t(\cdot))$-bounded, for any $0 < \epsilon \leq 1$, we have that $\mathrm{err}_{0-1}^{D_{\boldsymbol{x}}}(h_{\boldsymbol{u}}, h_{\boldsymbol{v}}) \leq U t(\epsilon)^2 \cdot \theta(\boldsymbol{v}, \boldsymbol{u}) + \epsilon$.*

Our main result is the first polynomial-time algorithm for learning halfspaces with Massart noise with respect to a broad class of well-behaved distributions. Before we formally state our algorithmic result, we define the family of distributions $\mathcal{F}$ for which our algorithm succeeds:

Definition 2.2 (Bounded distributions)**.** *Fix $U, R > 0$ and $t : (0, 1) \to \mathbb{R}_+$. An isotropic (i.e., zero mean and identity covariance) distribution $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ is called (U, R, t)-bounded if for any projection $(D_{\boldsymbol{x}})_V$ of $D_{\boldsymbol{x}}$ onto a 2-dimensional subspace V the corresponding pdf γ_V on $\mathbb{R}^2$ satisfies the following properties:*

1. *$\gamma_V(\boldsymbol{x}) \geq 1/U$, for all $\boldsymbol{x} \in V$ such that $\|\boldsymbol{x}\|_2 \leq R$ (anti-anti-concentration).*

2. *$\gamma_V(\boldsymbol{x}) \leq U$ for all $x \in V$ (anti-concentration).*

3. *For any $\epsilon \in (0,1)$, $\mathbf{Pr}_{\boldsymbol{x}\sim\gamma_V}[\|\boldsymbol{x}\|_2 \geq t(\epsilon)] \leq \epsilon$ (concentration).*

We say that $D_{\boldsymbol{x}}$ is (U,R)-bounded if concentration is not required to hold.

Our main result is the following theorem.

Theorem 2.3. *Let D be a distribution on $\mathbb{R}^d \times \{-1,+1\}$ such that the marginal $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ is $(U,R,t())$-bounded. Let $\eta < 1/2$ be an upper bound on the Massart noise rate. Algorithm 2 has the following performance guarantee: It draws $m = O((U/R)^{12} \cdot t^8(\epsilon/2)/(1-2\eta)^{10}) \cdot O(d/\epsilon^4)$ labeled examples from D, uses $O(m)$ gradient evaluations, and outputs a hypothesis vector $\bar{\boldsymbol{w}}$ that satisfies $\mathrm{err}_{0-1}^{D_x}(h_{\bar{\boldsymbol{w}}}, f) \leq \epsilon$ with probability at least $1-\delta$, where f is the target halfspace.*

2.2 Overview of Techniques

Our approach is extremely simple: We take an optimization view and leverage the structure of the learning problem to identify a simple *non-convex* surrogate loss $\mathcal{L}_\sigma(\boldsymbol{w})$ with the following property: *Any* approximate stationary point $\widehat{\boldsymbol{w}}$ of $\mathcal{L}_\sigma$ defines a halfspace $h_{\widehat{\boldsymbol{w}}}$, which is close to the target halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle)$. Our non-convex surrogate is smooth, by design. Therefore, we can use any first-order method to efficiently find an approximate stationary point.

We now proceed with a high-level intuitive explanation. For simplicity of this discussion, we consider the population versions of the relevant loss functions. The most obvious way to solve the learning problem is by attempting to directly optimize the population risk with respect to the $0-1$ loss, i.e., the misclassification error $\mathbf{Pr}_{(x,y)\sim D}[h_{\boldsymbol{w}}(\boldsymbol{x}) \neq y]$ as a function of the weight vector $\boldsymbol{w}$. Equivalently, we seek to minimize the function $F(\boldsymbol{w}) = \mathbf{E}_{(x,y)\sim D}[\mathbb{1}\{-y\boldsymbol{w}\cdot\boldsymbol{x} \geq 0\}]$, where $\mathbb{1}\{t \geq 0\}$ is the zero-one step function. This is of course a non-convex problem and it is unclear how to efficiently solve directly.

A standard recipe in machine learning to address non-convexity is to replace the $0-1$ loss $F(\boldsymbol{w})$ by an appropriate convex surrogate. This method seems to

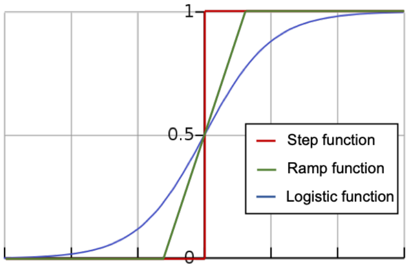


Figure 2.1: The step function and its surrogates.

inherently fail in our setting. However, we are able to find a *non-convex* surrogate that works. Even though finding a global optimum of a non-convex function is hard in general, we show that a much weaker requirement suffices for our learning problem. In particular, it suffices to find a point where our non-convex surrogate has small gradient. Our main structural result is that any such point is close to the target weight vector $\boldsymbol{w}^*$.

To obtain our non-convex surrogate loss $\mathcal{L}_\sigma$, we replace the step function $\mathbb{1}\{t \geq 0\}$ in $F(\boldsymbol{w})$ by a well-behaved approximation. That is, our surrogate is of the form $\mathcal{L}_\sigma(\boldsymbol{w}) = \mathbf{E}_{(\boldsymbol{x},y)\sim D}[r(-y\boldsymbol{w} \cdot \boldsymbol{x})]$, where $r(t)$ is an approximation (in some sense) of $\mathbb{1}\{t \geq 0\}$. A natural first idea is to approximate the step function by a piecewise linear (ramp) function. We show (Section 2.3) that this leads to a non-convex surrogate that indeed satisfies the desired structural property. The proof of this statement turns out to be quite clean, capturing the key intuition of our approach. Unfortunately, the non-convex surrogate obtained this way (i.e., using the ramp function as an approximation to the step function) is non-smooth and it is unclear how to efficiently find an approximate stationary point. A simple way to overcome this obstacle is to instead use an appropriately *smooth* approximation to the step function. Specifically, we use the logistic loss (Section 2.3), but several other choices would work. See Figure 2.1 for an illustration.

We note that our structural lemma (showing that any stationary point of a non-convex surrogate suffices) crucially leverages the underlying distributional assumptions (i.e., the fact that $D_{\boldsymbol{x}}$ is (U, R) bounded). It follows from a lower bound construction in Diakonikolas et al. (2019a) that the approach of this work does not extend to the distribution-independent setting. In particular, for any loss function $\mathcal{L}$, Diakonikolas et al. (2019a) constructs examples where there exist stationary points of $\mathcal{L}$ defining hypotheses that are far from the target halfspace.

2.3 Main Structural Result: Stationary Points Suffice

In this section, we prove our main structural result. In Section 2.3, we define a simple non-convex surrogate by replacing the step function by the (piecewise linear) ramp function and show that any approximate stationary point of this surrogate loss suffices. In Section 2.3, we prove our actual structural result for a smooth (sigmoid-based) approximation to the step function.

Warm-up: Non-convex surrogate based on ramp function

The main point of this subsection is to illustrate the key properties of a non-convex surrogate loss that allows us to argue that the stationary points of this loss are close to the true halfspace $\boldsymbol{w}^*$. To this end, we consider the *ramp function* $r_\sigma(t)$ with parameter $\sigma > 0$ – a piecewise linear approximation to the step function. The ramp function and its derivative are defined as follows:

$$r_\sigma(t) = \begin{cases} 0, & \text{for } t < -\sigma/2 \\ \frac{t}{\sigma} + \frac{1}{2}, & |t| \le \sigma/2 \\ 1, & t > \sigma/2 \end{cases} \qquad \text{and} \qquad r'_\sigma(t) = \frac{1}{\sigma}\mathbb{1}\{|t| \le \sigma/2\} \, . \tag{2.1}$$

Observe that as σ approaches 0, r_σ approaches the step function. Using the ramp function, we define the following non-convex surrogate loss function

$$\mathcal{L}_\sigma^{\mathrm{ramp}}(\boldsymbol{w}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}\left[r_\sigma\left(-y\frac{\boldsymbol{w}\cdot\boldsymbol{x}}{\|\boldsymbol{w}\|_2}\right)\right] \, . \tag{2.2}$$

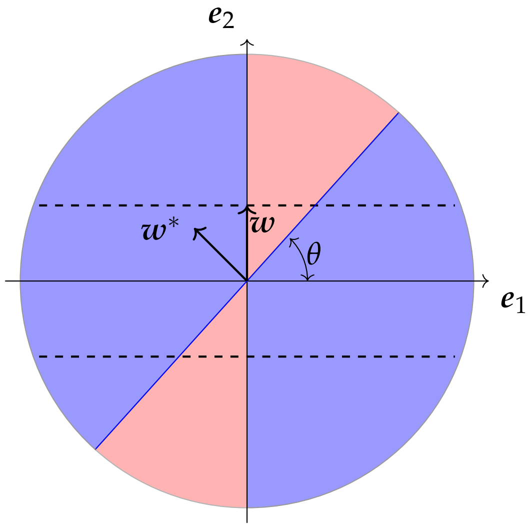


Figure 2.2: The sign of the two-dimensional gradient projection.

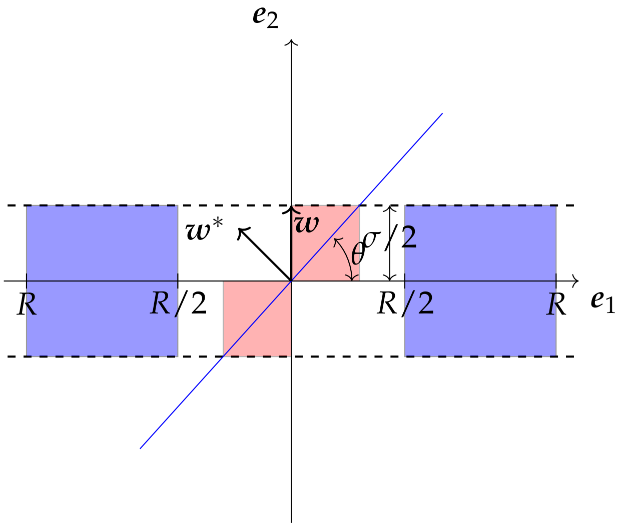


Figure 2.3: The "good" (blue) and "bad" (red) regions inside a band of size σ.

To simplify notation, we will denote the inner product of $\boldsymbol{x}$ and the normalized $\boldsymbol{w}$ as $\ell(\boldsymbol{w}, \boldsymbol{x}) = \frac{\boldsymbol{w} \cdot \boldsymbol{x}}{\|\boldsymbol{w}\|_2}$. By a straightforward calculation (see Appendix A.1), we get that the gradient of the objective $\mathcal{L}_\sigma^{\text{ramp}}(\boldsymbol{w})$ is

$$\nabla_{\boldsymbol{w}} \mathcal{L}_\sigma^{\text{ramp}}(\boldsymbol{w}) = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim D_x} \left[-r'_\sigma \left(\ell(\boldsymbol{w}, \boldsymbol{x}) \right) \; \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{x}) \; (1 - 2\eta(\boldsymbol{x})) \, \text{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x}) \right] \; . \tag{2.3}$$

Our goal is to establish a claim along the following lines.

Claim 2.4 (Informal)**.** *For every $\epsilon > 0$ there exists $\sigma > 0$ such that for any vector $\widehat{\boldsymbol{w}}$ with $\theta(\boldsymbol{w}^*, \widehat{\boldsymbol{w}}) > \epsilon$, it holds $\left\| \nabla_{w} \mathcal{L}_\sigma^{ramp}(\widehat{\boldsymbol{w}}) \right\|_2 \geq \epsilon$.*

The contrapositive of this claim implies that for every ϵ we can tune the parameter σ so that all points with sufficiently small gradient have angle at most ϵ with the optimal halfspace $\boldsymbol{w}^*$. This is a parameter distance guarantee that is easy to translate to missclafication error (using Claim 2.1).

Since it suffices to prove that the norm of the gradient of any "bad" hypothesis (i.e., one whose angle with the optimal is greater than ϵ) is large, we can restrict our attention to any subspace and bound from below the norm of the gradient in that subspace. Let $V = \mathrm{span}(\boldsymbol{w}^*, \boldsymbol{w})$ and note that the inner products $\boldsymbol{w}^* \cdot \boldsymbol{x}$, $\boldsymbol{w} \cdot \boldsymbol{x}$ do not change after the projection to this subspace. Write any point $\boldsymbol{x} \in \mathbb{R}^d$ as $\boldsymbol{v} + \boldsymbol{u}$, where $\boldsymbol{v} \in V$ is the projection of $\boldsymbol{x}$ onto V and $\boldsymbol{u} \in V^{\perp}$. Now, for each $\boldsymbol{v}$, we pick the worst-case $\boldsymbol{u}$ (the one that minimizes the norm of the gradient). We set $\eta_V(\boldsymbol{v}) = \eta_V(\boldsymbol{v} + \boldsymbol{u}(\boldsymbol{v}))$. Since $\eta(\boldsymbol{x}) \leq \eta$ for all $\boldsymbol{x}$, we also have that $\eta_V(\boldsymbol{v}) \leq \eta$, for all $\boldsymbol{v} \in V$. Therefore, we have

$$\left\|\nabla_{\boldsymbol{w}} \mathcal{L}_\sigma^{\mathrm{ramp}}(\boldsymbol{w})\right\|_2 \geq \left\|\mathrm{proj}_V \nabla_{\boldsymbol{w}} \mathcal{L}_\sigma^{\mathrm{ramp}}(\boldsymbol{w})\right\|_2 = \left\|\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D_V}[\nabla_{\boldsymbol{w}} \mathcal{L}_\sigma^{\mathrm{ramp}}(\boldsymbol{w})]\right\|_2 .$$

Without loss of generality, assume that $\widehat{\boldsymbol{w}} = \boldsymbol{e}_2$ and $\boldsymbol{w}^* = -\sin\theta \cdot \boldsymbol{e}_1 + \cos\theta \cdot \boldsymbol{e}_2$, see Figure 2.2. To simplify notation, in what follows we denote by $\eta(\boldsymbol{x})$ the function $\eta_V(\boldsymbol{x})$ after the projection. Observe that the gradient is always perpendicular to $\widehat{\boldsymbol{w}} = \boldsymbol{e}_2$ (this is also clear from the fact that $\mathcal{L}_\sigma^{\mathrm{ramp}}(\boldsymbol{w})$ does not depend on the length of $\boldsymbol{w}$). Therefore,

$$\begin{aligned}\left\|\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D_V}[\nabla_{\boldsymbol{w}} \mathcal{L}_\sigma^{\mathrm{ramp}}(\widehat{\boldsymbol{w}})]\right\|_2 &= |\nabla_{\boldsymbol{w}} \mathcal{L}_\sigma^{\mathrm{ramp}}(\widehat{\boldsymbol{w}}) \cdot \boldsymbol{e}_1| \\ &= \left|\mathop{\mathbf{E}}_{\boldsymbol{x}\sim (D_{\boldsymbol{x}})_V}[-r'_\sigma(\boldsymbol{x}_2)(1-2\eta(\boldsymbol{x}))\mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})\boldsymbol{x}_1]\right| . \quad (2.4)\end{aligned}$$

We partition $\mathbb{R}^2$ in two regions according to the sign of the pointwise gradient

$$g(\boldsymbol{x}) = -r'_\sigma(\boldsymbol{x}_2)(1-2\eta(\boldsymbol{x}))\mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})\boldsymbol{x}_1 .$$

Let

$$G = \{\boldsymbol{x} \in \mathbb{R}^2 : g(\boldsymbol{x}) \geq 0\} = \{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x}_1 \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x}) \leq 0\} ,$$

and let G^c be its complement. See Figure 2.2 for an illustration. To give some intuition behind this definition, imagine we were using SGD in this 2-dimensional setting, and at some step t we have $\boldsymbol{w}^{(t)} = \widehat{\boldsymbol{w}} = \boldsymbol{e}_2$. We draw a sample $(\boldsymbol{x}, y)$ from the distribution D and update the hypothesis. Then the expected update (with respect to the label y) is

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{e}_2 - g(\boldsymbol{x}) \cdot \boldsymbol{e}_1 \boldsymbol{e}_1 \ .$$

Therefore, assuming that $\theta(\boldsymbol{w}^*, \boldsymbol{e}_2) \in (0, \pi/2)$, the "good" points (region G) are those that decrease the $\boldsymbol{e}_1$ component (i.e., rotate the hypothesis counter-clockwise) and the "bad" points (region G^c) are those that try to increase the $\boldsymbol{e}_1$ component (rotate the hypothesis clockwise); see Figure 2.2.

We are now ready to explain the main idea behind the choice of the ramp function $r_\sigma(t)$. Recall that the derivative of the ramp function is the (scaled) indicator of a band of size $\sigma/2$ around 0, $r'_\sigma(t) = (1/\sigma)\mathbb{1}\{|t| \leq \sigma/2\}$. Therefore, the gradient of this loss function amplifies the contribution of points close to the current guess $\boldsymbol{w}$, that is, points inside the band $\mathbb{1}\{|\boldsymbol{x}_2| \leq \sigma/2\}$ in our 2-dimensional example of Figure 2.2. Assume for simplicity that the marginal distribution $\mathcal{D}_{\boldsymbol{x}}$ is the uniform distribution on the 2-dimensional unit ball. Then, no matter how small the angle of the true halfspace and our guess $\theta(\boldsymbol{w}^*, \widehat{\boldsymbol{w}})$ is, we can always pick σ sufficiently small so that the contribution of the "good" points (blue region in Figure 2.2) is much larger than the contribution of the "bad" points (red region).

Crucial in this argument is the fact that the distribution is "well-behaved" in the sense that the probability of every region is related to its area. This is where Definition 2.2 comes into play. To bound from below the contribution of "good" points, we require the anti-anti-concentration property of the distribution, namely a lower bound on the density function (in some bounded radius). To bound from above the contribution of "bad" points, we need the anti-concentration property of Definition 2.2, namely that the density is bounded from above (recall that we wanted the probability of a region to be related to its area).

We are now ready to show that our ramp-based non-convex loss works for all distributions satisfying Definition 2.2. In the following lemma, we prove that we can tune the parameter σ so that the stationary points of our non-convex loss are close to $\boldsymbol{w}^*$. The following lemma is a precise version of our initial informal goal, Claim 2.4.

Lemma 2.5 (Stationary points of $\mathcal{L}_\sigma^{\mathrm{ramp}}$ suffice)**.** *Let $D_{\boldsymbol{x}}$ be a (U,R)-bounded distribution on $\mathbb{R}^d$, and $\eta < 1/2$ be an upper bound on the Massart noise rate. Fix any $\theta \in (0,\pi/2)$. Let $\boldsymbol{w}^* \in \mathbb{S}^{d-1}$ be the normal vector to the optimal halfspace and $\widehat{\boldsymbol{w}} \in \mathbb{S}^{d-1}$ be such that $\theta(\widehat{\boldsymbol{w}},\boldsymbol{w}^*) \in (\theta, \pi-\theta)$. For $\sigma \le \frac{R}{2U}\sqrt{1-2\eta}\sin\theta$, we have that $\left\|\nabla_{\boldsymbol{w}} \mathcal{L}_\sigma^{ramp}(\widehat{\boldsymbol{w}})\right\|_2 \ge (1/8)R^2(1-2\eta)/U$.*

Proof. We will continue using the notation introduced in the above discussion. We let V be the 2-dimensional subspace spanned by $\boldsymbol{w}^*$ and $\widehat{\boldsymbol{w}}$. To simplify notation, we again assume without loss of generality that $\boldsymbol{w}^* = -\sin\theta\ \boldsymbol{e}_1 + \cos\theta\ \boldsymbol{e}_2$ and $\widehat{\boldsymbol{w}} = \boldsymbol{e}_2$, see Figure 2.2. Using the triangle inequality and Equation (2.4), we obtain

$$\left\| \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D_V} [\nabla_{\boldsymbol{w}} \mathcal{L}_\sigma^{\mathrm{ramp}}(\widehat{\boldsymbol{w}})] \right\|_2 \tag{2.5}$$

$$\ge \mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V} \left[r'_\sigma(\boldsymbol{x}_2)(1-2\eta(\boldsymbol{x}))|\boldsymbol{x}_1| \mathbb{1}_G(\boldsymbol{x}) \right] - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V} \left[r'_\sigma(\boldsymbol{x}_2)(1-2\eta(\boldsymbol{x}))|\boldsymbol{x}_1| \mathbb{1}_{\boldsymbol{x}\in G^c} \right]$$

$$= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V} \left[r'_\sigma(\boldsymbol{x}_2)(1-2\eta(\boldsymbol{x}))|\boldsymbol{x}_1| \right] - 2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V} \left[r'_\sigma(\boldsymbol{x}_2)(1-2\eta(\boldsymbol{x}))|\boldsymbol{x}_1| \mathbb{1}_{\boldsymbol{x}\in G^c} \right] . \tag{2.6}$$

We now bound from below the first term, as follows

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V} \left[r'_\sigma(\boldsymbol{x}_2)(1-2\eta(\boldsymbol{x}))|\boldsymbol{x}_1| \right] \\
&\ge (1-2\eta) \mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V} \left[\frac{\mathbb{1}\{|\boldsymbol{x}_2| \le \sigma/2\}}{\sigma} |\boldsymbol{x}_1| \right] \\
&\ge \frac{(1-2\eta)R}{2\sqrt{2}\sigma} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V} \left[\mathbb{1}\left\{ |\boldsymbol{x}_2| \le \frac{\sigma}{2}, \frac{R}{2\sqrt{2}} \le |\boldsymbol{x}_1| \le \frac{R}{\sqrt{2}} \right\} \right] \\
&\ge \frac{(1-2\eta)R}{2\sqrt{2}\sigma} \cdot \frac{R\sigma}{\sqrt{2}U} = \frac{R^2}{4U}(1-2\eta),
\end{aligned} \tag{2.7}$$

where the first inequality follows from the upper bound on the noise $\eta(\boldsymbol{x}) \le \eta$, and the third one from the lower bound on the 2-dimensional density function $1/U$ inside the ball $\|\boldsymbol{x}\|_2 \le R$ (see Definition 2.2).

We next bound from above the second term of Equation (2.6), that is the contribution of "bad" points. We have that

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim(D_x)_V}\left[r'_\sigma(x_2)(1-2\eta(x))|x_1|\mathbb{1}_{x\in G^c}\right] &\le \mathop{\mathbf{E}}_{x\sim(D_x)_V}\left[\frac{\mathbb{1}\{|x_2|\le\sigma/2\}}{\sigma}|x_1|\mathbb{1}\{x\in G^c\}\right]\\ &\le \frac{1}{\sigma}\mathop{\mathbf{E}}_{x\sim(D_x)_V}\left[|x_1|\mathbb{1}\{x\in G^c,|x_2|\le\sigma/2\}\right]\ .\end{aligned}$$

We now observe that for $\theta\in(0,\pi/2]$ it holds

$$\begin{aligned}G^c=\{x:x_1\mathrm{sign}(w^*\cdot x)>0\}&=\{x:x_1\mathrm{sign}(-x_1\sin\theta+x_2\cos\theta)>0\}\\ &\subseteq\{x:x_1x_2>0\}\ .\end{aligned}$$

On the other hand, if $\theta\in(\pi/2,\pi]$ we have $G^c\subseteq\{x:x_1x_2<0\}$. Assume first that $\theta\in(0,\pi/2]$ (the same argument works also for the other case). Then the intersection of the band $\{x:|x_2|\le\sigma/2\}$ and G^c is contained in the union of two rectangles $\mathcal{R}=\{x:|x_1|\le\sigma/(2\tan\theta),\ |x_2|\le\sigma/2,\ x_1x_2>0\}$, see Figure 2.3. Therefore,

$$\begin{aligned}&\mathop{\mathbf{E}}_{x\sim(D_x)_V}\left[r'_\sigma(x_2)(1-2\eta(x))|x_1|\mathbb{1}_{x\in G^c}\right] \qquad (2.8)\\ &\le\frac{1}{\sigma}\frac{\sigma}{2\tan\theta}\mathop{\mathbf{E}}_{x\sim(D_x)_V}\left[\mathbb{1}\{x\in G^c,|x_1|\le\frac{\sigma}{2\tan\theta},|x_2|\le\frac{\sigma}{2}\}\right]\\ &\le\frac{1}{\sigma}\frac{\sigma}{2\tan\theta}\mathop{\mathbf{E}}_{x\sim(D_x)_V}\left[\mathbb{1}\{x\in R\}\right]\le\frac{1}{2\tan\theta}\cdot\frac{U\sigma^2}{2\tan\theta}\\ &\le\frac{R^2}{16U}(1-2\eta)\ , \qquad (2.9)\end{aligned}$$

where for the last inequality we used our assumption that $\sigma\le\frac{R}{2U}\sqrt{1-2\eta}\sin\theta$. To finish the proof, we substitute the bounds (2.7), (2.9) in Equation (2.6). □

Main structural result: Non-convex surrogate via smooth approximation

In this subsection, we prove the structural result that is required for the correctness of our efficient gradient-descent algorithm in the following section. We consider the non-convex surrogate loss

$$\mathcal{L}_\sigma(\boldsymbol{w}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}\left[S_\sigma\left(-y\frac{\boldsymbol{w}\cdot\boldsymbol{x}}{\|\boldsymbol{w}\|_2}\right)\right], \tag{2.10}$$

where $S_\sigma(t) = \frac{1}{1+e^{-t/\sigma}}$ is the logistic function with growth rate $1/\sigma$. That is, we have replaced the step function by the sigmoid. As $\sigma \to 0$, $S_\sigma(t)$ approaches the step function. Formally, we prove the following:

Lemma 2.6 (Stationary points of $\mathcal{L}_\sigma$ suffice)**.** *Let $D_{\boldsymbol{x}}$ be a (U, R)-bounded distribution on $\mathbb{R}^d$, and $\eta < 1/2$ be an upper bound on the Massart noise rate. Fix any $\theta \in (0, \pi/2)$. Let $\boldsymbol{w}^* \in \mathbb{S}^{d-1}$ be the normal vector to the optimal halfspace and $\widehat{\boldsymbol{w}} \in \mathbb{S}^{d-1}$ be such that $\theta(\widehat{\boldsymbol{w}}, \boldsymbol{w}^*) \in (\theta, \pi-\theta)$. For $\sigma \le \frac{R}{8U}\sqrt{1-2\eta}\sin\theta$, we have that $\|\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\widehat{\boldsymbol{w}})\|_2 \ge \frac{1}{32U}R^2(1-2\eta)$.*

The proof of Lemma 2.6 is conceptually similar to the proof of Lemma 2.5 for the ramp function given in the previous subsection. The main difference is that, in the smoothed setting, it is harder to bound the contribution of each region of Figure 2.2 and the calculations end-up being more technical.

Proof of Lemma 2.6. Without loss of generality, we will assume that $\widehat{\boldsymbol{w}} = \boldsymbol{e}_2$ and $\boldsymbol{w}^* = -\sin\theta\cdot\boldsymbol{e}_1 + \cos\theta\cdot\boldsymbol{e}_2$. Using the same argument as in the proof of Section 2.3, we let $V = \mathrm{span}(\boldsymbol{w}^*, \boldsymbol{w})$ and have

$$\left\|\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D_V}[\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\widehat{\boldsymbol{w}})]\right\|_2 = \left|\mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V}[-S'_\sigma(|\boldsymbol{x}_2|)(1-2\eta(\boldsymbol{x}))\mathrm{sign}(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{x}_1]\right| . \tag{2.11}$$

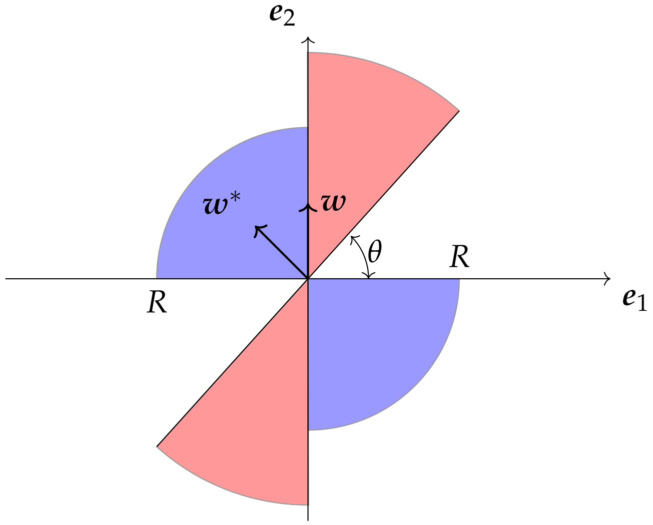


Figure 2.4: The “good" (blue) and “bad" (red) regions.

We partition $\mathbb{R}^2$ in two regions according to the sign of the gradient. Let

$$G = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 \mathrm{sign}(w^* \cdot x) > 0\} \ ,$$

and let G^c be its complement. Using the triangle inequality and Equation (2.11), we obtain

$$\begin{aligned}
&\left\| \mathop{\mathbf{E}}_{(x,y)\sim D_V} [\nabla_w \mathcal{L}_\sigma(\widehat{w})] \right\|_2 \\
&\geq \mathop{\mathbf{E}}_{x\sim (D_x)_V} \left[S'_\sigma(|x_2|)(1-2\eta(x))|x_1| \mathbb{1}_G(x) \right] - \mathop{\mathbf{E}}_{x\sim (D_x)_V} \left[S'_\sigma(|x_2|)(1-2\eta(x))|x_1| \mathbb{1}_{G^c}(x) \right] \\
&\geq \frac{(1-2\eta)}{4} \mathop{\mathbf{E}}_{x\sim (D_x)_V} \left[\frac{e^{-|x_2|/\sigma}}{\sigma} \cdot |x_1| \cdot \mathbb{1}_G(x) \right] - \mathop{\mathbf{E}}_{x\sim (D_x)_V} \left[\frac{e^{-|x_2|/\sigma}}{\sigma} \cdot |x_1| \cdot \mathbb{1}_{G^c}(x) \right] \ ,
\end{aligned} \tag{2.12}$$

where we used the upper bound on the Massart noise rate $\eta(x) \leq \eta$ and the fact that the sigmoid $S_\sigma(|t|)^2$ is bounded from above by 1 and bounded from below by 1/4.

We can now bound each term separately using the fact that the distribution is (U,R)-bounded. Assume first that $\theta(\boldsymbol{w}^*,\widehat{\boldsymbol{w}}) = \theta \in (0,\pi/2)$. Then we can express the region in polar coordinates as $G = \{(r,\phi) : \phi \in (0,\theta) \cup (\pi/2, \pi+\theta) \cup (3\pi/2, 2\pi)\}$. See Figure 2.4 for an illustration.

We denote by $\gamma(x,y)$ the density of the 2-dimensional projection on V of the marginal distribution $D_{\boldsymbol{x}}$. Since the integral is non-negative, we can bound from below the contribution of region G on the gradient by integrating over $\phi \in (\pi/2,\pi)$. Specifically, we have:

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V}\left[\frac{e^{-|\boldsymbol{x}_2|/\sigma}}{\sigma}\,|\boldsymbol{x}_1|\;\mathbb{1}_G(\boldsymbol{x})\right] &\geq \int_0^\infty \int_{\pi/2}^{\pi} \gamma(r\cos\phi, r\sin\phi) r^2 |\cos\phi| \frac{e^{-\frac{r\sin\phi}{\sigma}}}{\sigma} \mathrm{d}\phi \mathrm{d}r \\
&= \int_0^\infty \int_0^{\pi/2} \gamma(r\cos\phi, r\sin\phi) r^2 \cos\phi \frac{e^{-\frac{r\sin\phi}{\sigma}}}{\sigma} \mathrm{d}\phi \mathrm{d}r \\
&\geq \frac{1}{U}\int_0^R r^2 \mathrm{d}r \int_0^{\pi/2} \cos\phi \frac{e^{-\frac{R\sin\phi}{\sigma}}}{\sigma} \mathrm{d}\phi \\
&= \frac{1}{3U} R^2 \left(1 - e^{-\frac{R}{\sigma}}\right) \geq \frac{1}{3U} R^2 \left(1 - e^{-8}\right) , \qquad (2.13)
\end{aligned}$$

where for the second inequality we used the lower bound $1/U$ on the density function $\gamma(x,y)$ (see Definition 2.2) and for the last inequality we used that $\sigma \leq \frac{R}{8}$.

We next bound from above the contribution of the gradient in region G^c. Note that $G^c = \{(r,\phi) : \phi \in B_\theta = (\pi/2-\theta, \pi/2) \cup (3\pi/2-\theta, 3\pi/2)\}$. Hence, we can write:

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V}\left[\frac{e^{-|\boldsymbol{x}_2|/\sigma}}{\sigma}\,|\boldsymbol{x}_1|\;\mathbb{1}_{G^c}(\boldsymbol{x})\right] &= \int_0^\infty \int_{\phi\in B_\theta} \gamma(r\cos\phi, r\sin\phi) r^2 \cos\phi e^{-\frac{r\sin\phi}{\sigma}} \mathrm{d}\phi \mathrm{d}r \\
&\leq \frac{2U}{\sigma}\int_0^\infty \int_\theta^{\pi/2} r^2 \cos\phi e^{-\frac{r\sin\phi}{\sigma}} \mathrm{d}\phi \mathrm{d}r \\
&= \frac{2U\sigma^2\cos^2\theta}{\sin^2\theta} \\
&= \frac{(1-2\eta)R^2}{32U}\cos^2\theta , \qquad (2.14)
\end{aligned}$$

where the inequality follows from the upper bound U on the density $\gamma(x,y)$ (see

Definition 2.2) and the last inequality follows from our assumption that $\sigma \leq \frac{R}{8U}\sqrt{1-2\eta}\sin(\theta)$. Combining (2.13) and (2.14), we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V}\left[\frac{e^{-|x_2|/\sigma}}{\sigma}\,|\boldsymbol{x}_1|\,\mathbb{1}_{G^c}(\boldsymbol{x})\right] &\leq \frac{(1-2\eta)R^2}{32U}\cos^2\theta \\ &\leq \frac{(1-2\eta)R^2\left(1-e^{-8}\right)}{24U} \\ &\leq \frac{1}{2}\frac{(1-2\eta)}{4}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V}\left[\frac{e^{-|x_2|/\sigma}}{\sigma}\,|\boldsymbol{x}_1|\,\mathbb{1}_{G}(\boldsymbol{x})\right]\,, \quad (2.15)\end{aligned}$$

where the second inequality follows from $\cos^2\theta \leq 1$ and $\frac{1}{32} \leq \frac{\left(1-e^{-8}\right)}{24}$. Using (2.15) in (2.12), we obtain

$$\begin{aligned}\left\|\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D_V}\left[\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\widehat{\boldsymbol{w}})\right]\right\|_2 &\geq \frac{1}{2}\frac{(1-2\eta)}{4}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V}\left[\frac{e^{-|x_2|/\sigma}}{\sigma}\,|\boldsymbol{x}_1|\,\mathbb{1}_{G}(\boldsymbol{x})\right] \\ &\geq \frac{1}{32U}(1-2\eta)\,R^2\,.\end{aligned}$$

To conclude the proof, notice that the case where $\theta(\widehat{\boldsymbol{w}}, \boldsymbol{w}^*) \in (\pi/2, \pi-\theta)$ follows similarly. Finally, in the case where $\theta = \pi/2$, the region G^c is empty, and we again get the same lower bound on the gradient. This completes the proof of Lemma 2.6. □

2.4 Main Algorihtmic Result: Proof of Theorem 2.3

In this section, we prove our main algorithmic result.

Our algorithm proceeds by Projected Stochastic Gradient Descent (PSGD), with projection on the ℓ_2-unit sphere, to find an approximate stationary point of our non-convex surrogate loss. Since $\mathcal{L}_\sigma(\boldsymbol{w})$ is non-smooth for vectors $\boldsymbol{w}$ close to $\mathbf{0}$, at each step, we project the update on the unit sphere to avoid the region where the smoothness parameter is high.

Recall that a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is called L-Lipschitz if there is a parameter

$L > 0$ such that $\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2 \le L \|\boldsymbol{x} - \boldsymbol{y}\|_2$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. We will make use of the following folklore result on the convergence of projected SGD (for completeness, we provide a proof in Appendix A.2).

Algorithm 1 PSGD for $f(\boldsymbol{w}) = \mathbf{E}_{z \sim D}[g(\boldsymbol{z}, \boldsymbol{w})]$

```
procedure PSGD($f, T, \beta$)    ▷ $f(\boldsymbol{w}) = \mathbf{E}_{z\sim D}[g(\boldsymbol{z}, \boldsymbol{w})]$: loss, $T$: number of steps, $\beta$: step size.
    $\boldsymbol{w}^{(0)} \leftarrow \boldsymbol{e}_1$
    for $i = 1, \ldots, T$ do
        Sample $\boldsymbol{z}^{(i)}$ from $D$.
        $\boldsymbol{v}^{(i)} \leftarrow \boldsymbol{w}^{(i-1)} - \beta \nabla_{\boldsymbol{w}} g(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i-1)})$
        $\boldsymbol{w}^{(i)} \leftarrow \boldsymbol{v}^{(i)} / \left\|\boldsymbol{v}^{(i)}\right\|_2$
    return $(\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)})$.
```

Lemma 2.7 (PSGD). *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ with $f(\boldsymbol{w}) = \mathbf{E}_{z\sim D}[g(\boldsymbol{z}, \boldsymbol{w})]$ for some function $g : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$. Assume that for any vector $\boldsymbol{w}$, $g(\cdot, \boldsymbol{w})$ is positive homogeneous of degree-0 on $\boldsymbol{w}$. Let $\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_2 \ge 1\}$ and assume that f, g are continuously differentiable functions on $\mathcal{W}$. Moreover, assume that $|f(\boldsymbol{w})| \le R$, $\nabla_{\boldsymbol{w}} f(\boldsymbol{w})$ is L-Lipschitz on $\mathcal{W}$, $\mathbf{E}_{z\sim D}\left[\|\nabla_{\boldsymbol{w}} g(\boldsymbol{z}, \boldsymbol{w})\|_2^2\right] \le B$ for all $\boldsymbol{w} \in \mathcal{W}$. After T iterations the output $(\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)})$ of Algorithm 1 satisfies*

$$\mathop{\mathbf{E}}_{\boldsymbol{z}^{(1)},\ldots,\boldsymbol{z}^{(T)}\sim D}\left[\frac{1}{T}\sum_{i=1}^{T}\left\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i)})\right\|_2^2\right] \le \sqrt{\frac{LBR}{2T}} \, .$$

If, additionally, $\|\mathbf{E}_{z\sim D}[\nabla_{\boldsymbol{w}} g(\boldsymbol{z}, \boldsymbol{w})]\|_2^2 \le C$ for all $\boldsymbol{w} \in \mathcal{W}$, we have that with $T = (2LBR + 8C^2 \log(1/\delta))/\epsilon^4$ it holds $\min_{i=1,\ldots,T} \left\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i)})\right\|_2 \le \epsilon$, with probability at least $1 - \delta$.

We will require the following lemma establishing the smoothness properties of our loss (based on S_σ). See Appendix A.2 for the proof.

Lemma 2.8 (Sigmoid Smoothness). *Let $S_\sigma(t) = 1/(1+e^{-t/\sigma})$ and*

$$\mathcal{L}_\sigma(\boldsymbol{w}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}\left[S_\sigma\left(-y\frac{\boldsymbol{w}\cdot\boldsymbol{x}}{\|\boldsymbol{w}\|_2}\right)\right],$$

for $\boldsymbol{w} \in \mathcal{W}$, where $\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_2 \geq 1\}$. We have that $\mathcal{L}_\sigma(\boldsymbol{w})$ is continuously differentiable in $\mathcal{W}$, $|\mathcal{L}_\sigma(\boldsymbol{w})| \leq 1$, $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[\|\nabla_{\boldsymbol{w}} S_\sigma(\boldsymbol{w},\boldsymbol{x},y)\|_2^2] \leq 4d/\sigma^2$, $\|\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\boldsymbol{w})\|_2^2 \leq 4/\sigma^2$, and $\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\boldsymbol{w})$ is $(6/\sigma + 12/\sigma^2)$-Lipschitz.

Putting everything together gives Theorem 2.3.

Algorithm 2 Learning Halfspaces with Massart Noise

```
procedure ALG($\epsilon$, $U$, $R$, $t(\cdot)$)
    $C_1 \leftarrow \Theta(U^{12}/R^{12})$.
    $C_2 \leftarrow \Theta(R/U^2)$.
    $T \leftarrow C_1\, d\, t(\epsilon/2)^8/(\epsilon^4(1-2\eta)^{10})\log(1/\delta)$.        ▷ number of steps
    $\beta \leftarrow C_2^2\, d(1-2\eta)^3\epsilon^2/(t(\epsilon/2)^4 T^{1/2})$.        ▷ step size
    $\sigma \leftarrow C_2\sqrt{1-2\eta}\,\epsilon/t^2(\epsilon/2)$.
    $(\boldsymbol{w}^{(0)}, \boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)}) \leftarrow \mathrm{PSGD}(f, T, \beta)$.        ▷ $f(\boldsymbol{w}) = \mathbf{E}_{(\boldsymbol{x},y)\sim D}\left[S_\sigma\left(-y\frac{\boldsymbol{w}\cdot\boldsymbol{x}}{\|\boldsymbol{w}\|_2}\right)\right]$, (1)
    $L \leftarrow \{\pm\boldsymbol{w}^{(i)}\}_{i\in[T]}$.        ▷ $L$: List of candidate vectors
    Draw $N = O(\log(T/\delta)/(\epsilon^2(1-2\eta)^2))$ samples from $D$.
    $\bar{\boldsymbol{w}} \leftarrow \mathrm{argmin}_{\boldsymbol{w}\in L}\sum_{j=1}^N \mathbb{1}\{\mathrm{sign}(\boldsymbol{w}\cdot\boldsymbol{x}^{(j)}) \neq y^{(j)}\}$.
    return $\bar{\boldsymbol{w}}$.
```

Proof of Theorem 2.3. By Claim 2.1, to guarantee $\mathrm{err}_{0-1}^{D_x}(h_{\bar{\boldsymbol{w}}}, f) \leq \epsilon$ it suffices to show that the angle $\theta(\bar{\boldsymbol{w}}, \boldsymbol{w}^*) \leq O(\epsilon(1-2\eta)/(Ut^2(\epsilon/2))) =: \theta_0$. Using (the contrapositive of) Lemma 2.6, we get that with $\sigma = \Theta((R/U)\sqrt{1-2\eta}\theta_0)$, if the norm squared of the gradient of some vector $\boldsymbol{w} \in \mathbb{S}^{d-1}$ is smaller than $\rho = O((R^2/U)(1-2\eta))$, then $\boldsymbol{w}$ is close to either $\boldsymbol{w}^*$ or $-\boldsymbol{w}^*$ – that is, $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \leq \theta_0$ – or $\theta(\boldsymbol{w}, -\boldsymbol{w}^*) \leq \theta_0$. Therefore, it suffices to find a point $\boldsymbol{w}$ with gradient $\|\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\boldsymbol{w})\|_2 \leq \rho$.

From Lemma 2.8, we have that our PSGD objective function is bounded above by 1,

$$\mathbf{E}\left[\left\|\nabla_{\boldsymbol{w}} S_\sigma\Big(-y\frac{\boldsymbol{w}\cdot\boldsymbol{x}}{\|\boldsymbol{w}\|_2}\Big)\right\|_2^2\right] \le O(d/\sigma^2)\;,$$

$\left\|\mathbf{E}\left[\nabla_{\boldsymbol{w}} S_\sigma\Big(-y\frac{\boldsymbol{w}\cdot\boldsymbol{x}}{\|\boldsymbol{w}\|_2}\Big)\right]\right\|_2^2 \le O(1/\sigma^2)$, and that the gradient is Lipschitz with Lipschitz constant $O(1/\sigma^2)$. Using these bounds for the parameters of Lemma 2.7, we get that with $T = O(\frac{d}{\sigma^4\rho^4}\log(1/\delta))$ steps, the norm of the gradient of some vector in the list $(\boldsymbol{w}^{(0)},\ldots,\boldsymbol{w}^{(T)})$ will be at most ρ with probability $1-\delta$. Therefore, the required number of iterations is

$$T = O\left(d\frac{U^{12}}{R^{12}}\frac{t^8(\epsilon/2)\log(1/\delta)}{\epsilon^4(1-2\eta)^{10}}\right)\;.$$

We know that one of the hypotheses in the list L (line 8 of Algorithm 2) is ϵ-close to the true $\boldsymbol{w}^*$. We can evaluate all of them on a small number of samples from the distribution D to obtain the best among them. From Hoeffding's inequality, it follows that $N = O(\log(T/\delta)/(\epsilon^2(1-2\eta)^2))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most $\epsilon(1-2\eta)$. Using Fact A.2, for any hypotheses h, and the target concept f, it holds $\mathrm{err}_{0-1}^{D_x}(h,f) \le \frac{1}{(1-2\eta)}(\mathrm{err}_{0-1}^{D}(h) - \mathrm{opt})$, and therefore the chosen hypothesis achieves error at most 2ϵ. This completes the proof of Theorem 2.3. □

2.5 Strong Massart Noise Model

We start by defining the strong Massart noise model, which was considered in Zhang et al. (2017b) for the special case of the uniform distribution on the sphere. The main difference with the standard Massart noise model is that, in the strong model, the noise rate is allowed to approach arbitrarily close to 1/2 for points that lie very close to the separating hyperplane.

Definition 2.9 (Distribution-specific PAC Learning with Strong Massart Noise). *Let $\mathcal{C}$ be the concept class of halfspaces over $X = \mathbb{R}^d$, $\mathcal{F}$ be a* known family *of structured* *distributions on* X, $0 < c \leq 1$ *and* $0 < \epsilon < 1$. *Let* $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$ *be an unknown target function in* $\mathcal{C}$. *A* noisy example oracle, $\mathrm{EX}^{\mathrm{SMas}}(f, \mathcal{F}, \eta)$, *works as follows: Each time* $\mathrm{EX}^{\mathrm{SMas}}(f, \mathcal{F}, \eta)$ *is invoked, it returns a labeled example* $(\boldsymbol{x}, y)$, *such that: (a)* $\boldsymbol{x} \sim D_{\boldsymbol{x}}$, *where* $D_{\boldsymbol{x}}$ *is a fixed distribution in* $\mathcal{F}$, *and (b)* $y = f(\boldsymbol{x})$ *with probability* $1 - \eta(\boldsymbol{x})$ *and* $y = -f(\boldsymbol{x})$ *with probability* $\eta(\boldsymbol{x})$, *for an* unknown *parameter* $\eta(\boldsymbol{x}) \leq \max\{1/2 - c|\boldsymbol{w}^* \cdot \boldsymbol{x}|, 0\}$. *Let D denote the joint distribution on* $(\boldsymbol{x}, y)$ *generated by the above oracle. A learning algorithm is given i.i.d. samples from D and its goal is to output a hypothesis h such that with high probability the misclassification error of h is* ϵ*-close to the misclassfication error of* f, *i.e., it holds* $\mathrm{err}_{0-1}^{D}(h) \leq \mathrm{err}_{0-1}^{D}(f) + \epsilon$.

The main result of this section is the following theorem:

Theorem 2.10 (Learning Halfspaces with Strong Massart Noise). *Let D be a distribution on* $\mathbb{R}^d \times \{\pm 1\}$ *such that the marginal* $D_{\boldsymbol{x}}$ *on* $\mathbb{R}^d$ *is* $(U, R, t())$*-bounded. Let* $0 < c < 1$ *be the parameter of the strong Massart noise model. Algorithm 3 has the following performance guarantee: It draws* $m = O\left((U^{12}/R^{18})(t^8(\epsilon/2)/c^6)\right) O(d/\epsilon^4)$ *labeled examples from D, uses* $O(m)$ *gradient evaluations, and outputs a hypothesis vector* $\bar{\boldsymbol{w}}$ *that satisfies* $\mathrm{err}_{0-1}^{D}(h_{\bar{\boldsymbol{w}}}) \leq \mathrm{err}_{0-1}^{D}(f) + \epsilon$ *with probability at least* $1 - \delta$.

The proof of Theorem 2.10 follows along the same lines as in the previous sections. We show that any stationary point of our non-convex surrogate suffices and then use projected SGD.

The main structural result of this section generalizes Lemma 2.6:

Lemma 2.11 (Stationary points of $\mathcal{L}_\sigma$ suffice with strong Massart noise). *Let* $D_{\boldsymbol{x}}$ *be a* (U, R)*-bounded distribution on* $\mathbb{R}^d$, *and let* $c \in (0, 1)$ *be the parameter of strong Massart noise model. Let* $\theta \in (0, \pi/2)$. *Let* $\boldsymbol{w}^* \in \mathbb{S}^{d-1}$ *be the normal vector to an optimal halfspace and* $\widehat{\boldsymbol{w}} \in \mathbb{S}^{d-1}$ *be such that* $\theta(\widehat{\boldsymbol{w}}, \boldsymbol{w}^*) \in (\theta, \pi - \theta)$. *For* $\sigma \leq \frac{R}{24U}\sqrt{cR}\sin(\theta)$, *we have* $\|\nabla_{\boldsymbol{w}} \mathcal{L}_\sigma(\widehat{\boldsymbol{w}})\|_2 \geq \frac{1}{288U} c\, R^3$.

Proof. Without loss of generality, we can assume that $\widehat{\boldsymbol{w}} = \boldsymbol{e}_2$ and $\boldsymbol{w}^* = -\sin\theta \cdot \boldsymbol{e}_1 + \cos\theta \cdot \boldsymbol{e}_2$. Using the same argument as in the Section 2.3, for $V = \mathrm{span}(\boldsymbol{w}^*, \boldsymbol{w})$,

we have

$$\left\| \mathop{\mathbf{E}}_{(x,y)\sim D_V}[\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\widehat{\boldsymbol{w}})] \right\|_2 = |\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\widehat{\boldsymbol{w}}) \cdot \boldsymbol{e}_1| \\ = \left| \mathop{\mathbf{E}}_{x\sim D_x}[-S'_\sigma(|\boldsymbol{x}_2|)(1-2\eta(\boldsymbol{x}))\mathrm{sign}(\boldsymbol{w}^*\cdot\boldsymbol{x})\boldsymbol{x}_1] \right| \tag{2.16}$$

We partition $\mathbb{R}^2$ in two regions according to the sign of the gradient. Let $G = \{(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^2 : \boldsymbol{x}_1\mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x}) > 0\}$, and let G^c be its complement. Using the triangle inequality and Equation (2.16) we obtain

$$\left\| \mathop{\mathbf{E}}_{(x,y)\sim D_V}[\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\widehat{\boldsymbol{w}})] \right\|_2 \\ \geq \mathop{\mathbf{E}}_{x\sim D_x}[S'_\sigma(|\boldsymbol{x}_2|)(1-2\eta(\boldsymbol{x}))|\boldsymbol{x}_1|\mathbb{1}_G(\boldsymbol{x})] - \mathop{\mathbf{E}}_{x\sim D_x}[S'_\sigma(|\boldsymbol{x}_2|)(1-2\eta(\boldsymbol{x}))|\boldsymbol{x}_1|\mathbb{1}_{G^c}(\boldsymbol{x})] \\ \geq \frac{1}{4}\mathop{\mathbf{E}}_{x\sim D_x}\left[(1-2\eta(\boldsymbol{x}))\frac{e^{-|x_2|/\sigma}}{\sigma}\,|\boldsymbol{x}_1|\,\mathbb{1}_G(\boldsymbol{x})\right] - \mathop{\mathbf{E}}_{x\sim D_x}\left[\frac{e^{-|x_2|/\sigma}}{\sigma}\,|\boldsymbol{x}_1|\,\mathbb{1}_{G^c}(\boldsymbol{x})\right], \tag{2.17}$$

where we used the fact that the sigmoid $S_\sigma(|t|)^2$ is upper bounded by 1 and lower bounded by 1/4.

We can now bound each term using the fact that the distribution is (U, R)-bounded. Assume first that $\theta(\boldsymbol{w}^*, \boldsymbol{w}) = \theta \in (0, \pi/2)$. Then, (see Figure 2.2) we can express region G in polar coordinates as $G = \{(r, \phi) : \phi \in (0, \theta) \cup (\pi/2, \pi + \theta) \cup (3\pi/2, 2\pi)\}$. We denote by $\gamma(x, y)$ the density of the 2-dimensional projection on V of the marginal distribution $D_{\boldsymbol{x}}$. Since the integrand is non-negative we may bound from below the contribution of region G on the gradient by integrating over

$\phi \in (\pi/2, \pi)$.

$$
\begin{align}
&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[(1-2\eta(\boldsymbol{x}))\frac{e^{-|\boldsymbol{x}_2|/\sigma}}{\sigma}\;|\boldsymbol{x}_1|\;\mathbb{1}_G(\boldsymbol{x})\right] \tag{2.18}\\
&\geq \int_0^\infty \int_{\pi/2}^{\pi}(1-2\eta(\boldsymbol{x}))\gamma(r\cos\phi, r\sin\phi)r^2|\cos\phi|\frac{e^{-\frac{r\sin\phi}{\sigma}}}{\sigma}\mathrm{d}\phi\mathrm{d}r \tag{2.19}\\
&= \int_0^\infty \int_0^{\pi/2}(1-2\eta(\boldsymbol{x}))\gamma(r\cos\phi, r\sin\phi)r^2\cos\phi\frac{e^{-\frac{r\sin\phi}{\sigma}}}{\sigma}\mathrm{d}\phi\mathrm{d}r \notag\\
&\geq \int_{R/2}^R \int_0^{\pi/2} c|\boldsymbol{w}^*\cdot\boldsymbol{x}|\gamma(r\cos\phi, r\sin\phi)r^2\cos\phi\frac{e^{-\frac{r\sin\phi}{\sigma}}}{\sigma}\mathrm{d}\phi\mathrm{d}r \notag\\
&\geq c\frac{R}{6}\int_{R/2}^R \int_0^{\pi/2} \gamma(r\cos\phi, r\sin\phi)r^2\cos\phi\frac{e^{-\frac{r\sin\phi}{\sigma}}}{\sigma}\mathrm{d}\phi\mathrm{d}r \notag\\
&\geq c\frac{R}{6U}\int_{R/2}^R r^2\mathrm{d}r\int_0^{\pi/2}\cos\phi\frac{e^{-\frac{R\sin\phi}{\sigma}}}{\sigma}\mathrm{d}\phi \notag\\
&= c\frac{7}{144U}R^3\left(1-e^{-\frac{R}{\sigma}}\right) \geq c\frac{7}{144U}R^3\left(1-e^{-8}\right)\;, \tag{2.20}
\end{align}
$$

where for the third inequality we used that for $\|\boldsymbol{x}\|_2 \geq R/2$, we have that $\boldsymbol{w}^* \cdot \boldsymbol{x} = \frac{R}{2}(\cos(\theta)+\sin(\theta)) \geq R/6$, for the fourth inequality we used the lower bound $1/U$ on the density function $\gamma(r\cos\phi, r\sin\phi)$ (see Definition 2.2), and for the last inequality we used that $\sigma \leq R/8$.

We next bound from above the contribution of the gradient of region G^c. We have $G^c = \{(r,\phi) : \phi \in B_\theta = (\pi/2-\theta, \pi/2) \cup (3\pi/2-\theta, 3\pi/2)\}$

$$
\begin{align}
\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\frac{e^{-|\boldsymbol{x}_2|/\sigma}}{\sigma}\;|\boldsymbol{x}_1|\;\mathbb{1}_{G^c}(\boldsymbol{x})\right] &= \int_0^\infty\int_{\phi\in B_\theta}\gamma(r\cos\phi, r\sin\phi)r^2\cos\phi e^{-\frac{r\sin\phi}{\sigma}}\mathrm{d}\phi\mathrm{d}r \notag\\
&\leq \frac{2U}{\sigma}\int_0^\infty\int_\theta^{\pi/2} r^2\cos\phi e^{-\frac{r\sin\phi}{\sigma}}\mathrm{d}\phi\mathrm{d}r \notag\\
&= \frac{2U\sigma^2\cos^2\theta}{\sin^2\theta} = \frac{2R^3c\cos^2\theta}{24^2U}\;, \tag{2.21}
\end{align}
$$

where the inequality follows from the upper bound U on the density $\gamma(r\cos\phi, r\sin\phi)$ (see Definition 2.2), and the last equality follows from the value of σ. Combining

(2.20) and (2.21), we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V}\left[\frac{e^{-|\boldsymbol{x}_2|/\sigma}}{\sigma}\,|\boldsymbol{x}_1|\,\mathbb{1}_{G^c}(\boldsymbol{x})\right] &\le \frac{2R^3 c\cos^2\theta}{24^2 U}\\
&\le \frac{1}{8}\frac{7cR^3\left(1-e^{-8}\right)}{144U}\\
&\le \frac{1}{2}\frac{1}{4}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V}\left[\frac{e^{-|\boldsymbol{x}_2|/\sigma}}{\sigma}\,|\boldsymbol{x}_1|\,\mathbb{1}_{G}(\boldsymbol{x})\right]\,, \qquad (2.22)
\end{aligned}$$

where the second inequality follows from the identity $\cos^2\theta \le 1$ and $\frac{2}{24^2} \le \frac{1}{8}\frac{7\left(1-e^{-8}\right)}{144}$. Using (2.22) in (2.17), we obtain

$$\left\|\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D_V}\left[\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\widehat{\boldsymbol{w}})\right]\right\|_2 \ge \frac{1}{8}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim(D_{\boldsymbol{x}})_V}\left[\frac{e^{-|\boldsymbol{x}_2|/\sigma}}{\sigma}\,|\boldsymbol{x}_1|\,\mathbb{1}_{G}(\boldsymbol{x})\right] \ge \frac{cR^3}{288U}\,.$$

To conclude the proof, notice that the case where $\theta(\boldsymbol{w},\boldsymbol{w}^*) \in (\pi/2, \pi-\theta)$ follows by an analogous argument. Finally, in the case where $\theta = \pi/2$, the region G^c is empty and we can again get the same lower bound on the gradient norm.

□

Algorithm 3 Learning Halfspaces with Strong Massart Noise

```
procedure ALG($\epsilon, U, R, t(\cdot)$)
    $C_1 \leftarrow \Theta(U^{12}/R^{18})$.
    $C_2 \leftarrow \Theta(R^{3/2}/U^2)$.
    $T \leftarrow C_1\, d\, t(\epsilon/2)^8/(\epsilon^4 c^6)\log(1/\delta)$.        ▷ number of steps
    $\beta \leftarrow C_2^2\, d\, c^3\epsilon^2/(t(\epsilon/2)^4 T^{1/2})$.
    $\sigma \leftarrow C_2\, c^{1/2}\, \epsilon/t^2(\epsilon/2)$.
    $(\boldsymbol{w}^{(0)}, \boldsymbol{w}^{(1)}, \dots, \boldsymbol{w}^{(T)}) \leftarrow \mathrm{PSGD}(f, T, \beta)$.        ▷
  $f(\boldsymbol{w}) = \mathbf{E}_{(\boldsymbol{x},y)\sim D}\left[S_\sigma\Big(-y\frac{\boldsymbol{w}\cdot\boldsymbol{x}}{\|\boldsymbol{w}\|_2}\Big)\right]$, (1)
    $L \leftarrow \{\pm\boldsymbol{w}^{(i)}\}_{i\in[T]}$.        ▷ $L$: List of candinate vectors
    Draw $N = O(\log(T/\delta)/\epsilon^2)$ samples from $D$.
    $\bar{\boldsymbol{w}} \leftarrow \mathrm{argmin}_{\boldsymbol{w}\in L}\sum_{j=1}^{N} \mathbb{1}\{\mathrm{sign}(\boldsymbol{w}\cdot\boldsymbol{x}^{(j)}) \neq y^{(j)}\}$.
    return $\bar{\boldsymbol{w}}$.
```

Proof of Theorem 2.10. From Claim 2.1, we have that to make the $\mathrm{err}_{0-1}^{D_x}(h_{\bar{\boldsymbol{w}}}, f) \le \epsilon$ it suffices to prove that the angle $\theta(\bar{\boldsymbol{w}}, \boldsymbol{w}^*) \le O(\epsilon/(Ut^2(\epsilon/2))) =: \theta$. Using (the contrapositive of) Lemma 2.11 we get that with $\sigma \le \Theta(R/U\sqrt{cR\theta})$, if the norm squared of the gradient of some vector $\boldsymbol{w} \in \mathbb{S}^{d-1}$ is smaller than $\rho = O(R^3c/U)$, then $\boldsymbol{w}$ is close to either $\boldsymbol{w}^*$ or $-\boldsymbol{w}^*$, that is $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \le \theta$ or $\theta(\boldsymbol{w}, -\boldsymbol{w}^*) \le \theta$. Therefore, it suffices to find a point $\boldsymbol{w}$ with gradient $\|\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\boldsymbol{w})\|_2 \le \rho$.

From Lemma 2.8, we have that our PSGD objective function $\mathcal{L}_\sigma(\boldsymbol{w})$, is bounded by 1,

$$\mathbf{E}\left[\left\|\nabla_{\boldsymbol{w}} S_\sigma\Big(-y\frac{\boldsymbol{w}\cdot\boldsymbol{x}}{\|\boldsymbol{w}\|_2}\Big)\right\|_2^2\right] \le O(d/\sigma^2)\,,$$

$\left\|\mathbf{E}\left[\nabla_{\boldsymbol{w}} S_\sigma\Big(-y\frac{\boldsymbol{w}\cdot\boldsymbol{x}}{\|\boldsymbol{w}\|_2}\Big)\right]\right\|_2^2 \le O(1/\sigma^2)$, and that the gradient of $\mathcal{L}_\sigma(\boldsymbol{w})$ is Lipschitz with Lipschitz constant $O(1/\sigma^2)$. Using these bounds for the parameters of Lemma 2.7, we get that with $T = O(\frac{d}{\sigma^4\rho^4}\log(1/\delta))$ rounds, the norm of the gradient of some vector of the list $(\boldsymbol{w}^{(0)}, \dots, \boldsymbol{w}^{(T)})$ will be at most ρ with $1-\delta$

probability. Therefore, the required number of rounds is

$$T = O\left(\frac{U^{12}}{R^{18}}\frac{dt^8(\epsilon/2)\log(1/\delta)}{\epsilon^4 c^6}\right) \ .$$

Now that we know that one of the hypotheses in the list L (line 8 of Algorithm 3) is ϵ-close to the true $\boldsymbol{w}^*$, we can evaluate all of them on a small number of samples from the distribution D to obtain the best among them. The fact that $N = O(\log(T/\delta)/(\epsilon^2))$ samples are sufficient to guarantee that the excess error of the chosen hypothesis is at most ϵ with probability $1-\delta$ follows directly from Hoeffding's inequality. This completes the proof. □

3 LEARNING WITH TSYBAKOV NOISE IN QUASI-POLYNOMIAL TIME

3.1 Formal Statement of Results

Preliminaries

For a square matrix $\boldsymbol{M}$, we say that $\boldsymbol{M}$ is positive semi-definite if only if all the eigenvalues of $\boldsymbol{M}$ are non-negative. For $m \in \mathbb{Z}_+$, we denote $\mathcal{S}^m$ the set of symmetric matrices of dimension m. For an m-dimensional square matrix $\boldsymbol{A}$, let $\mathrm{tr}(\boldsymbol{A})$ be its trace.

Let $S = (s_1, s_2, \ldots, s_d)$ be a d-dimensional multi-index vector, where for all $i \in [d]$, s_i is non-negative integer. We denote $|S| = \sum_{i=1}^{d} s_i$ and for a d-dimensional vector $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d)$, we denote $\boldsymbol{w}^S = \prod_{i=1}^{d} \boldsymbol{w}_i^{s_i}$. For a degree-$k$ multivariate polynomial $p(\boldsymbol{x}) = \sum_{S:|S|\leq k} C_S \boldsymbol{x}^S$, let $\|p\|_2 :- \sqrt{\sum_{S:|S|\leq k} C_S^2}$ and $\|p\|_1 :- \sum_{S:|S|\leq k} |C_S|$. As we discussed in (Section 1.2), obtaining computationally efficient learning algorithms in the presence of Tsybakov noise in *any* non-trivial setting — that is, for any natural concept class and under any distributional assumptions — has been a long-standing open problem in learning theory. In this work, we make the first progress on this problem. Specifically, we give a learning algorithm for halfspaces that succeeds under a class of well-behaved distributions (including log-concave distributions) and runs in time *quasi-polynomial* in $1/\epsilon$. We start by describing the distribution family for which our algorithm succeeds. We remark that the following definition is a special case of the more general definition of bounded distributions Definition 2.2 (the difference is that now the tail must be sub-exponential).

Definition 3.1 (Sub-Exponential Bounded Distributions)**.** *For any set of parameters $L, R, B, \beta > 0$, an isotropic (i.e., zero mean and identity covariance) distribution $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ is called (L, R, B, β)*-bounded *if for any projection $(D_{\boldsymbol{x}})_V$ of $D_{\boldsymbol{x}}$ on a 2-dimensional* *subspace V, the corresponding pdf γ_V on $\mathbb{R}^2$ satisfies the following properties:*

1. *We have that $\gamma_V(\boldsymbol{x}) \geq L$, for all $\boldsymbol{x} \in V$ such that $\|\boldsymbol{x}\|_2 \leq R$ (anti-anti-concentration).*

2. *For any $t > 0$, we have that $\mathbf{Pr}_{\boldsymbol{x}\sim\gamma_V}[\|\boldsymbol{x}\|_2 \geq t] \leq B\exp(-\beta t)$ (concentration).*

Moreover, if there exists $U > 0$ such that for all $\boldsymbol{x} \in V$ we have that $\gamma_V(\boldsymbol{x}) \leq U$ (anti-concentration), then the distribution $D_{\boldsymbol{x}}$ is called (L, R, U, B, β)-bounded.

Definition 3.1 specifies the concentration and (anti-)anti-concentration properties on the underlying data distribution that are needed to prove the correctness of our algorithm. We note that the sample complexity and runtime of our algorithm depends on the values of these parameters.

For concreteness, we state a simplified version of our main result for the case that L, R, U, B, β are positive universal constants. We call such distributions *well-behaved*. We note that the class of well-behaved distributions is quite broad. In particular, it is easy to show (Fact 3.14) that every isotropic log-concave distribution is well-behaved. Moreover, the concentration and anti-concentration conditions of Definition 3.1 do not require a specific nonparametric constraint for the underlying density function, and are satisfied by many reasonable continuous distributions.

We show:

Theorem 3.2 (Learning Halfspaces with Tsybakov Noise). *Let $\mathcal{C}$ be the class of homogeneous halfspaces and $\mathcal{F}$ be a family of well-behaved distributions on $\mathbb{R}^d$. There is an algorithm with the following behavior: On input the error parameter $\epsilon > 0$ and oracle access to a Tsybakov example oracle $\mathrm{EX}^{\mathrm{Tsyb}}(f, \mathcal{F})$ with parameters (α, A), where $f \in \mathcal{C}$ is the target concept, the algorithm draws $N = d^{O\left((1/\alpha^2)\log^2(1/\epsilon)\right)}$ labeled examples, runs in $\mathrm{poly}(N, d)$ time, and computes a hypothesis $h \in \mathcal{C}$ that with high probability is ϵ-close to f.*

See Theorem 3.13 for a more detailed statement that takes into account the dependence on the parameters L, R, U, B, β.

3.2 Overview of Techniques

In this subsection, we give an intuitive description of our techniques that lead to Theorem 3.2 in tandem with a brief comparison to prior techniques and why they fail in our context.

It is instructive to begin by explaining where algorithms for the related problem of learning with Massart noise fall apart. The Massart noise model corresponds to the special case of Tsybakov noise where the label of each example $\boldsymbol{x}$ is independently flipped with probability $\eta(\boldsymbol{x}) \leq \eta$, where $\eta < 1/2$ is a parameter of the model. A line of work has developed efficient algorithms for learning halfspaces in this model, with the recent works Zhang et al. (2020a); Diakonikolas et al. (2020c) being the state-of-the-art. (See Section 3.5 for more details.)

We start by briefly describing the underlying idea behind several previous algorithms for learning halfspaces with Massart noise Zhang et al. (2020a); Diakonikolas et al. (2020c). These algorithms are typically iterative: In each iteration t, we have a current guess $\boldsymbol{w}$ for the normal vector $\boldsymbol{w}^*$ to the true halfspace, and our goal is to perform a local step to improve our guess (in expectation). To perform these updates, the algorithms aim to boost the contribution of the disagreement region A between the halfspaces corresponding to $\boldsymbol{w}$ and $\boldsymbol{w}^*$. This is achieved by considering points only around a small band around $\boldsymbol{w}$, i.e., all $\boldsymbol{x}$ with $|\boldsymbol{w} \cdot \boldsymbol{x}| < T$. This idea suffices to obtain efficient algorithms for the Massart noise model under well-behaved (e.g., log-concave) distributions as the total contribution of those points is amplified.

For the case of Tsybakov noise however, the situation is much more challenging. Even though the probability mass of the points in region A increases by restricting to a band around the current guess, it does not guarantee that the angle between $\boldsymbol{w}$ and $\boldsymbol{w}^*$ improves. This is because in the Tsybakov noise model, it is possible that all points in region A have flipping probabilities $\eta(\boldsymbol{x}) \approx 1/2$, which grow closer to $1/2$ the more the band shrinks. Thus, even though the conditional probability of region A increases with smaller band size T, the signal that these points provide to improve the angle may not be strong enough to overcome the effect that the remaining points have.

Our main idea to overcome this obstacle is to increase the contribution of points in region A by appropriately reweighting them (see Figure 3.1). A key observation that drives our algorithm (see Fact 3.3) is to find a weighting scheme that *certifies* whether a given guess $\boldsymbol{w}$ is (near-)optimal. In more detail, if there exists a non-negative weighting function $F(\boldsymbol{x})$ such that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[F(\boldsymbol{x})y \ \mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x})] < 0$, then the weight vector $\boldsymbol{w}$ is not optimal. Conversely, if $\boldsymbol{w}$ is not optimal, a weighting function F that makes the above expectation negative always exists (take for example the indicator of the disagreement region between $\boldsymbol{w}$ and $\boldsymbol{w}^*$).

Our first technical contribution is making the aforementioned certificate algorithmic. In more detail, we show that in order to certify that a guess $\boldsymbol{w}$ is ϵ-far from optimal, it suffices to consider weighting functions of a particular form, equal to the square of a multivariate polynomial restricted on a band close to $\boldsymbol{w}$. In particular, we show (Theorem 3.4) that it suffices to consider polynomials of degree at most $k = O(\log^2(1/\epsilon)/\alpha^2)$. We provide an explicit construction of such a multivariate polynomial with bounded coefficients, making critical use of Chebyshev polynomials.

Given this structural result, we can efficiently check the validity of a particular guess by searching all functions of the aforementioned form. Drawing sufficiently many samples so that all functions in the class converge uniformly, we can identify a good weighting (if one exists) by solving a semidefinite program to check the required condition over all squares of polynomials of degree-k. The sample complexity required to find our certificate is $d^{O(k)}$ and can be achieved in sample-polynomial time (Lemma 3.11).

We note that while our algorithm searches over multivariate polynomials that certify the error of our estimate, our approach differs significantly from other approaches for learning halfspaces by approximating them by polynomial threshold functions, like the L_1-regression algorithm of Kalai et al. (2008). Our use of polynomials is done in order to certify whether a candidate halfspace is sufficiently accurate, instead of searching a larger class of hypotheses. Remaining within the class of halfspaces allows us to use geometric properties of the underlying data distributions and the setting we consider, like the relationship of the misclassification error and the angle between the guess and the optimal halfspace. Additionally, while the L_1-regression can be written as a linear program, our approach requires searching over squares of polynomials and inherently relies on solving SDPs for obtaining a certificate.

Finally, turning the above algorithm for obtaining certificates into a learning algorithm is not immediate. To achieve this, we rely on online convex optimization with a similar approach to the one used in Zhang et al. (2020a). In contrast to an offline method like stochastic gradient descent, online convex optimization allows us to change the distribution of examples with which we penalize the guess, and the distribution is allowed to depend on the current guess. For every guess $\boldsymbol{w}$, we compute a loss function according to the reweighted distribution of points given by our certificate. We set up the objective so that any guess that is not close to optimal incurs a large loss, while the optimal guess always incurs a very small loss. By the guarantees of online convex optimization, after few iterations, the average loss of our guesses must be very close to the optimal loss. This means that one of the guesses must be near-optimal (see Lemma 3.18). This property will cause the certificate algorithm to accept this guess as close to optimal. A complication that arises in designing the loss function is that guessing 0 must give a large loss compared to the optimal, which we ensure by making the loss sufficiently negative at the optimal linear classifier.

3.3 Certifying Optimality in Quasi-Polynomial Time

We now describe our quasi-polynomial algorithm to test whether a given candidate hypothesis $\boldsymbol{w}$ is close to the optimal hypothesis $\boldsymbol{w}^*$. Our approach is based on the following observation.

Fact 3.3 (Certifying Function)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that: (a) For any pair of distinct unit vectors $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^d$, we have that $\mathbf{Pr}_{\boldsymbol{x} \sim D_{\boldsymbol{x}}}[h_{\boldsymbol{v}}(\boldsymbol{x}) \neq h_{\boldsymbol{u}}(\boldsymbol{x})] > 0$.*

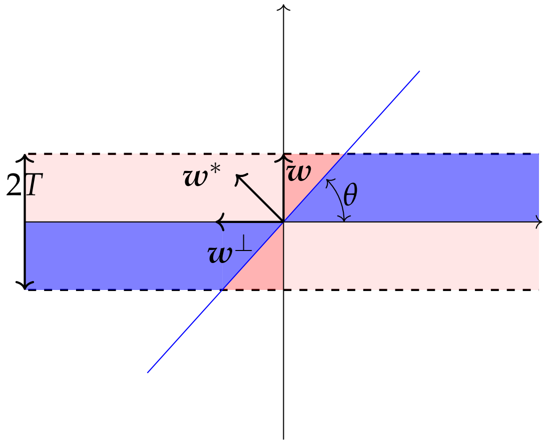


Figure 3.1: The disagreement region A ("blue") of the halfspaces $\boldsymbol{w}$ and $\boldsymbol{w}^*$. Our reweighting boosts points in region A: lower opacity means lower weight.

(b) D satisfies the Tsybakov noise condition with optimal classifier $f(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}^*, \boldsymbol{x}\rangle)$. Then we have:

1. *For any $T : \mathbb{R}^d \mapsto \mathbb{R}_+$, we have that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\, y\boldsymbol{w}^* \cdot \boldsymbol{x}] \geq 0$.*

2. *For any non-zero vector $\boldsymbol{w} \in \mathbb{R}^d$ such that $\theta(\boldsymbol{w}, \boldsymbol{w}^*) > 0$, there exists a function $T : \mathbb{R}^d \mapsto \mathbb{R}_+$ satisfying $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\, y\boldsymbol{w} \cdot \boldsymbol{x}] < 0$.*

Proof. For the first statement, note that

$$\begin{aligned}\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\, y\boldsymbol{w}^* \cdot \boldsymbol{x}] &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[T(\boldsymbol{x})|\boldsymbol{w}^* \cdot \boldsymbol{x}|(1-\eta(\boldsymbol{x}))] - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[T(\boldsymbol{x})|\boldsymbol{w}^* \cdot \boldsymbol{x}|\,\eta(\boldsymbol{x})] \\ &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[T(\boldsymbol{x})|\boldsymbol{w}^* \cdot \boldsymbol{x}|\,(1-2\eta(\boldsymbol{x}))] \geq 0\,,\end{aligned}$$

where we used the fact that $\eta(\boldsymbol{x}) \leq 1/2$ and $T(\boldsymbol{x}) \geq 0$.

For the second statement, let $\boldsymbol{w} \neq \mathbf{0}$ and $\theta(\boldsymbol{w}, \boldsymbol{w}^*) > 0$. By picking as a certifying function T the indicator function of the disagreement region between f and $h_{\boldsymbol{w}}$, i.e., $T(\boldsymbol{x}) :- \mathbb{1}\{h_{\boldsymbol{w}}(\boldsymbol{x}) \neq f(\boldsymbol{x})\}$, we have that

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\, y\boldsymbol{w} \cdot \boldsymbol{x}] = -\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[T(\boldsymbol{x})|\boldsymbol{w} \cdot \boldsymbol{x}|\,(1-2\eta(\boldsymbol{x}))]\ .$$

We claim that $\mathbf{E}_{x \sim D_x}[T(x)|w \cdot x| \,(1 - 2\eta(x))] > 0$, which proves the second statement. To see this, we use our assumption that the symmetric difference between any pair of distinct homogeneous halfspaces has positive probability mass. First, we note that from the Tsybakov condition (for any choice of parameters) we have that $\mathbf{Pr}_{x \sim D_x}[\eta(x) = 1/2] = 0$. So, it suffices to show that $\mathbf{E}_{x \sim D_x}[T(x)|w \cdot x|] > 0$.

Let w' be a non-zero vector such that the hyperplane $\{x : \langle w', x \rangle = 0\}$ is contained in the disagreement region $\{x : h_w(x) \neq f(x)\}$ and $\theta(w, w'), \theta(w^*, w') > 0$. This implies that $\{x : h_w(x) \neq f(x)\} \supset \{x : h_{w'}(x) \neq f(x)\}$ and $\mathbf{Pr}_{x \sim D_x}[h_{w'}(x) \neq f(x)] > 0$. Note that $|\langle w, x \rangle| > 0$ for all x with $h_{w'}(x) \neq f(x)$. Therefore, we get that

$$\mathop{\mathbf{E}}_{x \sim D_x}[T(x)|w \cdot x|] \geq \mathop{\mathbf{E}}_{x \sim D_x}[\mathbb{1}\{h_{w'}(x) \neq f(x)\}|w \cdot x|] > 0 \,.$$

This completes the proof of Fact 3.3. □

From Fact 3.3, we see that, given a hypothesis vector w that is not optimal, there exists a non-negative function that will make the expression of Item 2 of Fact 3.3 negative. One such function is $F(x) = \mathbb{1}\{\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)\}$, in which case we have $\mathbf{E}_{(x,y) \sim D}[F(x) w \cdot x y] = -\mathbf{E}_{x \sim D_x}[|w \cdot x|(1 - 2\eta(x))] < 0$. Since we cannot efficiently search over the space of all non-negative functions, we need to restrict our search space of certifying functions to some parametric class, ideally with a small number of parameters. In Section 3.3, we show that considering squares of low-degree polynomials suffices. In Section 3.3, we show that we can efficiently search in the space of (squares of) low-degree polynomials and find a certifying one.

Existence of a Low-Degree Polynomial Certificate

We start by showing that given a candidate hypothesis w that is "far" from being optimal, that is the angle $\theta(w, w^*)$ is bounded away from zero, we can construct a *low complexity* certificate F that will satisfy $\mathbf{E}_{(x,y) \sim D}[F(x) w \cdot x y] < 0$. In particular, we construct a certificate that is the product of a square of a low degree non-negative polynomial and an indicator function that depends on the hypothesis w.

This result is formally stated in the lemma bellow, which is the main result of this subsection.

Theorem 3.4 (Low-Degree Polynomial Certificate)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the Tsybakov noise condition with parameters (α, A) and the marginal $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ is (L, R, B, β)-bounded. Fix any $\theta \in (0, \pi/2]$. Let $\boldsymbol{w}^* \in \mathbb{S}^{d-1}$ be the normal vector to the optimal halfspace and $\widehat{\boldsymbol{w}} \in \mathbb{S}^{d-1}$ be such that $\theta(\widehat{\boldsymbol{w}}, \boldsymbol{w}^*) \geq \theta$. There exists polynomial $p : \mathbb{R}^d \mapsto \mathbb{R}$ of degree*

$$k = O\left(\frac{1}{\alpha^2 R \beta} \log^2\left(\frac{BA}{LR\theta}\right)\right)$$

satisfying $\|p\|_2^2 \leq d^{O(k)}$ such that

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}\left[p(\boldsymbol{x})^2\, \mathbb{1}\{0 \leq \boldsymbol{w}\cdot\boldsymbol{x} \leq \theta R/4\}\, y\boldsymbol{w}\cdot\boldsymbol{x}\right] \leq -\frac{\theta R}{4}\,.$$

We are going to use the following simple fact about Tsybakov noise that shows that large probability regions will also have large integral even if we weight the integral with the noise function $1 - 2\eta(\boldsymbol{x}) > 0$. Notice that larger noise $\eta(\boldsymbol{x})$ makes $1 - 2\eta(\boldsymbol{x})$ closer to 0, and therefore tends to reduce the probability mass of the regions where $\eta(\boldsymbol{x})$ is large. A similar lemma can be found in Tsybakov (2004). Since the definition of $\eta(\boldsymbol{x})$ is slightly different than ours, we provide the proof for completeness in Appendix B.1.

Lemma 3.5. *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the Tsybakov noise condition with parameters (α, A). Then for every measurable set $S \subseteq \mathbb{R}^d$ it holds $\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})(1 - 2\eta(\boldsymbol{x}))] \geq C_\alpha^A \left(\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})]\right)^{\frac{1}{\alpha}}$, where $C_\alpha^A = \alpha\left(\frac{1-\alpha}{A}\right)^{\frac{1-\alpha}{\alpha}}$.*

Using the lemma above, we can bound from below and above the $\mathrm{err}_{0-1}^{D}(h)$ with the $\mathrm{err}_{0-1}^{D_x}(h, f)$ between our current hypothesis h and the optimal f.

Corollary 3.6. *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the Tsybakov noise condition with parameters (α, A) and $f(\boldsymbol{x})$ be the optimal halfspace. Then for any halfspace*

$h(\boldsymbol{x})$, it holds

$$\mathop{\mathbf{Pr}}_{(\boldsymbol{x},y)\sim D}[h(\boldsymbol{x}) \neq y] \leq \mathop{\mathbf{Pr}}_{(\boldsymbol{x},y)\sim D}[f(\boldsymbol{x}) \neq y] + \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]$$

and

$$\mathop{\mathbf{Pr}}_{(\boldsymbol{x},y)\sim D}[h(\boldsymbol{x}) \neq y] \geq \mathop{\mathbf{Pr}}_{(\boldsymbol{x},y)\sim D}[f(\boldsymbol{x}) \neq y] + C_\alpha^A \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]^{\frac{1}{\alpha}} .$$

Proof. Let $S = \{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) \neq h(\boldsymbol{x})\}$ then

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{(\boldsymbol{x},y)\sim D}[h(\boldsymbol{x}) \neq y] &= \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[\mathbb{1}\{h(\boldsymbol{x}) \neq y\}] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}\{h(\boldsymbol{x}) \neq f(\boldsymbol{x})\}(1-\eta(\boldsymbol{x}))] + \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}\{h(\boldsymbol{x}) = f(\boldsymbol{x})\}\eta(\boldsymbol{x})] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}\{h(\boldsymbol{x}) \neq f(\boldsymbol{x})\}(1-2\eta(\boldsymbol{x}))] + \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\eta(\boldsymbol{x})] .
\end{aligned}$$

The first inequality follows from the fact that $1 - 2\eta(\boldsymbol{x}) \leq 1$ and the second one from Lemma 3.5. □

Central role in our construction play the Chebyshev polynomials. In the next fact, we collect the properties of Chebyshev polynomials that we are going to use in our argument, and we prove some of them in Appendix B.1.

Fact 3.7 (Chebyshev Polynomials Mason and Handscomb (2002))**.** *We denote by $T_k(t)$ the degree-k Chebyshev polynomial of the first kind. It holds*

$$T_k(t) = \begin{cases} \cos(k \arccos t) , & |t| \leq 1 \\ \frac{1}{2}\left(\left(t - \sqrt{t^2-1}\right)^k + \left(t + \sqrt{t^2-1}\right)^k\right) , & |t| \geq 1 . \end{cases}$$

Moreover, it holds $\|T_k\|_2^2 \leq 2^{6k+\log k+4}$.

Given a univariate polynomial $p(t)$, the following simple lemma bounds the blow-up of the square norm of the multivariate polynomial $q(\boldsymbol{x}) = p(\boldsymbol{w} \cdot \boldsymbol{x})$. We

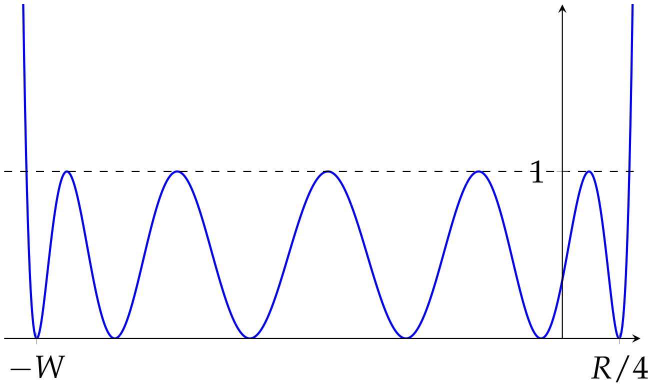


Figure 3.2: Plot of the polynomial $(T_k(g(t)))^2$ used in the proof of Theorem 3.4. Observe that this polynomial boosts the contribution of points in the blue region of Figure 3.1: points in A_2 have significantly boosted contribution because their density is lower bounded by some constant and the polynomial takes very large values in A_2, see Fact 3.9. In A_0, even though the polynomial has large value, the exponential tails of the distribution cancel the contribution of these points (given that W is sufficiently large).

also give a simple bound on the coefficient norm blow-up under shift of the argument of a univariate polynomial.

Lemma 3.8. *Let $p(t) = \sum_{i=0}^{k} c_i t^i$ be a degree-k univariate polynomial. Given $\boldsymbol{w} \in \mathbb{R}^d$ with $\|\boldsymbol{w}\|_2 \leq 1$, define the multivariate polynomial $q(\boldsymbol{x}) = p(\boldsymbol{w} \cdot \boldsymbol{x}) = \sum_{S:|S|\leq k} C_S \boldsymbol{x}^S$. Then we have that $\sum_{S:|S|\leq k} C_S^2 \leq d^{2k} \sum_{i=0}^{k} c_i^2$. Moreover, let $r(t) = p(at+b) = \sum_{i=0}^{k} d_i t^i$ for some $a, b \in \mathbb{R}$. Then $\|r\|_2^2 \leq (2\max(1,a)\max(1,b))^{2k} \|p\|_2^2$.*

The proof of this lemma is given in Appendix B.1. We can now proceed to the proof of the main technical theorem.

Proof of Theorem 3.4. Let V be the 2-dimensional subspace spanned by $\boldsymbol{w}^*$ and $\boldsymbol{w}$. To simplify notation, let θ be the angle between $\boldsymbol{w}^*$ and $\boldsymbol{w}$. First, we assume that $\theta \leq \pi/2$. Without loss of generality, assume $\boldsymbol{w} = \boldsymbol{e}_2$ and $\boldsymbol{w}^* = -a\boldsymbol{e}_1 + b\boldsymbol{e}_2$, where $\boldsymbol{e}_1, \boldsymbol{e}_2$ are the standard basis vectors of $\mathbb{R}^2$. For some parameter $W > 0$ to be specified later, we define the linear transformation

$$g(t) = 1 + 2\frac{t - R/4}{W + R/4}.$$

Set $p(\boldsymbol{x}) = T_k(g(\boldsymbol{x}_1))$, where T_k is the degree-k Chebyshev polynomial of Fact 3.7, and define the following partition of $\mathbb{R}^d$

$$A_0 = \{\boldsymbol{x} : \boldsymbol{x}_1 \in [-\infty, -W]\}, A_1 = \{\boldsymbol{x} : \boldsymbol{x}_1 \in [-W, R/4]\},$$
$$A_2 \qquad\qquad = \{\boldsymbol{x} : \boldsymbol{x}_1 \in [R/4, +\infty]\} .$$

We first investigate the behavior of $p(\boldsymbol{x})$ in each of these three regions.

Fact 3.9. *For the polynomial $p(\boldsymbol{x})$ defined above, the following properties hold in each region:*

1. *For all $\boldsymbol{x} \in A_0$, $p(\boldsymbol{x})^2 \le (2g(\boldsymbol{x}_1))^{2k}$.*
2. *For all $\boldsymbol{x} \in A_1$, $p(\boldsymbol{x})^2 \le 1$.*
3. *For all $\boldsymbol{x}$ such that $\boldsymbol{x}_1 \ge R/2$, it holds that $p(\boldsymbol{x})^2 \ge \frac{1}{2}\left(1 + \sqrt{\frac{R}{2W+R/2}}\right)^{2k}$.*

Proof. By Fact 3.7, for the univariate Chebyshev polynomials of degree-k, we know that for all $t \le -1$ it holds

$$|T_k(t)| = \left|\frac{1}{2}((t - \sqrt{t^2-1})^k + (t + \sqrt{t^2-1})^k)\right| \le (2t)^k .$$

Observe that for all $\boldsymbol{x} \in A_0$, we have $g(\boldsymbol{x}_1) \le -1$, thus $p(\boldsymbol{x})^2 \le (2g(\boldsymbol{x}_1))^{2k}$. For all $\boldsymbol{x} \in A_1$, we have $-1 \le g(\boldsymbol{x}_1) \le 1$, which leads to $p(\boldsymbol{x})^2 \le 1$.

Finally, from the definition of the Chebyshev polynomial T_k (Fact 3.7), we have that for all $t \ge 0$ it holds

$$T_k(1+t) \ge \frac{1}{2}(1 + t + \sqrt{t^2 + 2t})^k \ge \frac{1}{2}(1 + \sqrt{t})^k.$$

Moreover, all the roots of $T_k(t)$ lie in the interval $[-1,1]$ and hence, for $t \geq 0$, the polynomial $(T_k(1+t))^2$ is increasing in t. Therefore, for any $\boldsymbol{x}$ with $\boldsymbol{x}_1 \geq R/2$ it holds that

$$p(\boldsymbol{x})^2 = T_k(g(\boldsymbol{x}_1)) \geq T_k(g(R/2)) \geq \frac{1}{2}\left(1+\sqrt{\frac{R}{2W+R/2}}\right)^{2k} .$$

□

We bound the expectation $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[p(\boldsymbol{x})^2 \boldsymbol{w}\cdot\boldsymbol{x} y \operatorname{sign}(\boldsymbol{w}\cdot\boldsymbol{x})\mathbb{1}\{0 \leq \boldsymbol{w}\cdot\boldsymbol{x} \leq \frac{\theta R}{4}\}]$ in each of the three regions separately. We start from A_0, where we have

$$\begin{aligned}
I_0 &= \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[p(\boldsymbol{x})^2 \boldsymbol{w}\cdot\boldsymbol{x}\, y\, \mathbb{1}\{\boldsymbol{w}\cdot\boldsymbol{x} \in [0,\theta R/4]\}\, \mathbb{1}_{A_0}(\boldsymbol{x})] \\
&= \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[T_k(g(\boldsymbol{x}_1))^2\, \boldsymbol{x}_2 y\, \mathbb{1}\{\boldsymbol{x}_2 \in [0,\theta R/4]\}\, \mathbb{1}\{\boldsymbol{x}_1 \leq -W\}] \\
&\leq \frac{\theta R}{4} \mathop{\mathbf{E}}_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\sim D_V}[(2g(\boldsymbol{x}_1))^{2k}\mathbb{1}\{\boldsymbol{x}_1 \leq -W\}] ,
\end{aligned}$$

where to get the last inequality we used that $\boldsymbol{x}_2\mathbb{1}[\boldsymbol{x}_2 \in [0,\theta R/4] \leq \theta R/4$ and Item 1 of Fact 3.9. Using the fact that for any real random variable X it holds $\mathbf{E}[|X|^m] = \int_0^\infty m t^{m-1}\,\mathbf{Pr}[|X| \geq t]\mathrm{d}t$ and the exponential concentration of D_V (see Definition 3.1), we obtain

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\sim D_V}[g(\boldsymbol{x}_1)^{2k}\mathbb{1}\{\boldsymbol{x}_1 \leq -W\}] \\
&\qquad = \int_0^\infty 2kt^{2k-1} \mathop{\mathbf{Pr}}_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\sim D_V}[|g(\boldsymbol{x}_1)\mathbb{1}\{\boldsymbol{x}_1 \leq -W\}| \leq t]\mathrm{d}t \\
&\qquad = \int_0^1 2kt^{2k-1}e^{-\beta W}\mathrm{d}t \;+\; \int_1^\infty 2kt^{2k-1}e^{-\beta\frac{t+1}{2}\left(W+\frac{R}{4}\right)+\beta\frac{R}{4}}\mathrm{d}t .
\end{aligned}$$

We observe that for all $t > 1, R > 0, W > 0$ it holds

$$\frac{t+1}{2}\left(W+\frac{R}{4}\right) - \frac{R}{4} \geq \frac{tW}{2} .$$

Therefore,

$$\begin{aligned}\int_1^\infty 2kt^{2k-1}e^{-\beta\frac{t+1}{2}\left(W+\frac{R}{4}\right)+\beta\frac{R}{4}}\mathrm{d}t &\le \int_1^\infty 2kt^{2k-1}e^{-t\beta W/2}\mathrm{d}t\\ &\le \int_0^\infty 2kt^{2k-1}e^{-t\beta W/2}\mathrm{d}t \le \left(\frac{W\beta}{2}\right)^{-2k}(2k)!\,.\end{aligned}$$

Combining the above inequalities we obtain

$$\begin{aligned}I_0 &\le \frac{\theta RB2^{2k}}{4}\left(\int_0^1 2kt^{2k-1}e^{-\beta W}\mathrm{d}t\ +\ \int_1^\infty 2kt^{2k-1}e^{-\beta\frac{t+1}{2}\left(W+\frac{R}{4}\right)+\beta\frac{R}{4}}\mathrm{d}t\right)\\ &= \frac{\theta RB2^{2k}}{4}\left(e^{-\beta W}+\ (W\beta/2)^{-2k}(2k)!\right)\,.\end{aligned}$$

We now set $W=8k/\beta$ and get

$$I_0 \le \frac{\theta RB}{4}(2^{2k}e^{-8k}+(2k)^{-2k}(2k)!) \le \frac{\theta RB}{4}(e^{-6k}+e^{-2k+1}\sqrt{2k}) \le \frac{\theta RB}{4}\,,$$

where we used Stirling's approximation, i.e., $(2k)! \le e\sqrt{2k}e^{-2k}(2k)^{2k}$, and the fact that $e^{-6k}+e^{-2k+1}\sqrt{2k}\le 1$, for all $k\ge 1$.

Bounding the contribution of region A_1 is quite simple. Using from Fact 3.9, that $p(\boldsymbol{x})^2\le 1$ for all $\boldsymbol{x}\in A_1$, we obtain

$$I_1 = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[p(\boldsymbol{x})^2\boldsymbol{w}\cdot\boldsymbol{x}\,y\,\mathbb{1}\{\boldsymbol{w}\cdot\boldsymbol{x}\in[0,\theta R/4]\}\,\mathbb{1}_{A_1}(\boldsymbol{x})] \le \frac{\theta R}{4}.$$

We finally bound the contribution of region A_2. We have

$$\begin{aligned}I_2 &= \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[p(\boldsymbol{x})^2\boldsymbol{w}\cdot\boldsymbol{x}\,y\,\mathbb{1}\{\boldsymbol{w}\cdot\boldsymbol{x}\in[0,\theta R/4]\}\,\mathbb{1}_{A_2}(\boldsymbol{x})]\\ &= -\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[p(\boldsymbol{x})^2\boldsymbol{w}\cdot\boldsymbol{x}\,(1-2\eta(\boldsymbol{x}))\,\mathbb{1}\{\boldsymbol{w}\cdot\boldsymbol{x}\in[0,\theta R/4]\}\,\mathbb{1}_{A_2}(\boldsymbol{x})]\\ &\le -\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[p(\boldsymbol{x})^2\boldsymbol{w}\cdot\boldsymbol{x}\,(1-2\eta(\boldsymbol{x}))\,\mathbb{1}\{\boldsymbol{w}\cdot\boldsymbol{x}\in[\theta R/8,\theta R/4]\}\,\mathbb{1}\{\boldsymbol{x}_1\ge R/2\}]\\ &\le -\frac{\theta R}{8}T_k(g(R/2))^2\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\,(1-2\eta(\boldsymbol{x}))\,\mathbb{1}\{\boldsymbol{w}\cdot\boldsymbol{x}\in[\theta R/8,\theta R/4]\}\,\mathbb{1}\{\boldsymbol{x}_1\ge R/2\}]\,,\end{aligned}$$

where we used Item 3 of Fact 3.9. Using Lemma 3.5, we obtain that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\,(1-2\eta(\boldsymbol{x}))\,\mathbb{1}\{\boldsymbol{w}\cdot\boldsymbol{x}\in[\theta R/8,\theta R/4]\}\,\mathbb{1}\{\boldsymbol{x}_1\geq R/2\}]\geq C_\alpha^A(L\theta R/16)^{1/\alpha}\,.$$

From Item 3 of Fact 3.9, we obtain

$$\begin{aligned} I_2 &\leq -C_\alpha^A T_k(g(R/2))^2\frac{\theta R}{8}\left(L\frac{\theta R^2}{16}\right)^{1/\alpha} \\ &\leq -\frac{\theta R}{4}(B+2)\frac{C_\alpha^A}{2(B+2)}\left(1+\sqrt{\frac{R}{2W+R/2}}\right)^{2k}\left(L\frac{\theta R^2}{16}\right)^{\frac{1}{\alpha}}\,. \end{aligned}$$

Using the inequality $1+t\geq e^{t/2}$ for all $t\leq 2$, we obtain that in order to prove that $I_0+I_1+I_2\leq -\theta R/4$, it suffices to pick the degree k so that

$$\frac{C_\alpha^A}{2(B+2)}e^{\sqrt{\frac{Rk^2}{2W+R/2}}}\left(L\frac{\theta R^2}{16}\right)^{\frac{1}{\alpha}}\geq 1.$$

By our choice of $W=8k/\beta$, it follows that setting the degree of the polynomial to

$$k=O\left(\frac{1}{\alpha^2 R\beta}\log^2\left(\frac{BA}{LR\theta}\right)\right)$$

suffices. To complete the proof, we need to provide an upper bound on the magnitude of the coefficients of the polynomial p. From Fact 3.7, we have that $\|T_k(x)\|_2^2\leq 2^{6k+2\log k+4}$. Using Lemma 3.8, we obtain that $\|T_k(g(x))\|_2^2\leq 2^{2k}\cdot 2^{6k+2\log k+4}=2^{8k+2\log k+4}$. Moreover, from the Lemma 3.8, we can derive an upper bound on the square norm of the multivariate polynomial p, which is $\|p\|_2^2\leq d^{2k}2^{8k+2\log k+4}=d^{O(k)}$.

Moreover, for the case where $\pi\geq\theta>\pi/2$, we can prove with the same argument that

$$\mathop{\mathbf{E}}_{(x,y)\sim D}\left[p(\boldsymbol{x})^2\,\mathbb{1}\{0\leq\boldsymbol{w}\cdot\boldsymbol{x}\leq\pi R/8\}\,y\boldsymbol{w}\cdot\boldsymbol{x}\right]\leq-\frac{\pi R}{8}\,.$$

This follows from the fact that the expectation over the partitions A_0 and A_1 are at most their values for the case of $\theta = \pi/2$, and the expectation over A_2 is the same.

□

Efficiently Computing the Certificate

In this section, we show that we can efficiently compute our polynomial certificate given labeled examples from the target distribution. For the rest of this section, let $Q = d^{\Theta(k)}$ and let $\mathbb{1}_B(\boldsymbol{x})$ be the indicator function of the region $B = \{\boldsymbol{x} : 0 \leq \boldsymbol{w} \cdot \boldsymbol{x} \leq \theta R/4\}$. Denote by $\boldsymbol{m}(\boldsymbol{x})$ the vector containing all monomials up to degree k, such that $\boldsymbol{m}_S(\boldsymbol{x}) :- \boldsymbol{x}^S$, indexed by the multi-index S satisfying $|S| \leq k$. The dimension of $\boldsymbol{m}(\boldsymbol{x}) \in \mathbb{R}^m$ is $m = \binom{d+k}{k}$. For a real matrix $\boldsymbol{A} \in \mathbb{R}^{m \times m}$, we define the following function

$$\mathcal{L}_w(\boldsymbol{A}) = \mathop{\mathbf{E}}_{(x,y)\sim D}\left[\boldsymbol{m}(\boldsymbol{x})^T \boldsymbol{A}\, \boldsymbol{m}(\boldsymbol{x}) \mathbb{1}_B(\boldsymbol{x}) \boldsymbol{w} \cdot \boldsymbol{x} y\right] = \operatorname{tr}\left(\boldsymbol{A}\boldsymbol{M}\right) \;, \tag{3.1}$$

where $\boldsymbol{M} = \mathbf{E}_{(x,y)\sim D}\left[\boldsymbol{m}(\boldsymbol{x})\boldsymbol{m}(\boldsymbol{x})^T \mathbb{1}_B(\boldsymbol{x}) \boldsymbol{w} \cdot \boldsymbol{x} y\right]$. Notice that $\mathcal{L}_w$ is linear in its variable $\boldsymbol{A}$. From the discussion of the previous subsection, and in particular from Theorem 3.4, we know that if $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \geq \theta$, then there exists a polynomial $p(\boldsymbol{x})$ and a vector $\boldsymbol{b}$ of coefficients such that $p(\boldsymbol{x}) = \boldsymbol{b} \cdot \boldsymbol{m}(\boldsymbol{x})$ and $\mathcal{L}_w(\boldsymbol{b}\boldsymbol{b}^T) \leq -\theta R/4$. It follows that there exists a positive semi-definite rank-1 matrix $\boldsymbol{B} = \boldsymbol{b}\boldsymbol{b}^T$ such that $\mathcal{L}_w(\boldsymbol{B}) \leq -\theta R/4$. Moreover, we have that $\left\|p^2(\boldsymbol{x})\right\|_2^2 \leq Q$, which translates to $\|\boldsymbol{B}\|_F^2 \leq Q$. Therefore, we can formulate the following semi-definite program, which is feasible when $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \geq \theta$.

$$\begin{aligned} \operatorname{tr}(\boldsymbol{A}\boldsymbol{M}) &\leq -\theta R/4 \\ \|\boldsymbol{A}\|_F^2 &\leq Q \\ \boldsymbol{A} &\succeq 0 \end{aligned} \tag{3.2}$$

We define $\widetilde{\boldsymbol{M}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}(\boldsymbol{x}^{(i)})\boldsymbol{m}(\boldsymbol{x}^{(i)})^T \mathbb{1}_B(\boldsymbol{x}^{(i)}) y^{(i)} \boldsymbol{w} \cdot \boldsymbol{x}^{(i)}$, the empirical estimate of $\boldsymbol{M}$ using N samples from D. We can now replace the matrix $\boldsymbol{M}$ in Equation (3.1) with the estimate $\widetilde{\boldsymbol{M}}$ and define the following “empirical" SDP

$$\begin{aligned}\operatorname{tr}(\boldsymbol{A}\widetilde{\boldsymbol{M}}) &\leq -\frac{3\theta R}{16} \\ \|\boldsymbol{A}\|_F^2 &\leq Q \\ \boldsymbol{A} &\succeq 0\end{aligned} \tag{3.3}$$

In the following lemma, we bound the sample size required so that $\widetilde{\boldsymbol{M}}$ is sufficiently close to $\boldsymbol{M}$.

Lemma 3.10 (Estimation of $\boldsymbol{M}$)**.** *Let $\Omega = \{\boldsymbol{A} \in \mathcal{S}^m : \boldsymbol{A} \succeq 0, \|\boldsymbol{A}\|_F \leq Q\}$. There exists an algorithm that draws*

$$N = O\left(\frac{BQ^2}{\epsilon^2}\frac{(d+k)^{3k+2}}{(\beta/2)^{2k}}\log(1/\delta)\right)$$

samples from D, runs in $\operatorname{poly}(N, d)$ *time and with probability at least $1-\delta$ outputs a matrix $\widetilde{\boldsymbol{M}}$ such that*

$$\mathbf{Pr}\left[\sup_{\boldsymbol{A}\in\Omega}\left|\operatorname{tr}(\boldsymbol{A}\widetilde{\boldsymbol{M}}) - \operatorname{tr}(\boldsymbol{A}\boldsymbol{M})\right| \geq \epsilon\right] \leq 1-\delta\,.$$

Proof. Recall that $\widetilde{\boldsymbol{M}}$ is the empirical estimate of $\boldsymbol{M}$, that is $\boldsymbol{M} = \mathbf{E}_{(x,y)\sim D}[\boldsymbol{m}(\boldsymbol{x})\boldsymbol{m}(\boldsymbol{x})^T\mathbb{1}_B(\boldsymbol{x})y\boldsymbol{w}\cdot \boldsymbol{x}]$

$$\widetilde{\boldsymbol{M}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{m}(\boldsymbol{x}^{(i)})\boldsymbol{m}(\boldsymbol{x}^{(i)})^T\mathbb{1}_B(\boldsymbol{x}^{(i)})y^{(i)}\boldsymbol{w}\cdot\boldsymbol{x}^{(i)}\,. \tag{3.4}$$

Using the Cauchy-Schwarz inequality, we get

$$\operatorname{tr}\left(\boldsymbol{A}(\boldsymbol{M}-\widetilde{\boldsymbol{M}})\right) \leq \|\boldsymbol{A}\|_F\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|_F\,.$$

Therefore, it suffices to bound the probability that $\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|_F \geq \epsilon/Q$. From

Markov's inequality, we have

$$\mathbf{Pr}\left[\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|_F \geq \epsilon/Q\right] \leq \frac{Q^2}{\epsilon^2}\,\mathbf{E}\left[\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|_F^2\right]. \tag{3.5}$$

Using multi-indices S_1, S_2 that correspond to the monomials $\boldsymbol{x}^{S_1}, \boldsymbol{x}^{S_2}$ (as indices of the matrix $\boldsymbol{M}$), we have

$$\mathbf{E}\left[\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|_F^2\right] = \sum_{S_1,S_2:|S_1|,|S_2|\leq k} (\boldsymbol{M}_{S_1,S_2} - \widetilde{\boldsymbol{M}}_{S_1,S_2})^2 = \sum_{S_1,S_2:|S_1|,|S_2|\leq k} \mathbf{Var}[\widetilde{\boldsymbol{M}}_{S_1,S_2}].$$

Using the fact that the samples $(\boldsymbol{x}^{(i)}, y^{(i)})$ are independent, we can bound from above the variance of each entry (S_1, S_2) of $\widetilde{\boldsymbol{M}}$

$$\begin{aligned}\mathbf{Var}[\widetilde{\boldsymbol{M}}_{S_1,S_2}] &\leq \frac{1}{N}\mathop{\mathbf{E}}_{(x,y)\sim D}\left[\boldsymbol{x}^{2(S_1+S_2)}\left(\mathbb{1}_B(\boldsymbol{x})\boldsymbol{w}\cdot \boldsymbol{x}y\right)^2\right] \\ &\leq \frac{1}{N}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\boldsymbol{x}^{2(S_1+S_2)}\,\|\boldsymbol{x}\|_2^2\right] \\ &\leq \frac{1}{N}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[(\|\boldsymbol{x}\|_2^2)^{|S_1+S_2|+1}\right].\end{aligned}$$

To bound the higher-order moments, we are going to use the (two-dimensional) exponential tails of $D_{\boldsymbol{x}}$ of Definition 3.1. For all $t \geq t_0$, it holds

$$\mathbf{Pr}[\|\boldsymbol{x}\|_2 \geq t] = \mathbf{Pr}[\|\boldsymbol{x}\|_2^2 \geq t^2] \leq \sum_{i=1}^{d}\mathbf{Pr}\left[|\boldsymbol{x}_i|^2 \geq \frac{t^2}{d}\right] \leq Bde^{-\beta t/\sqrt{d}},$$

where β, B are the parameters of Definition 3.1. For every $\ell \geq 1$, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[(\|\boldsymbol{x}\|_2^2)^{\ell}\right] = \int_{t=0}^{\infty} 2\ell t^{2\ell-1}\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\|\boldsymbol{x}\|_2 \geq t]\mathrm{d}t \leq Bd^{\ell+1}\beta^{-2\ell}(2\ell)!\,.$$

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

$$\begin{aligned}\mathbf{E}\left[\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|_F^2\right] &\leq \frac{1}{N} B d^{k+1} \beta^{-2k} (2k)!\, m^2 = \frac{1}{N} B d^{k+1} \beta^{-2k} (2k)! \binom{d+k}{k}^2 \\ &\leq \frac{1}{N} B (\beta/2)^{-2k} (d+k)^{3k+1}\,, \end{aligned} \tag{3.6}$$

where we used the inequality $(2n)!/(n!)^2 \leq 4^n$. Combining Equations (3.5) and (3.6) we obtain that with $N \geq BQ^2(\beta/2)^{-2k}(d+k)^{3k+1}/(4\epsilon^2)$ samples we can estimate $\boldsymbol{M}$ 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 ℓ times to obtain estimates $\widetilde{\boldsymbol{M}}^{(1)}, \ldots, \widetilde{\boldsymbol{M}}^{(\ell)}$ and keep the coordinate-wise median as our final estimate. It follows that $\ell = O(\log(m/\delta))$ repetitions suffice to guarantee confidence probability at least $1-\delta$.

□

The following is the main lemma of this subsection, where we bound the number of samples and the runtime needed to construct the certificate given samples from the distribution D.

Lemma 3.11. *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the Tsybakov noise condition with parameters (α, A) and the marginal $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ is (L, R, B, β)-bounded. Let $\boldsymbol{w}^* \in \mathbb{S}^{d-1}$ be the normal vector to the optimal halfspace and $\boldsymbol{w} \in \mathbb{S}^{d-1}$. Fix any $\theta \in (0, \pi/2]$ and assume that $\theta(\boldsymbol{w}^*, \boldsymbol{w}) \geq \theta$. Let*

$$k = O\left(\frac{1}{\alpha^2 R \beta} \log^2\left(\frac{BA}{LR\theta}\right)\right),$$

and $Q = d^{\Theta(k)}$. There exists an algorithm that draws $N = d^{O(k)} \log(1/\delta)$ samples from D, runs in time $\mathrm{poly}(N, d)$, and with probability $1-\delta$ returns a positive semi-definite matrix $\boldsymbol{A}$ such that $\|\boldsymbol{A}\|_F^2 \leq Q$ and $\mathrm{tr}(\boldsymbol{A}\boldsymbol{M}) \leq -\theta R/16$.

Proof. From Lemma 3.10, we obtain that with N samples we can get a matrix $\widetilde{\boldsymbol{M}}$ such that $|\mathrm{tr}(\boldsymbol{A}\widetilde{\boldsymbol{M}} - \mathrm{tr}(\boldsymbol{A}\boldsymbol{M})| \leq \theta R/16$ with probability at least $1-\delta$. From

Theorem 3.4, we know that with the given bound for k and $\|\boldsymbol{A}\|_F$, there exists $\boldsymbol{A}^*$ such that

$$\operatorname{tr}(\boldsymbol{A}^*\boldsymbol{M}) \leq -\theta R/4.$$

Therefore, the SDP (3.2) is feasible. Moreover, from Lemma 3.10 we get that

$$\operatorname{tr}(\boldsymbol{A}^*\widetilde{\boldsymbol{M}}) \leq -\theta R/4 + \theta R/16 \leq -\frac{3\theta R}{16} .$$

Thus, the following SDP is also feasible

$$\begin{aligned} \operatorname{tr}(\boldsymbol{A}\widetilde{\boldsymbol{M}}) &\leq -\frac{3\theta R}{16} \\ \|\boldsymbol{A}\|_F^2 &\leq Q \\ \boldsymbol{A} &\succeq 0 \end{aligned} \tag{3.7}$$

Since the dimension of the matrix $\boldsymbol{A}$ 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 using tolerance $\theta R/16$, we obtain an almost feasible $\widetilde{\boldsymbol{A}}$, in the sense that $\operatorname{tr}(\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{M}}) \leq -3\theta R/16 + \theta R/16 = -\theta R/8$. Using again the guarantee of Lemma 3.10, we get that solving the SDP (3.7), we obtain a positive-semi definite matrix $\widetilde{\boldsymbol{A}}$ such that $\operatorname{tr}(\widetilde{\boldsymbol{A}}\boldsymbol{M}) \leq -\theta R/8 + \theta R/16 = -\theta R/16$.

□

3.4 Learning the Optimal Halfspace via Online Gradient Descent

In this section, we give a quasi-polynomial time algorithm that can learn a unit vector $\widehat{\boldsymbol{w}}$ with small angle from the normal vector of the optimal halfspace $\boldsymbol{w}^*$. Our main result of this section is the following theorem.

Theorem 3.12 (Parameter Estimation under (L, R, B, β)-bounded distributions)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the Tsybakov noise condition with* *parameters* (α, A) *and the marginal* $D_{\mathbf{x}}$ *on* $\mathbb{R}^d$ *is* (L, R, B, β)*-bounded. Moreover, let* $\mathbf{w}^* \in \mathbb{S}^{d-1}$ *be the normal vector to the optimal halfspace. There exists an algorithm that draws* $N = d^{O(k)} \log(1/\delta)$ *examples from* D *where* $k = O\left(\frac{1}{\alpha^2 R\beta} \log^2\left(\frac{BA}{\epsilon LR}\right)\right)$*, runs in* $\mathrm{poly}(N, d)$ *time, and computes a vector* $\widehat{\mathbf{w}}$ *such that* $\theta(\widehat{\mathbf{w}}, \mathbf{w}^*) \leq \epsilon$*, with probability* $1 - \delta$*.*

Note here that we do not need the U bounded assumption for Theorem 3.12. This corresponds to an anti-concentration assumption. If we have this additional property, we immediately get Theorem 3.13, which is the main result of this paper. Specifically, with this additional structure on the distribution, one can translate the small angle guarantee of Theorem 3.12 to the zero-one loss of the hypothesis that our algorithm outputs.

Theorem 3.13 (PAC-Learning under (L, R, U, B, β)-bounded distributions)**.** *Let* D *be a distribution on* $\mathbb{R}^d \times \{\pm 1\}$ *that satisfies the Tsybakov noise condition with parameters* (α, A) *and the marginal* $D_{\mathbf{x}}$ *on* $\mathbb{R}^d$ *is* (L, R, U, B, β)*-bounded. Moreover, let* $\mathbf{w}^* \in \mathbb{S}^{d-1}$ *be the normal vector to the optimal halfspace. There exists an algorithm that draws* $N = d^{O(k)} \log(1/\delta)$ *examples from* D *where* $k = O\left(\frac{1}{\alpha^2 R\beta} \log^2\left(\frac{B\,UA}{\epsilon LR\beta}\right)\right)$*, runs in* $\mathrm{poly}(N, d)$ *time, and computes a vector* $\widehat{\mathbf{w}}$ *such that* $\mathrm{err}_{0-1}^{D_x}(h_{\widehat{\mathbf{w}}}, f) \leq \epsilon$*, with probability* $1 - \delta$*, where* f *is the target halfspace.*

A corollary of the above theorem is that we can PAC learn halfspaces when the marginal distribution $D_{\mathbf{x}}$ is log-concave. The following known fact (see, e.g., Fact A.4 of Diakonikolas et al. (2020c)) shows that the family of log-concave distributions is indeed (L, R, U, B, β)-bounded for constant values of the parameters.

Fact 3.14. *An isotropic log-concave distribution on* $\mathbb{R}^d$ *is* $(2^{-12}, 1/9, e2^{17}, c, 1)$*-bounded, where* c *is an absolute constant.*

From Thereom 3.13 and Fact 3.14, we obtain the following corollary.

Corollary 3.15 (PAC-Learning under Isotropic Log-Concave Distributions)**.** *Let* D *be a distribution on* $\mathbb{R}^d \times \{\pm 1\}$ *that satisfies the Tsybakov noise condition with parameters*

(α, A) and the marginal $D_{\boldsymbol{x}}$ is an isotropic log-concave distribution. There exists an algorithm that draws $N = d^{O(k)} \log(1/\delta)$ examples from D where $k = O\left(\frac{1}{\alpha^2} \log^2(A/\epsilon)\right)$, runs in $\mathrm{poly}(N, d)$ *time, and computes a vector $\widehat{\boldsymbol{w}}$ such that $\mathrm{err}_{0-1}^{D_x}(h_{\widehat{\boldsymbol{w}}}, f) \le \epsilon$, with probability $1-\delta$, where f is the target halfspace.*

We now provide a high-level sketch of the proof of Theorem 3.12 for constant values of the parameters L, R, B, and β. For every candidate halfspace $\boldsymbol{w}$, that has angle greater than ϵ with the optimal hypothesis vector $\boldsymbol{w}^*$, our main structural result, Theorem 3.4, guarantees that there exists a polynomial p of degree $k = O((\log(1/\epsilon)/\alpha)^2)$ such that

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[p^2(\boldsymbol{x})\mathbb{1}_B(\boldsymbol{x})\boldsymbol{w}\cdot\boldsymbol{x}y] \le -\Omega(\epsilon)\ .$$

Moreover, from Lemma 3.10, we get that, given a candidate $\boldsymbol{w}$, we can compute a witnessing polynomial p in time $d^{O(k)}$. The next step is to use the certificate to improve the candidate $\boldsymbol{w}$. We are going to use Online Projected Gradient Decent (OPGD) to do this.

Lemma 3.16 (see, e.g., Theorem 3.1 of Hazan (2016))**.** *Let $\mathcal{V} \subseteq \mathbb{R}^n$ a non-empty closed convex set with diameter K. Let $\ell_1, \ldots, \ell_T$ be a sequence of T convex functions $\ell_t : \mathcal{V} \mapsto \mathbb{R}$ differentiable in open sets containing $\mathcal{V}$, and let $G = \max_{t\in[T]} \|\nabla_{\boldsymbol{w}}\ell_t\|_2$. Pick any $\boldsymbol{w}_1 \in \mathcal{V}$ and set $\eta_t = \frac{K}{G\sqrt{t}}$ for $t \in [T]$. Then, for all $\boldsymbol{u} \in \mathcal{V}$, we have that*

$$\sum_{t=1}^{T}(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u})) \le \frac{3}{2}GK\sqrt{T}\ .$$

In particular, let p_t be the re-weighting function returned by Lemma 3.10 for a candidate $\boldsymbol{w}^{(t)}$. If $\boldsymbol{w}^{(t)} = \mathbf{0}$, we set p_t to be the zero function. The objective function that we give to the online gradient descent algorithm, in the t-th step, is an estimator of $\ell_t(\boldsymbol{w}^{(t)}) = -\mathbf{E}_{(\boldsymbol{x},y)\sim D}[(p_t(\boldsymbol{w}) + \lambda)\boldsymbol{w}\cdot\boldsymbol{x}y]$, where λ is a non-negative parameter. Using ℓ_t, we perform a gradient update and project to get a new candidate $\boldsymbol{w}^{(t+1)}$. The OPGD guarantees that after roughly $d^{\Theta(k)}$ steps, there exists a t, where the value of function ℓ_t for our candidate is close to the value of the optimal one. From Theorem 3.4, we know that this is possible only if the angle between the candidate and the optimal is less than ϵ. For each iteration t, Step 15 of Algorithm 4 uses the OPGD algorithm, and the remaining steps are used to calculate the function ℓ_t.

Algorithm 4 Learning Halfspaces with Tsybakov Noise

```
1:  procedure ALG($\epsilon, \delta$)                                   ▷ $\epsilon$: accuracy, $\delta$: confidence
2:      $\boldsymbol{w}^{(0)} \leftarrow \boldsymbol{e}_1$
3:      $k \leftarrow \Theta\left(\frac{1}{\alpha^2 R\beta} \log^2\left(\frac{BA}{\epsilon LR}\right)\right)$
4:      $T \leftarrow d^{\Theta(k)}$
5:      for $t = 1, \ldots, T$ do
6:          $\eta_t \leftarrow \frac{1}{d^{\Theta(k)}\sqrt{t}}$
7:          If $\boldsymbol{w}^{(t-1)} = \mathbf{0}$ then
8:              $p_t \leftarrow 0$
9:          Else
10:             $p_t$ gets the output of SDP (3.3) with input $\boldsymbol{w}^{(t-1)} / \left\|\boldsymbol{w}^{(t-1)}\right\|_2$     ▷ Lemma 3.11
11:         If SDP fails and $\boldsymbol{w}^{(t-1)} \neq \mathbf{0}$ then
12:             return $\boldsymbol{w}^{(t-1)}$
13:         Draw $N = d^{\Theta(k)} \log(T/\delta)$ samples $\{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, y^{(N)})\}$ from $D$
14:         Set $\hat{\ell}_t(\boldsymbol{w})$ according to Lemma 3.17
15:         $\boldsymbol{w}^{(t)} \leftarrow \Pi_{\mathcal{V}}\left(\boldsymbol{w}^{(t-1)} - \eta_t \nabla_{\boldsymbol{w}} \hat{\ell}_t\left(\boldsymbol{w}^{(t-1)}\right)\right)$   ▷ $\mathcal{V} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq 1\}$
```

For the set $\mathcal{V}$, i.e., the unit ball with respect the $\|\cdot\|_2$, the diameter K equals to 2. We are going to show that in fact the optimal vector $\boldsymbol{w}^*$ and our current candidate vector $\boldsymbol{w}^{(t)}$ have indeed a separation in the value of ℓ_t. Because we do not have access to ℓ_t to optimize, we need a function $\hat{\ell}_t$, which is close to ℓ_t with high probability. The following lemma, which is proven in Appendix B.1, gives us an efficient way to compute an approximation $\hat{\ell}_t$ of ℓ_t.

Lemma 3.17 (Estimating the function ℓ_t)**.** *Let $p_t(\boldsymbol{x})$ be the non-negative function, given from the SDP* (3.3)*. Then taking $d^{O(k)} \log(1/\delta)$ samples, where $k = O\left(\frac{1}{\alpha^2 R\beta} \log^2\left(\frac{BA}{\epsilon LR}\right)\right)$,* *we can efficiently compute a function $\hat{\ell}_t(\boldsymbol{w})$ such that with probability at least $1-\delta$, the following conditions hold*

- $|\hat{\ell}_t(\boldsymbol{w}) - \mathbf{E}_{(\boldsymbol{x},y)\sim D}[(p_t(\boldsymbol{x})+\lambda)y\boldsymbol{w}\cdot\boldsymbol{x}]| \le \epsilon$*, for any* $\lambda > 0$ *and* $\boldsymbol{w} \in \mathcal{V}$*,*
- $\left\|\nabla_{\boldsymbol{w}}\hat{\ell}_t\right\|_2 \le d^{O(k)}$ *.*

The last thing we need to proceed to our main proof is to show that when the Algorithm 4 in Step 10 returns a function p_t, then there exists a function ℓ_t for which our current candidate vector $\boldsymbol{w}^{(t)}$ and the optimal one $\boldsymbol{w}^*$ are not close.

Lemma 3.18 (Error of ℓ_t)**.** *Let $\boldsymbol{w}^{(t)}$ be a vector in $\mathcal{V}$ and $\boldsymbol{w}^*$ be the optimal vector. Let $g_t(\boldsymbol{x}) = -(p_t(\boldsymbol{x})+\lambda)$ and $\ell_t(\boldsymbol{w}) = \mathbf{E}_{(\boldsymbol{x},y)\sim D}[g_t(\boldsymbol{x})y\boldsymbol{x}\cdot\boldsymbol{w}]$, where $p_t(\boldsymbol{x})$ is a non-negative function such that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[p_t(\boldsymbol{x})y\boldsymbol{w}^{(t)}\cdot\boldsymbol{x}] \le -\left\|\boldsymbol{w}^{(t)}\right\|_2\frac{\theta R}{16}$ and λ a non-negative parameter. Then it holds*

$$\ell_t(\boldsymbol{w}^*) \le -\lambda\frac{R}{4}C_\alpha^A\left(\frac{R^2\,L}{16}\right)^{1/\alpha} \quad \text{and} \quad \ell_t(\boldsymbol{w}^{(t)}) \ge \left\|\boldsymbol{w}^{(t)}\right\|_2\left(\frac{R\theta}{16}-\lambda\right).$$

Proof. Without loss of generality, let $\boldsymbol{w}^* = \boldsymbol{e}_1$. From Fact 3.3 and the definition of $\eta(\boldsymbol{x})$, for every $t \in [T]$, it holds $\ell_t(\boldsymbol{w}^*) \le -\lambda\,\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|\boldsymbol{w}^*\cdot\boldsymbol{x}|(1-2\eta(\boldsymbol{x}))]$. To bound from above the expectation, we use the (L,R,B,β)-bound properties. We have

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|\boldsymbol{w}^*\cdot\boldsymbol{x}|(1-2\eta(\boldsymbol{x}))] &\ge \frac{R}{2\sqrt{2}}\int_{R/2}^{R/\sqrt{2}}\int_{R/(2\sqrt{2})}^{R/\sqrt{2}}(1-2\eta(x_1,x_2))\gamma(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2 \qquad (3.8)\\ &\ge \frac{R}{4}C_\alpha^A\left(\frac{R^2\,L}{16}\right)^{1/\alpha},\end{aligned}$$

where in the last inequality we used Lemma 3.5. Thus, $\ell_t(\boldsymbol{w}^*) \le -\lambda\frac{R}{4}C_\alpha^A\left(\frac{R^2\,L}{16}\right)^{1/\alpha}$.

From Lemma 3.4, we have that

$$\ell_t(\boldsymbol{w}^{(t)}) = - \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}\left[(p_t(\boldsymbol{x})+\lambda)\boldsymbol{w}^{(t)}\cdot \boldsymbol{x}y\right] \geq \left\|\boldsymbol{w}^{(t)}\right\|_2 \frac{R\theta}{16} - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\lambda \boldsymbol{w}^{(t)}\cdot \boldsymbol{x}y\right]$$
$$\geq \left\|\boldsymbol{w}^{(t)}\right\|_2 \frac{R\theta}{16} - \lambda\sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\boldsymbol{w}^{(t)}\cdot \boldsymbol{x}^2\right]} \geq \left\|\boldsymbol{w}^{(t)}\right\|_2 \left(\frac{R\theta}{16} - \lambda\right) ,$$

where we used the Cauchy-Schwarz inequality and the fact that $\boldsymbol{x}$ is in isotropic position. □

We are now ready to prove our main results.

Proof of Theorem 3.12. We start by setting all the parameters that we use in the proof. Let $k = \Theta\left(\frac{1}{\alpha^2 R\beta}\log^2\left(\frac{BA}{\epsilon LR}\right)\right)$ and $\epsilon' = \epsilon\frac{R^2}{512}C_\alpha^A\left(\frac{R^2 L}{16}\right)^{\frac{1}{\alpha}}$. Assume, in order to reach a contradiction, that for all steps t, $\theta\left(\boldsymbol{w}^{(t)}, \boldsymbol{w}^*\right) \geq \epsilon$. Let $p_t(\boldsymbol{x})$ be the non-negative function output by the algorithm in Step 10. Then, from Lemma 3.11, we have that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[p_t(\boldsymbol{x})y\boldsymbol{w}^{(t)}\cdot\boldsymbol{x}] \leq -\left\|\boldsymbol{w}^{(t)}\right\|_2 \epsilon\frac{R}{16}$. Let $\hat{\ell}_t(\boldsymbol{w})$ be as in Lemma 3.17. Then $\ell_t(\boldsymbol{w}) = \mathbf{E}[\hat{\ell}_t(\boldsymbol{w})] = -\mathbf{E}_{(\boldsymbol{x},y)\sim D}[(p_t(\boldsymbol{x})+\lambda)\,y\boldsymbol{x}\cdot\boldsymbol{w}]$. Now using Lemma 3.17, for $N = \frac{d^{O(k)}}{\epsilon'^2}\log\left(\frac{T}{\delta}\right)$ samples, we have $\mathbf{Pr}\left[|\hat{\ell}_t(\boldsymbol{w}^{(t)}) - \ell_t(\boldsymbol{w}^{(t)})| \geq \epsilon'\right] \leq \frac{\delta}{2T}$ and $\mathbf{Pr}\left[|\hat{\ell}_t(\boldsymbol{w}^*) - \ell_t(\boldsymbol{w}^*)| \geq \epsilon'\right] \leq \frac{\delta}{2T}$. From Lemma 3.18, for $\lambda = \epsilon\frac{R}{32}$, in each step t we have $\ell_t(\boldsymbol{w}^{(t)}) \geq \left\|\boldsymbol{w}^{(t)}\right\|_2 \frac{R}{32}\epsilon$ and $\ell_t(\boldsymbol{w}^*) \leq -4\epsilon'$. From Lemma 3.16, for $G = d^{O(k)}$ and $K = 2$, we get

$$\sum_{t=1}^{T}\frac{\hat{\ell}_t\left(\boldsymbol{w}^{(t)}\right)}{T} - \sum_{t=1}^{T}\frac{\hat{\ell}_t\left(\boldsymbol{w}^*\right)}{T} \leq \frac{3d^{O(k)}}{\sqrt{T}} .$$

By the union bound, it follows that with probability at least $1-\delta$, we have that

$$\sum_{t=1}^{T}\frac{\ell_t\left(\boldsymbol{w}^{(t)}\right)}{T} - \sum_{t=1}^{T}\frac{\ell_t\left(\boldsymbol{w}^*\right)}{T} \leq \frac{3d^{O(k)}}{\sqrt{T}} + 2\epsilon' .$$

Thus, if the number of steps is $T = d^{\Theta(k)}/\epsilon'^2$ then, with probability at least $1-\delta$

we have that, $\frac{1}{T}\sum_{t=1}^{T}\ell_t\left(\boldsymbol{w}^{(t)}\right)-\ell_t\left(\boldsymbol{w}^*\right)\leq 3\epsilon'$. This means that there exists $t\in[T]$ such that $\ell_t\left(\boldsymbol{w}^{(t)}\right)-\ell_t\left(\boldsymbol{w}^*\right)\leq 3\epsilon'$, which implies that $\ell_t\left(\boldsymbol{w}^{(t)}\right)<-\epsilon'$ because from Lemma 3.18 it holds $\ell_t\left(\boldsymbol{w}^*\right)\leq -4\epsilon'$. Using the contrapositive of Theorem 3.4, it follows that Step 10 does not return a witnessing function and also the $\boldsymbol{w}^{(t)}$ is not zero because then $\ell_t(\boldsymbol{w}^{(t)})=0$, which lead us to a contradiction. Therefore, we have that for the last t it holds $\theta\left(\boldsymbol{w}^{(t)},\boldsymbol{w}^*\right)\leq\epsilon$. Moreover, the number of samples is $O(TN)=(dk)^{O(k)}\log(1/\delta)$, and since k is smaller than the dimension we use $d^{O(k)}\log(1/\delta)$ samples. □

To prove the Theorem 3.13, we need the following claim for the (L,R,U,B,β)-bounded distributions.

Claim 3.19 (Claim 2.1 of Diakonikolas et al. (2020c)). *Let $D_{\boldsymbol{x}}$ be an (L,R,U,B,β)-bounded distribution on $\mathbb{R}^d$. Then, for any $0<\epsilon\leq 1$, we have that $\mathrm{err}_{0-1}^{D_{\boldsymbol{x}}}(h_{\boldsymbol{u}},h_{\boldsymbol{v}})\leq U\frac{\log^2\left(\frac{B}{\epsilon}\right)}{\beta^2}\cdot\theta(\boldsymbol{v},\boldsymbol{u})+\epsilon$.*

Proof of Theorem 3.13. We run Algorithm 4 for $\epsilon'=\frac{\epsilon\beta^2}{2U}\frac{1}{\log(2/\epsilon)}$. From Theorem 3.12, Algorithm 4 outputs a $\hat{\boldsymbol{w}}$ such that $\theta(\hat{\boldsymbol{w}},\boldsymbol{w}^*)\leq\frac{\epsilon\beta^2}{2U}\frac{1}{2\log(1/\epsilon)}$. From Claim 3.19, we have that $\mathrm{err}_{0-1}(h_{\hat{\boldsymbol{w}}},f)\leq\epsilon$. This completes the proof. □

3.5 Further Related Work

It is instructive to compare the Tsybakov noise model with two other classical noise models, namely the agnostic model Haussler (1992); Kearns et al. (1994a) and the bounded (or Massart) noise model Sloan (1988); Massart and Nedelec (2006). The Tsybakov noise model lies in between these two models.

In the agnostic model Haussler (1992); Kearns et al. (1994a), the learner is given access to iid labeled examples from an arbitrary distribution D on labeled examples $(\boldsymbol{x},y)\in\mathbb{R}^d\times\{\pm 1\}$ and the goal of the learner is to output a hypothesis h such that the misclassification error $\mathrm{err}_{0-1}^{D}(h):-\mathbf{Pr}_{(\boldsymbol{x},y)\sim D}[h(\boldsymbol{x})\neq y]$ is as small as possible. In more detail, we want to achieve $\mathrm{err}_{0-1}^{D}(h)\leq\mathrm{opt}+\epsilon$, where

$\mathrm{opt} :- \inf_{g \in \mathcal{C}} \mathrm{err}_{0-1}^{D}(g)$ is the minimum possible misclassification error by any function in the class $\mathcal{C}$. Agnostic noise is the most challenging noise model in the literature. Without assumptions on the marginal distribution $D_{\boldsymbol{x}}$ on the (unlabaled) points, (even weak) agnostic learning is known to be computationally intractable Guruswami and Raghavendra (2006); Feldman et al. (2006a); Daniely (2016a).

On the other hand, if $D_{\boldsymbol{x}}$ is known to be well-behaved, in a precise sense, dimension-efficient agnostic algorithms are known. Specifically, the L_1-regression algorithm of Kalai et al. (2008) agnostically learns halfspaces under the standard Gaussian and, more generally, any isotropic log-concave distribution, with sample complexity and runtime $d^{m(1/\epsilon)}$, for an appropriate function m. In more detail, if $D_{\boldsymbol{x}}$ is the standard Gaussian $N(0, I)$, then $m(1/\epsilon) = \tilde{\Theta}(1/\epsilon^2)$ (see, e.g., Diakonikolas et al. (2010a,b)) and if $D_{\boldsymbol{x}}$ is any isotropic log-concave distribution, then $m(1/\epsilon) = 2^{\Theta(\mathrm{poly}(1/\epsilon))}$. These runtime bounds are tight for the L_1-regression approach, as they rely on the minimum degree of certain polynomial approximations of the univariate sign function. Moreover, recent work Diakonikolas et al. (2020b); Goel et al. (2020b) has shown Statistical Query lower bounds of $d^{\mathrm{poly}(1/\epsilon)}$ for agnostically learning halfspaces, even under Gaussian marginals.

Prior to this work, the only known algorithms for Tsybakov noise are the ones obtained via the straightforward reduction to agnostic learning. Specifically, by applying the L_1-regression algorithm Kalai et al. (2008) for $\epsilon' = \Theta(\epsilon^{1/\alpha})$ in place of ϵ, where $\alpha \in (0, 1]$ is the Tsybakov noise parameter of Definition 1.5, we have (see, e.g., Corollary 3.6) that the output hypothesis h satisfies $\mathbf{Pr}_{\boldsymbol{x} \sim D_{\boldsymbol{x}}}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})] \leq \epsilon$. This straightforward reduction leads to algorithms with runtimes $d^{\mathrm{poly}\left(1/\epsilon^{1/\alpha}\right)}$ for Gaussian marginals, and $d^{2^{\mathrm{poly}\left(1/\epsilon^{1/\alpha}\right)}}$ for log-concave marginals.

We acknowledge a related line of work Klivans et al. (2009a); Awasthi et al. (2017); Daniely (2015); Diakonikolas et al. (2018b) that gave efficient algorithms for learning halfspaces with agnostic noise under similar distributional assumptions. While these algorithms run in time $\mathrm{poly}(d/\epsilon)$, they achieve a "semi-agnostic" error guarantee of $O(\mathrm{opt}) + \epsilon$ — instead of $1 \cdot \mathrm{opt} + \epsilon$. This guarantee is significantly weaker for our purposes and cannot be used to obtain a hypothesis that is arbitrarily close to the target halfspace.

Finally, it should be noted that this work is part of the broader agenda of designing robust estimators for a range of generative models with respect to various noise models. A recent line of work Klivans et al. (2009a); Awasthi et al. (2017); Diakonikolas et al. (2016a); Lai et al. (2016a); Diakonikolas et al. (2017a, 2018a,b); Klivans et al. (2018); Diakonikolas et al. (2019c,b) has given efficient robust estimators for a range of learning tasks (both supervised and unsupervised) in the presence of a small constant fraction of adversarial corruptions.

4 LEARNING WITH TSYBAKOV NOISE IN POLYNOMIAL TIME

4.1 Formal Statement of Results

Preliminaries We will denote by $\text{proj}_U(\boldsymbol{x})$ the projection of $\boldsymbol{x}$ onto the subspace $U \subset \mathbb{R}^d$. For a subspace $U \subset \mathbb{R}^d$, let $U^\perp$ be the orthogonal complement of U. For a vector $\boldsymbol{w} \in \mathbb{R}^d$, we use $\boldsymbol{w}^\perp$ to denote the subspace spanned by vectors orthogonal to $\boldsymbol{w}$, i.e., $\boldsymbol{w}^\perp = \{\boldsymbol{u} \in \mathbb{R}^d : \boldsymbol{w} \cdot \boldsymbol{u} = 0\}$. Finally, we denote by $\boldsymbol{w}^{\perp_{\boldsymbol{v}}}$ the projection of the vector $\boldsymbol{w}$ on the subspace $\boldsymbol{v}^\perp$ after normalization, i.e., $\boldsymbol{w}^{\perp_{\boldsymbol{v}}} = \frac{\boldsymbol{w} - \boldsymbol{w} \cdot \boldsymbol{v}\, \boldsymbol{v}}{\|\boldsymbol{w} - \boldsymbol{w} \cdot \boldsymbol{v}\, \boldsymbol{v}\|_2}$.

In this chapter we present the first polynomial-time algorithm for learning halfspaces with Tsybakov noise. Starting from a non-trivial warm-start, our algorithm performs a novel "win-win" iterative process which, at each step, either finds a valid certificate or improves the angle between the current halfspace and the true one. Our warm-start algorithm for isotropic log-concave distributions involves a number of analytic tools that may be of broader interest. These include a new efficient method for reweighting the distribution in order to recenter it and a novel characterization of the spectrum of the degree-2 Chow parameters. We start by defining the distribution family for which our algorithms succeed.

Definition 4.1 (Well-Behaved Distributions)**.** *For $L, R, U > 0$ and $k \in \mathbb{Z}_+$, a distribution $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ is called (k, L, R, U)-well-behaved if for any projection $(D_{\boldsymbol{x}})_V$ of $D_{\boldsymbol{x}}$ on a k-dimensional subspace V of $\mathbb{R}^d$, the corresponding pdf γ_V on V satisfies the following properties: (i) $\gamma_V(\boldsymbol{x}) \geq L$, for all $\boldsymbol{x} \in V$ with $\|\boldsymbol{x}\|_2 \leq R$ (anti-anti-concentration), and (ii) $\gamma_V(\boldsymbol{x}) \leq U$ for all $\boldsymbol{x} \in V$ (anti-concentration). If, additionally, there exists $\beta \geq 1$ such that, for any $t > 0$ and unit vector $\boldsymbol{w} \in \mathbb{R}^d$, we have that $\mathbf{Pr}_{\boldsymbol{x} \sim D_{\boldsymbol{x}}}[|\boldsymbol{w} \cdot \boldsymbol{x}| \geq t] \leq \exp(1 - t/\beta)$ (sub-exponential concentration), we call $D_{\boldsymbol{x}}$ (k, L, R, U, β)-well-behaved.*

We focus on the case that the marginal distribution $D_{\boldsymbol{x}}$ on the examples is well-behaved for some values of the relevant parameters. Definition 4.1 specifies the concentration and anti-concentration conditions on the low-dimensional projections of the data distribution that are required for our learning algorithm. Throughout this paper, we will take $k = 3$, i.e., we only require 3-dimensional projections to have such properties.

Interestingly, the class of well-behaved distributions is quite broad. In particular, it is easy to show that the broad class of isotropic log-concave distributions is well-behaved for L, R, U, β being universal constants. Moreover, as Definition 4.1 does not require a specific functional form for the underlying density function, it encompasses a much more general set of distributions.

Since the complexity of our algorithm depends (polynomially) on $1/L, 1/R, U, \beta$, we state here a simplified version of our main result for the case that these parameters are bounded by a universal constant. To simplify the relevant theorem statements, we will sometimes say that a distribution D of labeled examples in $\mathbb{R}^d \times \{\pm 1\}$ is well-behaved to mean that its marginal distribution $D_{\boldsymbol{x}}$ is well-behaved. We show:

Theorem 4.2 (Learning Tsybakov Halfspaces under Well-Behaved Distributions)**.** *Let D be a well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$. There exists an algorithm that draws $N = O_{A,\alpha}(d/\epsilon)^{O(1/\alpha)}$ samples from D, runs in $\mathrm{poly}(N, d)$ time, and computes a vector $\widehat{\boldsymbol{w}}$ such that, with high probability we have that $\mathrm{err}_{0-1}^{D_x}(h_{\widehat{\boldsymbol{w}}}, f) \leq \epsilon$.*

See Theorem 4.39 for a more detailed statement.

For the class of log-concave distributions, we give a significantly more efficient algorithm:

Theorem 4.3 (Learning Tsybakov Halfspaces under Log-concave Distributions)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$ and is such that $D_{\boldsymbol{x}}$ is isotropic log-concave. There exists an algorithm that draws $N = \mathrm{poly}(d)\, O(A/\epsilon)^{O(1/\alpha^2)}$ samples* *from D, runs in* $\mathrm{poly}(N, d)$ *time, and computes a vector* $\widehat{\boldsymbol{w}}$ *such that, with high probability, we have that* $\mathrm{err}_{0-1}^{D_x}(h_{\widehat{\boldsymbol{w}}}, f) \leq \epsilon$.

See Theorem 4.40 for a more detailed statement. Since the sample complexity of the problem is $\mathrm{poly}(d, 1/\epsilon^{1/\alpha})$, the algorithm of Theorem 4.3 is qualitatively close to best possible.

4.2 Overview of Techniques

Here we give an intuitive summary of our techniques in tandem with a comparison to the most relevant prior work. A more detailed technical discussion is provided in the proceeding sections.

Our learning algorithms employ the certificate-based framework of Diakonikolas et al. (2021b). At a high-level, this framework allows us to efficiently reduce the problem of *finding* a near-optimal halfspace $h_{\widehat{\boldsymbol{w}}}(\boldsymbol{x}) = \mathrm{sign}(\langle \widehat{\boldsymbol{w}}, \boldsymbol{x} \rangle)$ to the (easier) problem of *certifying* whether a candidate halfspace $h_{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$ is "far" from the optimal halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle)$. The idea is to use a certificate algorithm (as a black-box) and combine it with an online convex optimization routine. Roughly speaking, starting from an initial guess $\boldsymbol{w}_0$ for $\boldsymbol{w}^*$, a judicious combination of these two ingredients allows us to efficiently compute a near-optimal halfspace $\widehat{\boldsymbol{w}}$, i.e., one that the certifying algorithm cannot reject. We note that a similar approach has been used in Chen et al. (2020a) for converting non-proper learners to proper learners in the Massart noise model.

With the aforementioned approach as the starting point, the learning problem reduces to that of designing an efficient certifying algorithm. In recent work Diakonikolas et al. (2021b), the authors developed a certifying algorithm for Tsybakov halfspaces based on high-dimensional polynomial regression. This method leads to a certifying algorithm with sample complexity and runtime $d^{\mathrm{polylog}(1/\epsilon)}$, i.e., a quasi-polynomial upper bound. As we will explain in Section 4.3, the Diakonikolas et al. (2021b) approach is inherently limited to quasi-polynomial time and new ideas are needed to obtain a polynomial time algorithm. *The main contribution of* *this paper is the design of a polynomial-time certificate algorithm for Tsybakov halfspaces under well-behaved distributions.*

The key idea to design a certificate in the Tsybakov noise model is the following simple but crucial observation: If $\boldsymbol{w}^*$ is the normal vector to true halfspace, then for any non-negative function $T(\boldsymbol{x})$, it holds that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})y\,\boldsymbol{w}^* \cdot \boldsymbol{x}] \geq 0$. On the other hand, for any $\boldsymbol{w} \neq \boldsymbol{w}^*$ there exists a non-negative function $T(\boldsymbol{x})$ such that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\,y\,\boldsymbol{w} \cdot \boldsymbol{x}] < 0$. In other words, there exists a *reweighting of the space* that makes the expectation of $y\boldsymbol{w} \cdot \boldsymbol{x}$ negative (Fact 3.3). Note that we can always use as $T(\boldsymbol{x})$ the indicator of the disagreement region between the candidate halfspace $h_{\boldsymbol{w}}(\boldsymbol{x})$ and the optimal halfspace $f(\boldsymbol{x}) = h_{\boldsymbol{w}^*}(\boldsymbol{x})$. Of course, since optimizing over the space of non-negative functions is intractable, we need to restrict our search space to a "simple" parametric family of functions. In Diakonikolas et al. (2021b), squares of low-degree polynomials were used, which led to a quasi-polynomial upper bound.

In this work, we consider certifying functions of the form:

$$T(\boldsymbol{x}) = \frac{1}{\boldsymbol{w} \cdot \boldsymbol{x}} \mathbb{1}\left\{\sigma_1 \leq \boldsymbol{w} \cdot \boldsymbol{x} \leq \sigma_2\,, -t_1 \leq \boldsymbol{v} \cdot \mathrm{proj}_{\boldsymbol{w}^\perp} \frac{\boldsymbol{x}}{\boldsymbol{w} \cdot \boldsymbol{x}} \leq -t_2\right\}$$

that are parameterized by a vector $\boldsymbol{v}$ and scalar thresholds $\sigma_1, \sigma_2, t_1, t_2 > 0$. Here $\mathrm{proj}_{\boldsymbol{w}^\perp}$ denotes the orthogonal projection on the subspace orthogonal to $\boldsymbol{w}$. It will be important for our approach that functions of this form are specified by $O(d)$ parameters.

Of course, it may not be a priori clear why functions of this form can be used as certifying functions in our setting. The intuition behind choosing functions of this simple form is given in Section 4.3. In particular, in Claim 4.6, we show that for any incorrect guess $\boldsymbol{w}$ there *exists* a *certifying vector* $\boldsymbol{v}$ that makes the expectation $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\,y\,\boldsymbol{w} \cdot \boldsymbol{x}]$ negative. In fact, the vector $\boldsymbol{v} = \mathrm{proj}_{\boldsymbol{w}^\perp}\boldsymbol{w}^* / \left\|\mathrm{proj}_{\boldsymbol{w}^\perp}\boldsymbol{w}^*\right\|_2 := (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ suffices for this purpose.

The key challenge is in finding such a certifying vector $\boldsymbol{v}$ algorithmically. We note that our algorithm in general does not find $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. But it does find a vector $\boldsymbol{v}$ with similar behavior, in the sense of making the $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\,y\,\boldsymbol{w} \cdot \boldsymbol{x}]$ sufficiently negative. To achieve this goal, we take a two-step approach: The first step involves computing an initialization vector $\boldsymbol{v}_0$ that has non-trivial correlation with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. In our second step, we give a perceptron-like update rule that iteratively improves the initial guess until it converges to a certifying vector $\boldsymbol{v}$. While this algorithm is relatively simple, its correctness relies on a win-win analysis (Lemma 4.14) whose proof is quite elaborate. In more detail, we show that for any *non-certifying* vector $\boldsymbol{v}$ that is sufficiently correlated with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$, we can efficiently compute a direction that improves its correlation to $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. We then argue (Lemma 4.19) that by choosing an appropriate step size this iteration converges to a certifying vector within a small number of steps.

A subtle point is that the aforementioned analysis does not take place in the initial space, where the underlying distribution is well-behaved and the labels are Tsybakov homogeneous halfspaces, but in a transformed space. The transformed space is obtained by restricting our points in a band and then performing an appropriate "perspective" projection on the subspace orthogonal to $\boldsymbol{w}$ (Section 4.3). Fortunately, we are able to show (Proposition 4.8) that this transformation preserves the structure of the problem: The transformed distribution remains well-behaved (albeit with somewhat worse parameters) and satisfies the Tsybakov noise condition (again with somewhat worse parameters) with respect to a potentially biased halfspace. In fact, this consideration motivated our use of the perspective projection in the definition of $T(\boldsymbol{x})$.

It remains to argue how to compute an initialization vector $\boldsymbol{v}_0$ that acts as a warm-start for our algorithm. Naturally, the sample complexity and runtime of our certificate algorithm depend on the quality of the initialization. The simplest way to initialize is by using a random unit vector. With random initialization, we achieve initial correlation roughly $1/\sqrt{d}$, which leads to a certifying algorithm with complexity $(d/\epsilon)^{O(1/\alpha)}$ (Theorem 4.5). This simple initialization suffices to obtain Theorem 4.2 for the general class of well-behaved distributions.

To obtain our faster algorithm for log-concave marginals (Theorem 4.3), we use the exact same approach described above starting from a better initialization. Our algorithm to obtain a better starting vector leverages additional structural properties of log-concave distributions. Our initialization algorithm runs in $\mathrm{poly}(d)$ time (independent of $1/\alpha$) and computes a unit vector whose correlation with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ is $\Omega(\epsilon^{1/\alpha})$ (Theorem 4.24).

Specifically, our initialization algorithm works as follows:

1. It starts by conditioning on a random sufficiently narrow band around the current candidate $\boldsymbol{w}$ and projecting the samples on the subspace $\boldsymbol{w}^\perp$.

2. It transforms the resulting distribution to ensure that it is isotropic log-concave through rescaling and rejection sampling.

3. It then computes the degree-2 *Chow parameters* and uses them to construct a low-dimensional subspace V inside which $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ has sufficiently large projection. This subspace V is the span of the degree-1 Chow vector and the large eigenvectors of the degree-2 Chow matrix.

4. Finally, the algorithm outputs a uniformly random vector in V that can be shown to have the desired correlation with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$.

The resulting distribution after the initial conditioning in Step 1 is still log-concave and approximately satisfies the Tsybakov noise condition with respect to a near-origin centered halfspace orthogonal to $\boldsymbol{w}$. However, the distribution may no longer be zero-centered and may contain a tiny amount of non-Tsybakov noise — in the sense that we may end with points $\boldsymbol{x}$ having $\eta(\boldsymbol{x}) > 1/2$. As we can control the total non-Tsybakov noise, the latter is not a significant issue. We address the former issue by reweighting the distribution to make it isotropic. We do this by applying rejection sampling with probability $\min(1, \exp(-\langle \boldsymbol{x}, \boldsymbol{r} \rangle))$, for some vector $\boldsymbol{r}$ that we compute via SGD (so that the resulting mean is near-zero) and then rescaling by the inverse covariance matrix.

After the first two steps, our goal is to find any vector with non-trivial correlation $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$, given that the underlying distribution is isotropic log-concave. We show that the labels y must correlate with some degree-2 polynomial in $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x}$ (Lemma 4.31). Our algorithm crucially exploits this property, along with recently established "thin shell" estimates Lee and Vempala (2017) for log-concave distributions, to show that a large part of this correlation is explained by the vector of degree-1 Chow parameters and the top few eigenvectors of the degree-2 Chow matrix (Lemma 4.32). This implies that the subspace V spanned by those vectors contains a non-trivial part of $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$, and thus a random vector from V has non-trivial correlation with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ with constant probability.

4.3 Efficiently Certifying Non-Optimality

In this section, we give an efficient algorithm that can certify whether a candidate weight vector $\boldsymbol{w}$ defines a halfspace $h_{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$ that is far from the optimal halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle)$. Before we formally describe and analyze our algorithm, we provide some intuition.

Main Result of this Section. Fact 3.3 shows that a certifying function exists. However, in general, finding such a function is information-theoretically and computationally hard. By leveraging our distributional assumptions, we show that a certifying function of a specific simple form exists and can be computed in polynomial time.

For the rest of this section, we work with distributions that are $(3, L, R, \beta)$-well-behaved. These distributions satisfy the same properties as those in Definition 4.1, except the anti-concentration condition. (The anti-concentration condition is only required at the end of our analysis in Section 4.5 to deduce that small angle between two halfspaces implies small 0-1 error.)

Definition 4.4. *For $L, R > 0$, $\beta \geq 1$, and $k \in \mathbb{Z}_+$, a distribution $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ is called (k, L, R, β)-well-behaved if the following conditions hold: (i) For any projection $(D_{\boldsymbol{x}})_V$ of $D_{\boldsymbol{x}}$ on a k-dimensional subspace V of $\mathbb{R}^d$, the corresponding pdf γ_V on V satisfies $\gamma_V(\boldsymbol{x}) \geq L$, for all $\boldsymbol{x} \in V$ with $\|\boldsymbol{x}\|_2 \leq R$ (anti-anti-concentration). (ii) For any $t > 0$ and unit vector $\boldsymbol{w} \in \mathbb{R}^d$, we have that $\mathbf{Pr}_{\boldsymbol{x} \sim D_{\boldsymbol{x}}}[|\boldsymbol{w} \cdot \boldsymbol{x}| \geq t] \leq \exp(1 - t/\beta)$ (sub-exponential concentration).*

Specifically, we have:

Theorem 4.5 (Efficiently Certifying Non-Optimality)**.** *Let D be a $(3, L, R, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$. Let $\boldsymbol{w}$ be a unit vector with $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \geq \theta$, where $\theta \in (0, \pi]$. There is an algorithm that, given as input $\boldsymbol{w}$, θ, and $N = ((A/(LR)) \cdot (d/\theta))^{O(1/\alpha)} \log(1/\delta)$ samples from D, it runs in $\mathrm{poly}(N, d)$ time, and with probability at least $1 - \delta$ returns a certifying function $T_{\boldsymbol{w}} : \mathbb{R}^d \mapsto \mathbb{R}_+$ such that*

$$\mathop{\mathbf{E}}_{(x,y)\sim D} [T_{\boldsymbol{w}}(\boldsymbol{x})\, y\boldsymbol{w} \cdot \boldsymbol{x}] \leq -\frac{1}{\beta} \left(\frac{LR\,\theta}{A\,d} \right)^{O(1/\alpha)} . \tag{4.1}$$

Intuition and Roadmap of the Proof

In this subsection, we give an intuitive proof overview of Theorem 4.5 along with pointers to the corresponding subsections where the proof of each component appears. First, we discuss the specific form of the certifying function that we compute. The proof of Fact 3.3 shows that a valid choice for the certifying function would be the characteristic function of the disagreement region between the candidate hypothesis $\boldsymbol{w}$ and the optimal halfspace $\boldsymbol{w}^*$, i.e., $T_{\boldsymbol{w}}(\boldsymbol{x}) = \mathbb{1}\{\mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x}) \neq \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x}\}$. Unfortunately, we do not know $\boldsymbol{w}^*$ (this is the vector we are trying to approximate!), and therefore it is unclear how to algorithmically use this certifying function.

Our goal is to judiciously define a parameterized family of "simple" certifying functions and optimize over this family to find one that acts similarly to the indicator of the disagreement region. A natural attempt to construct a certifying function for a guess $\boldsymbol{w}$ would be to focus on a small "band" around the candidate halfspace $\boldsymbol{w}$. This idea bears some similarity with the technique of "localization", an approach going back to Bartlett et al. (2005), which has previously seen success for the problem of efficiently learning homogeneous halfspaces with Massart noise Awasthi et al. (2015, 2016a); Zhang et al. (2020a); Diakonikolas et al. (2020c). Unfortunately, this idea is inherently insufficient to provide us with a certifying

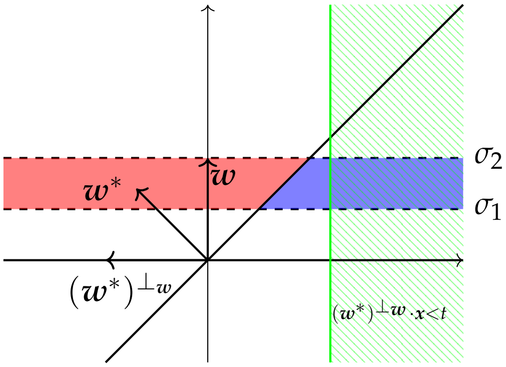

σ_2
σ_1
$\boldsymbol{w}^*$
$\boldsymbol{w}$
$(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$
$(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x} < t$

Figure 4.1: The indicator of a band $\{\boldsymbol{x} : \sigma_1 \le \boldsymbol{w} \cdot \boldsymbol{x} \le \sigma_2\}$ cannot be used as a certificate even when there is no noise and the underlying distribution is the standard Gaussian: the contribution of the positive points (red region) is larger than the contribution of the negative points (blue region). On the other hand, taking the intersection of the band and the halfspace with normal vector $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ and a sufficiently negative threshold $t < 0$ gives us a subset of the disagreement region (intersection of blue and green regions).

function for the following reason: Even an arbitrarily thin band around $\boldsymbol{w}$ will assign more probability mass on points that do not belong in the disagreement region, and therefore the expectation $\mathbf{E}_{(x,y)\sim D}[\mathbb{1}\{\sigma_1 \le \boldsymbol{w} \cdot \boldsymbol{x} \le \sigma_2\} y \boldsymbol{w} \cdot \boldsymbol{x}]$ will be positive. See Figure 4.1 for an illustration.

Intuitively, we need a way to *boost* the contribution of the disagreement region. One way to achieve this is by constructing a smooth reweighting of the space. In particular, we can look in the direction of the projection of $\boldsymbol{w}^*$ on the orthogonal complement of $\boldsymbol{w}$, i.e., the vector

$$(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} = \frac{\mathrm{proj}_{\boldsymbol{w}^\perp}(\boldsymbol{w}^*)}{\left\|\mathrm{proj}_{\boldsymbol{w}^\perp}(\boldsymbol{w}^*)\right\|_2},$$

that lies in the 2-dimensional subspace spanned by $\boldsymbol{w}$ and $\boldsymbol{w}^*$; see Figure 4.1. Notice that the disagreement region is a subset of the points that have negative inner product with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. Therefore, a candidate reweighting can be obtained by using a polynomial $p((\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x})$ of moderately large degree that will boost the points that lie in the disagreement region. This was the approach used in the recent work Diakonikolas et al. (2021b). Since $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ is not known, one needs to formulate a convex program (SDP) over the space of all d-variate polynomials of sufficiently large degree k implying that the corresponding SDP has $d^{\Omega(k)}$ variables. Unfortunately, it is not hard to show that the required degree cannot be smaller than $\Omega(\log(1/\epsilon))$. Therefore, this approach can only give a $d^{\Omega(\log(1/\epsilon))}$, i.e., quasi-polynomial, certificate algorithm.

In this work, we instead use a *hard threshold function* together with a band to isolate (a non-trivial subset of) the disagreement region. In more detail, we consider a function of the form $\mathbb{1}\{(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x} < t\}$ for some scalar threshold t; see Figure 4.1. Since $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ is unknown, we need to find a certifying vector $\boldsymbol{v}$ that is perpendicular to $\boldsymbol{w}$, i.e., $\boldsymbol{v} \in \boldsymbol{w}^{\perp}$ and acts similarly to $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. This leads us to the following **non-convex** optimization problem

$$\min_{t \in \mathbb{R}, \boldsymbol{v} \in \boldsymbol{w}^{\perp}} \mathop{\mathbf{E}}_{(\boldsymbol{x},y) \sim D} [\mathbb{1}\{\sigma_1 \leq \boldsymbol{w} \cdot \boldsymbol{x} \leq \sigma_2\} \mathbb{1}\{\boldsymbol{v} \cdot \boldsymbol{x} < t\} \boldsymbol{w} \cdot \boldsymbol{x}] .$$

Thus far, we have succeeded in reducing the number of parameters that we want to compute down to $O(d)$, but now we are faced with a non-convex optimization problem. Our main result is an efficient algorithm that computes a *certifying vector* $\boldsymbol{v}$ and a threshold t that does not necessarily minimize the above non-convex objective, but still suffice to make the corresponding expectation sufficiently negative.

We now describe the main steps we use to compute the certifying vector $\boldsymbol{v}$. The first obstacle we need to overcome is that, for $\boldsymbol{v} \in \boldsymbol{w}^{\perp}$, the corresponding instance fails to satisfy the Tsybakov noise condition. In particular, when we project the datapoints on $\boldsymbol{w}^{\perp}$, the region close to the boundary of the optimal halfspace becomes “fuzzy" even without noise: Points with different labels are mapped to the same point of $\boldsymbol{w}^{\perp}$; see Figure 4.2. We bypass this difficulty by using a *perspective projection* to map the datapoints onto $\boldsymbol{w}^{\perp}$. For non-zero vectors $\boldsymbol{w}, \boldsymbol{x} \in \mathbb{R}^d$, the perspective projection of $\boldsymbol{x}$ on $\boldsymbol{w}$ is defined as follows:

$$\pi_{\boldsymbol{w}}(\boldsymbol{x}) :- \mathrm{proj}_{\boldsymbol{w}^{\perp}} \frac{\boldsymbol{x}}{\boldsymbol{w} \cdot \boldsymbol{x}} . \tag{4.2}$$

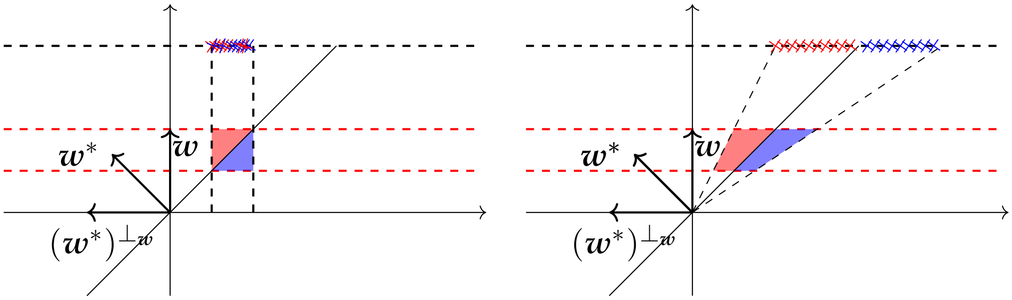

$\boldsymbol{w}^*$
$\boldsymbol{w}$
$(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$
$\boldsymbol{w}^*$
$\boldsymbol{w}$
$(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$

Figure 4.2: The dotted line on top of the figures corresponds to the subspace $\boldsymbol{w}^{\perp}$. When we project the points to $\boldsymbol{w}^{\perp}$ orthogonally (shown in left figure), we map points with different labels to the same point of $\boldsymbol{w}^{\perp}$ and obtain the “fuzzy" region where blue points (classified as negative by $\boldsymbol{w}^*$) overlap with red points (positive according to $\boldsymbol{w}^*$). On the other hand, the perspective projection (shown in the right figure) defined in Equation 4.2 preserves linear separability.

Notice that without noise the perspective projection keeps the dataset linearly separable (see Figure 4.2), which means that after we perform this projection the label noise of the resulting instance will again satisfy the Tsybakov noise condition. In addition, we show that this transformation will preserve the crucial distributional properties (concentration, anti-anti-concentration) of the underlying marginal distribution $D_{\boldsymbol{x}}$. For a detailed discussion and analysis of this data transformation, see Subsection 4.3.

Given this setup, the certificate that our algorithm will compute for a candidate weight vector $\boldsymbol{w} \in \mathbb{R}^d$ is a function of the form

$$T_{\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{\boldsymbol{w}\cdot\boldsymbol{x}} \mathbb{1}\left\{\sigma_1 \le \boldsymbol{w}\cdot\boldsymbol{x} \le \sigma_2\,, -t_1 \le \boldsymbol{v}\cdot\pi_{\boldsymbol{w}}(\boldsymbol{x}) \le -t_2\right\} =: \frac{\psi(\boldsymbol{x})}{\boldsymbol{w}\cdot\boldsymbol{x}}, \qquad (4.3)$$

for some vector $\boldsymbol{v} \in \mathbb{R}^d$ and scalars $\sigma_1, \sigma_2, t_1, t_2 > 0$. For an illustration, in Figure 4.2 we plot the set of the indicator function $\psi(\boldsymbol{x})$ which is a (high-dimensional) trapezoid.

It is not difficult to verify that by choosing $\boldsymbol{v} = (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ and appropriately picking $\sigma_1, \sigma_2, t_1, t_2$, the corresponding certificate function $T_{\boldsymbol{w}}$ resembles the indicator function of the disagreement region and certifies the *non-optimality* of the candidate halfspace $\boldsymbol{w}$. In the following claim, we prove that for any non-optimal halfspace there exists a certifying function of the above form.

Claim 4.6. *Let D be a $(3, L, R, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$. Fix any non-zero vector $\boldsymbol{w}$ such that $\theta(\boldsymbol{w}, \boldsymbol{w}^*) > 0$. Then, by setting $\boldsymbol{v} = (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ in the definition* (4.3) *of $T_{\boldsymbol{w}}(\boldsymbol{x})$, there exist $\sigma_1, \sigma_2, t_1, t_2 > 0$ such that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T_{\boldsymbol{w}}(\boldsymbol{x})\, y \boldsymbol{w} \cdot \boldsymbol{x}] < 0$.*

We note here that the proof of Claim 4.6 is sketched below for the sake of intuition and is not required for the subsequent analysis.

Proof Sketch. Setting $\boldsymbol{v} = (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ in (4.3), we have

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[T_{\boldsymbol{w}}(\boldsymbol{x})\, y \boldsymbol{w} \cdot \boldsymbol{x}] = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[\psi(\boldsymbol{x})\, y] = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[\psi(\boldsymbol{x})\,(1 - 2\eta(\boldsymbol{x}))\operatorname{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})] \ .$$

We will show that by appropriate choices of $\sigma_1, \sigma_2, t_1, t_2$ the indicator $\psi(\boldsymbol{x})$ above corresponds to a subset of the disagreement region $\{\boldsymbol{x} : \operatorname{sign}(\boldsymbol{w} \cdot \boldsymbol{x}) \neq \operatorname{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})\}$. See Figure 4.3 for an illustration. More precisely, since the distribution satisfies an anti-anti-concentration property, we can choose $\sigma_1, \sigma_2 = \Theta(R)$, so that inside the band $\{\sigma_1 \leq \boldsymbol{w} \cdot \boldsymbol{x} \leq \sigma_2\}$ there is non-zero probability mass. In particular, by setting $\sigma_1 = \rho R/2$ and $\sigma_2 = \rho R/\sqrt{2}$, for some $\rho \in (0, 1]$, we have that the band has mass roughly $\Omega(\rho R^3)$. For these choices of σ_1 and σ_2, we can pick $t_1 = \Theta(R/\rho)$ and guarantee that the slope of the corresponding line in the two-dimensional subspace is sufficiently small, so that we get a trapezoid whose intersection with the aforementioned horizontal band is large (see Figure 4.3). It remains to tune the parameter t_2. Since $\theta = \theta(\boldsymbol{w}, \boldsymbol{w}^*)$ is known, we may pick $t_2 = \Theta(R \tan\theta/\rho)$ in order to make sure that the trapezoid is a subset of the disagreement region between $\boldsymbol{w}^*$ and $\boldsymbol{w}$. □

From the above proof, it is clear that one does not really need to optimize the scalars σ_1, σ_2, t_1. Their values can be chosen according to the parameters of the underlying well-behaved distribution. Our optimization problem will be with

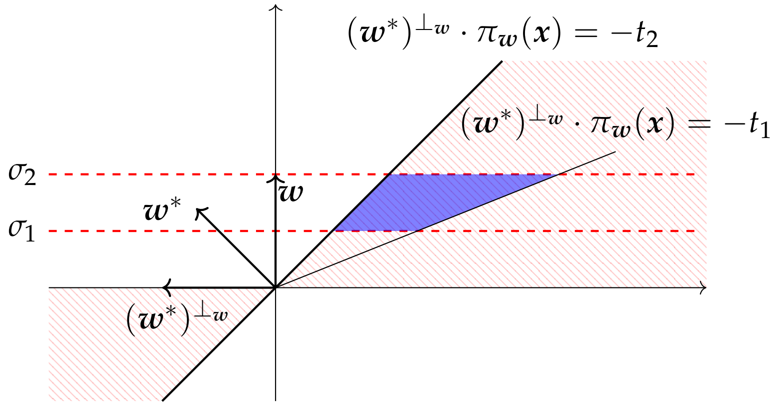

$(w^*)^{\perp_w} \cdot \pi_w(x) = -t_2$
$(w^*)^{\perp_w} \cdot \pi_w(x) = -t_1$
σ_2
w
w^*
σ_1
$(w^*)^{\perp_w}$

Figure 4.3: The function $\psi(x)$ for $v = (w^*)^{\perp_w} = \frac{\mathrm{proj}_{w^\perp}(w^*)}{\|\mathrm{proj}_{w^\perp}(w^*)\|_2}$ defined in (4.3) and appropriate scalars $\sigma_1, \sigma_2, t_1, t_2$ is the indicator of a subset of the disagreement region $\{x : \mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)\}$.

respect to the vector v and the threshold t_2. However, optimizing the expectation of the certifying function T_w of Equation (4.3) is still a non-convex problem. Given a candidate certifying vector v_0 that has non-trivial correlation with $(w^*)^{\perp_w}$, our main structural result is a **win-win** statement showing that either there exists a threshold t_2 that, together with v_0, makes the corresponding expectation of T_w sufficiently negative, or a perceptron-like update rule *will improve the correlation* between $(w^*)^{\perp_w}$ and w. In particular, we show that after roughly $\mathrm{poly}(d/\epsilon)$ updates the correlation between the guess v and $(w^*)^{\perp_w}$ will be sufficiently large so that there exists some threshold t_2 that makes v a certifying vector. Having such a vector v, it is easy to optimize over all possible thresholds and find a value for t_2 that works. For the formal statement of this claim and its proof, see Subsection 4.3 and Proposition 4.13.

Data Transformation

In this subsection, we show that we can simplify the problem of searching for a certifying vector v in $T_w(x)$ defined in Equation (4.3) by projecting the samples to an appropriate $(d-1)$-dimensional subspace via the perspective projection (4.2).

The main proposition of this subsection (Proposition 4.8) shows that this operation in some sense preserves the structure of the problem. In more detail, the transformed distribution remains well-behaved and satisfies the Tsybakov noise condition (albeit with somewhat worse parameters).

The transformation we perform is as follows:

1. We first condition on the band $B = \{\boldsymbol{x} : \boldsymbol{x} \cdot \boldsymbol{w} \in [\sigma_1, \sigma_2]\}$, for some positive parameters σ_1, σ_2.
2. We then perform the perspective projection on the samples, $\pi_{\boldsymbol{w}}(\cdot)$, defined in Equation (4.2).

To facilitate the proceeding formal description, we introduce the following definition.

Definition 4.7 (Transformed Distribution)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, $B \subseteq \mathbb{R}^d$ and $(\boldsymbol{x}, y) \sim D$.*

- *We use D_B to denote D conditioned on $\boldsymbol{x}$ being in the set B.*
- *Let $q : \mathbb{R}^d \mapsto \mathbb{R}^d$. We denote by D^q the distribution of the random variable $(q(\boldsymbol{x}), y)$.*

With the above notation, D_B^q is the distribution obtained by first conditioning on B and then applying the transformation $q(\cdot)$ to D_B.

With Definition 4.7 in place, the distribution obtained from D after we condition on the band B is D_B, and the distribution obtained from D_B after we perform the perspective projection is $D_B^{\pi_w}$. We can now state the main proposition of this subsection.

Proposition 4.8 (Properties of $D_B^{\pi_w}$)**.** *Let D be a $(3, L, R, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$. Fix any unit vector $\boldsymbol{w}$ such that $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \theta$, and let $B = \{\boldsymbol{x} : \boldsymbol{x} \cdot \boldsymbol{w} \in [\rho R/2, \rho R/\sqrt{2}]\}$, for some $\rho \in (0, 1]$. Then, for some $c = (LR)^{O(1)}$, the following conditions hold:*

1. *The distribution* $D_B^{\pi_w}$ *on* $\mathbb{R}^d \times \{\pm 1\}$ *is* $\left(2, c\rho^3, \frac{1}{\rho}, \frac{\beta}{c\rho}\log\frac{1}{\rho}\right)$*-well-behaved.*
2. *The distribution* $D_B^{\pi_w}$ *satisfies the* $\left(\alpha, \frac{A}{c\rho}\right)$*-Tsybakov noise condition with optimal classifier* $\operatorname{sign}\left((\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x} + 1/\tan\theta\right)$.

The rest of this subsection is devoted to the proof of Proposition 4.8. Before we proceed with the proof, we express the problem of finding a certifying vector $\boldsymbol{v}$ satisfying (4.3) in the transformed domain. Indeed, it is not hard to see that after we condition on B and perform the perspective projection $\pi_{\boldsymbol{w}}$, our goal is to find a vector $\boldsymbol{v}$ and scalars $t_1, t_2 > 0$ such that

$$\mathop{\mathbf{E}}_{(z,y)\sim D_B^{\pi_w}}[\mathbb{1}\{-t_1 \le \boldsymbol{v}\cdot\boldsymbol{z} \le -t_2\}\, y] < 0\,. \tag{4.4}$$

More formally, we have the following simple lemma showing that if we find a certifying vector $\boldsymbol{v}$ and parameters t_1, t_2 in the transformed instance $D_B^{\pi_w}$ satisfying Equation (4.4), the same vector and parameters will be a certificate with respect to the initial well-behaved distribution D. The relevant expectation remains negative but is slightly closer to zero.

Lemma 4.9. *Let* D *be a* $(3, L, R, \beta)$*-well-behaved distribution on* $\mathbb{R}^d$ *and let* $B = \{\boldsymbol{x} : \boldsymbol{x}\cdot\boldsymbol{w} \in [\rho R/2, \rho R/\sqrt{2}]\}$*, for some* $\rho \in (0, 1]$*. Let* $\boldsymbol{w} \in \mathbb{R}^d$ *be a unit vector and let* $\boldsymbol{v} \in \boldsymbol{w}^\perp$*,* $t_1, t_2 > 0$ *be such that* $\mathbf{E}_{(z,y)\sim D_B^{\pi_w}}[\mathbb{1}\{-t_1 \le \boldsymbol{v}\cdot\boldsymbol{z} \le -t_2\}\, y] < -C$*, for some* $C > 0$*. Then we have that* $\mathbf{E}_{(x,y)\sim D}[T_{\boldsymbol{w}}(\boldsymbol{x})\, y\boldsymbol{w}\cdot\boldsymbol{x}] = -\Omega(CLR^3\rho)$*.*

Proof. It holds

$$\begin{aligned}\mathop{\mathbf{E}}_{(z,y)\sim D_B^{\pi_w}}[\mathbb{1}\{-t_1 \le \boldsymbol{v}\cdot\boldsymbol{z} \le -t_2\}y] &= \mathop{\mathbf{E}}_{(x,y)\sim D_B}[\mathbb{1}\{-t_1 \le \boldsymbol{v}\cdot\pi_{\boldsymbol{w}}(\boldsymbol{x}) \le -t_2\}y] \\ &= \frac{1}{\mathbf{Pr}_D[B]}\mathop{\mathbf{E}}_{(x,y)\sim D}[T_{\boldsymbol{w}}(\boldsymbol{x})\boldsymbol{w}\cdot\boldsymbol{x}y]\,.\end{aligned}$$

Using the anti-anti concentration property of $D_{\boldsymbol{x}}$, we can bound $\mathbf{Pr}_D[B]$ from below. Observe that since the lower bound L on the 3-dimensional marginal density holds inside a ball of radius R, to bound the above probability from below, we can multiply L by the volume of the intersection of B with the ball of radius R. Using the formula for the volume of spherical segments, we obtain $\mathbf{Pr}_D[B] = \Omega(LR^3\rho)$. This completes the proof. □

Proof of Proposition 4.8. Our goal is to compute a certificate of the form (4.3). As we already discussed, if we had chosen to simply project the points on the subspace $\boldsymbol{w}^\perp$, we would have obtained an instance that is not linearly separable — even if the noise rate $\eta(\boldsymbol{x})$ was identically zero. By first conditioning on the set $B = \{\boldsymbol{x} : \boldsymbol{x} \cdot \boldsymbol{w} \in [\sigma_1, \sigma_2]\}$, where $\sigma_1, \sigma_2 > 0$, and then performing the perspective projection $\pi_{\boldsymbol{w}}$, we keep the dataset linearly separable (with respect to the noiseless distribution, i.e., for $\eta(\boldsymbol{x}) = 0$), albeit by a *biased* linear classifier.

We have the following lemma.

Lemma 4.10. *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that for $(\boldsymbol{x}, y) \sim D$ we have that $y = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$. Let $\boldsymbol{w}$ be any unit vector such that $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \theta \in (0, \pi]$. For $(\boldsymbol{z}, y) \sim D_B^{\pi_w}$ it holds $y = \mathrm{sign}\left((\boldsymbol{w}^*)^{\perp_w} \cdot \boldsymbol{z} + \frac{1}{\tan\theta}\right)$, i.e., the transformed distribution is linearly separable by a biased hyperplane.*

Proof. Observe that $\boldsymbol{w}^* = \lambda_1 (\boldsymbol{w}^*)^{\perp_w} + \lambda_2 \boldsymbol{w}$, where $\lambda_1 > 0$. We then have

$$\begin{aligned}\mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x}) &= \mathrm{sign}\left(\lambda_1 (\boldsymbol{w}^*)^{\perp_w} \cdot \boldsymbol{x} + \lambda_2 \boldsymbol{w} \cdot \boldsymbol{x}\right) = \mathrm{sign}\left(\lambda_1 \boldsymbol{w} \cdot \boldsymbol{x} \left(\frac{(\boldsymbol{w}^*)^{\perp_w} \cdot \boldsymbol{x}}{\boldsymbol{w} \cdot \boldsymbol{x}} + \frac{\lambda_2}{\lambda_1}\right)\right) \\ &= \mathrm{sign}\left((\boldsymbol{w}^*)^{\perp_w} \cdot \pi_{\boldsymbol{w}}(\boldsymbol{x}) + \frac{\lambda_2}{\lambda_1}\right) ,\end{aligned}$$

where to get the last equality we use the fact that λ_1 and $\boldsymbol{w} \cdot \boldsymbol{x}$ are both positive given that we conditioned on the band B. Observe that if the angle between $\boldsymbol{w}$ and $\boldsymbol{w}^*$ is θ, then $\lambda_1 = \sin\theta$ and $\lambda_2 = \cos\theta$. This completes the proof. □

We next show that conditioning on the band B will not make the Tsybakov noise condition substantially worse.

Lemma 4.11. *Let D be a $(3, L, R, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace*

$f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$. Let $B = \{\boldsymbol{x} : \boldsymbol{x} \cdot \boldsymbol{w} \in [\rho R/2, \rho R/\sqrt{2}]\}$, for some $\rho \in (0,1]$. Then D_B satisfies the Tsybakov noise condition with parameters $(\alpha, O(A/(R^3 L\rho)))$ and optimal linear classifier $\boldsymbol{w}^*$.

Proof. We have that $\mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[1-2\eta(\boldsymbol{x}) > t | \boldsymbol{x} \in B] \le \mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[1-2\eta(\boldsymbol{x}) > t] / \mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[B]$. From the proof of Lemma 4.9, we have seen that we can use the anti-anti-concentration property of $D_{\boldsymbol{x}}$ to bound $\mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[B]$ from below. Specifically, we have $\mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[B] \ge \Omega(LR^3\rho)$. Therefore, D_B satisfies the Tsybakov noise condition with parameters $(\alpha, O(A/(R^3\rho L))$. □

Finally, we show that the transformation of Equation (4.2) also preserves the anti-anti-concentration and concentration properties of the marginal distribution $D_{\boldsymbol{x}}$.

Lemma 4.12. *Let D be a $(3, L, R, \beta)$-well-behaved distribution. Fix any unit vector $\boldsymbol{w}$ and let $B = \{\boldsymbol{x} : \boldsymbol{x} \cdot \boldsymbol{w} \in [\rho R/2, \rho R/\sqrt{2}]\}$, for some $\rho \in (0,1]$. Then the transformed distribution $D_B^{\pi_{\boldsymbol{w}}}$ is $\big(2, \Omega(L\rho^3 R^3), 1/\rho, O(\beta/(R\rho)\log(1/(LR\rho)))\big)$-well-behaved.*

Proof. Let $\gamma(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}_+$ be the probability density function of $D_{\boldsymbol{x}}$ and $B = \{\boldsymbol{x} : \rho R/2 \le \boldsymbol{w} \cdot \boldsymbol{x} \le \rho R/\sqrt{2}\}$. Note that the conditional distribution $(D_{\boldsymbol{x}})_B$ of the random vector $\boldsymbol{x} \sim D_{\boldsymbol{x}}$ on the band B has density $\gamma_B(\boldsymbol{x}) = \mathbb{1}_B(\boldsymbol{x})\gamma(\boldsymbol{x})/(\int_B \gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x})$. Since the transformation $\pi_{\boldsymbol{w}}(\cdot)$ is not injective, we consider the transformation $\phi(\boldsymbol{x}) = (\boldsymbol{w} \cdot \boldsymbol{x}, \pi_{\boldsymbol{w}}(\boldsymbol{x}))$ and observe that $\phi(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}^d$ is injective. Denote by $\boldsymbol{U}$ the random variable corresponding to the image of $\boldsymbol{x}$, $\boldsymbol{x} \sim (D_{\boldsymbol{x}})_B$, under ϕ. Without loss of generality, we may assume that $\boldsymbol{w} = \boldsymbol{e}_1$. By computing the Jacobian of the above one-to-one transformation. we get that the density function of the random vector $\boldsymbol{U}$ is given by $\gamma_{\boldsymbol{U}}(\boldsymbol{u}) = |\boldsymbol{u}_1|^{d-1}\gamma_B(\boldsymbol{u}_1(1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d))$. We can marginalize out the "dummy" variable $\boldsymbol{u}_1$ to obtain the density function g of $\boldsymbol{z} \sim (D_{\boldsymbol{x}})_B^{\pi_{\boldsymbol{w}}}$, i.e.,

$$g(\boldsymbol{z}) = \int_{-\infty}^{\infty} |\boldsymbol{u}_1|^{d-1}\gamma_B(\boldsymbol{u}_1(1, \boldsymbol{z}))\mathrm{d}\boldsymbol{u}_1 .$$

Let V be any 2-dimensional subspace of $\boldsymbol{w}^\perp$. Without loss of generality, we may assume that $V = \mathrm{span}(\boldsymbol{e}_2, \boldsymbol{e}_3)$. Denote $\boldsymbol{z}_{[3,d-1]} = (\boldsymbol{z}_3, \ldots, \boldsymbol{z}_{d-1})$, $U = \mathrm{span}(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$, and $U^{\perp} = \mathrm{span}(\boldsymbol{e}_4, \ldots, \boldsymbol{e}_d)$. The marginal density of $\boldsymbol{z} \sim (D_{\boldsymbol{x}})_B^{\pi_w}$ on V is then given by

$$\begin{aligned} g_V(z_1, z_2) &= \int_{U^{\perp}} \int_{-\infty}^{\infty} |\boldsymbol{u}_1|^{d-1} \gamma_B(\boldsymbol{u}_1(1, \boldsymbol{z})) \mathrm{d}\boldsymbol{u}_1 \, \mathrm{d}\boldsymbol{z}_{[3,d-1]} \\ &= \int_{-\infty}^{\infty} |\boldsymbol{u}_1|^{d-1} \int_{U^{\perp}} \gamma_B(\boldsymbol{u}_1(1, \boldsymbol{z})) \mathrm{d}\boldsymbol{z}_{[3,d-1]} \, \mathrm{d}\boldsymbol{u}_1 \\ &= \frac{1}{\int_B \gamma(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}} \int_{\rho R/2}^{\rho R/\sqrt{2}} |\boldsymbol{u}_1|^{d-1} \int_{U^{\perp}} \gamma(\boldsymbol{u}_1(1, \boldsymbol{z})) \mathrm{d}\boldsymbol{z}_{[3,d-1]} \mathrm{d}\boldsymbol{u}_1 \\ &= \frac{1}{\int_B \gamma(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}} \int_{\rho R/2}^{\rho R/\sqrt{2}} |\boldsymbol{u}_1|^{2} \gamma_U(\boldsymbol{u}_1(1, z_1, z_2)) \mathrm{d}\boldsymbol{u}_1 \,, \end{aligned}$$

where to get the third equality we used the definition of the conditional density on B and the fact that the set B only depends on the first coordinate. The last equality follows by a change of variables. Since $D_{\boldsymbol{x}}$ is $(3, L, R, \beta)$-well-behaved, we have that if $\boldsymbol{u}_1^2(1 + z_2^2 + z_3^2) \leq R^2$ we have that $\gamma_U(\boldsymbol{u}_1(1, z_1, z_2)) \geq L$. Therefore, using the fact that $\boldsymbol{u}_1^2 \leq \rho^2 R^2/2$, we obtain that for $z_1^2 + z_2^2 \leq 2/\rho^2 - 1$ it holds $\gamma_U(\boldsymbol{u}_1(1, z_1, z_2)) \geq L$. Observe that since $\rho \leq 1$, we can get the slightly looser bound $z_2^2 + z_3^2 \leq 1/\rho^2$. Note that $\int_B \gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq 1$ and also $\int_{\rho R/2}^{\rho R/\sqrt{2}} |\boldsymbol{u}_1|^2 \mathrm{d}\boldsymbol{u}_1 = \Omega(\rho^3 R^3)$. Combining these bounds, we obtain that $g_V(z_1, z_2) \geq \Omega(L\rho^3 R^3)$.

It remains to prove that the transformed distribution still has exponentially decaying tails. In the proof of Lemma 4.11, we have already argued that the probability mass of B is bounded below by $C_B = \Omega(LR^3\rho)$. Therefore, the distribution $(D_{\boldsymbol{x}})_B$ obtained after conditioning has exponential concentration with parameter $\beta(1 - \log C_B)$. After we perform the perspective projection (Equation (4.2)) to obtain $(D_{\boldsymbol{x}})_B^{\pi_w}$, the concentration parameter becomes $2\beta(1 - \log C_B)/(\rho R)$, since we divide each coordinate of $\boldsymbol{x}$ by a quantity that is bounded from below by $R\rho/2$. This completes the proof of Lemma 4.12. □

Proposition 4.8 follows by combining Lemmas 4.10, 4.11, 4.12.

Efficient Certificate Computation Given Initialization

In this subsection, we give our main algorithm for computing a non-optimality certificate in the transformed instance, i.e., a vector $\boldsymbol{v}$ and parameters $t_1, t_2 > 0$ satisfying Equation (4.4). Recall that after the perspective projection transformation of Subsection 4.3, we now have sample access to i.i.d. labeled examples $(\boldsymbol{x}, y)$ from a well-behaved distribution D on $\mathbb{R}^d \times \{\pm 1\}$ satisfying the Tsybakov noise condition (albeit with somewhat worse parameters) with the optimal classifier being a non-homogeneous halfspace (see Proposition 4.8.)

Our certificate algorithm in this subsection assumes the existence of an initialization vector, i.e., a vector that has non-trivial correlation with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. The simplest way to find such a vector is by picking a uniformly random unit vector. A random initialization suffices for the guarantees of this subsection (and in particular for Theorem 4.5). We note that for the family of log-concave distributions, we can leverage additional structure to design a fairly sophisticated initialization algorithm that in turn leads to a faster certificate algorithm (see Section 4.4).

The main algorithmic result of this section is an efficient algorithm to compute a certifying vector satisfying Equation (4.4). Note that we are essentially working in $(d-1)$ dimensions, since we have already projected the examples to the subspace $\boldsymbol{w}^{\perp}$. As shown in Proposition 4.8, the transformed distribution $D_B^{\pi_{\boldsymbol{w}}}$ is still well-behaved and follows the Tsybakov noise condition, but with somewhat worse parameters than the initial distribution D.

To avoid clutter in the relevant expressions, we overload the notation and use D instead of $D_B^{\pi_{\boldsymbol{w}}}$ in the rest of this section. Moreover, we use the notation (L, R, β) and (α, A) to denote the well-behaved distribution's parameters and the Tsybakov noise parameters. The actual parameters of $D_B^{\pi_{\boldsymbol{w}}}$ (quantified in Proposition 4.8) are used in the proof of Theorem 4.5. To simplify notation, we will henceforth denote by $\boldsymbol{v}^*$ the vector $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. We show:

Proposition 4.13. *Let D be a $(2, L, R, \beta)$-well-behaved distribution on $\mathbb{R}^d \times \{\pm 1\}$ satisfying the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{v}^* \cdot \boldsymbol{x} + b)$. Let $\boldsymbol{v}_0 \in \mathbb{R}^d$ be a unit vector such that $\boldsymbol{v}_0 \cdot \boldsymbol{v}^* \geq 4b/R$. There* *is an algorithm (Algorithm 5) with the following performance guarantee: Given $\boldsymbol{v}_0$ and $N = d\,\frac{\beta^2 R^2}{b^2}\left(\frac{A}{RL}\right)^{O(1/\alpha)} \log(1/\delta)$ samples from D, the algorithm runs in $\mathrm{poly}(N, d)$ time, and with probability at least $1-\delta$ returns a unit vector $\boldsymbol{v} \in \mathbb{R}^d$ and a scalar $t \in \mathbb{R}_+$ such that*

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}\left[\mathbb{1}[-R \le \boldsymbol{v}\cdot\boldsymbol{x} \le -t]\,y\right] \le -\frac{b}{R\beta}\left(\frac{RL}{A}\right)^{O(1/\alpha)} .$$

Algorithm 5 employs a "perceptron-like" update rule that in polynomially many rounds succeeds in improving the angle between the initial guess $\boldsymbol{v}_0$ and the target vector $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} = \boldsymbol{v}^*$. While the algorithm is relatively simple, its proof of correctness relies on a novel structural result (Lemma 4.14) whose proof is the main technical contribution of this section. Roughly speaking, our structural result establishes the following win-win statement: Given a vector whose correlation with $\boldsymbol{v}^*$ is non-trivial, either this vector is already a certifying vector (see Item 1 of Lemma 4.14 and Lemma 4.9) or the update step will improve the angle with $\boldsymbol{v}^*$ (Item 2 of Lemma 4.14).

In more detail, starting with a vector $\boldsymbol{v}_0$ that has non-trivial correlation with $\boldsymbol{v}^*$, we consider the following update rule

$$\boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)} + \lambda \boldsymbol{g}\,, \tag{4.5}$$

where $\lambda > 0$ is an appropriately chosen step size and

$$\boldsymbol{g} = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}\left[\mathbb{1}\{-R \le \langle \boldsymbol{v}^{(t)}, \boldsymbol{x}\rangle \le -R/2\}\, y\, \mathrm{proj}_{(\boldsymbol{v}^{(t)})^\perp}(\boldsymbol{x})\right] ,$$

where $\mathrm{proj}_{(\boldsymbol{v}^{(t)})^\perp}(\boldsymbol{x})$ is the projection of $\boldsymbol{x}$ to the subspace $(\boldsymbol{v}^{(t)})^\perp$. In Lemma 4.19, we show that if $\boldsymbol{v}^{(t)}$ is not a certifying vector, i.e., it does not satisfy Item 1 of Lemma 4.19, then there exists an appropriately small step size λ that improves the correlation with $\boldsymbol{v}^*$ after the update. This is guaranteed by Item 2 of Lemma 4.19, which shows that $\boldsymbol{g}$ has positive correlation with $(\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}}$ (the normalized projection of $\boldsymbol{v}^*$ onto $\boldsymbol{v}^\perp$), and thus will turn $\boldsymbol{v}^{(t)}$ towards the direction of $\boldsymbol{v}^*$ decreasing the angle between them.

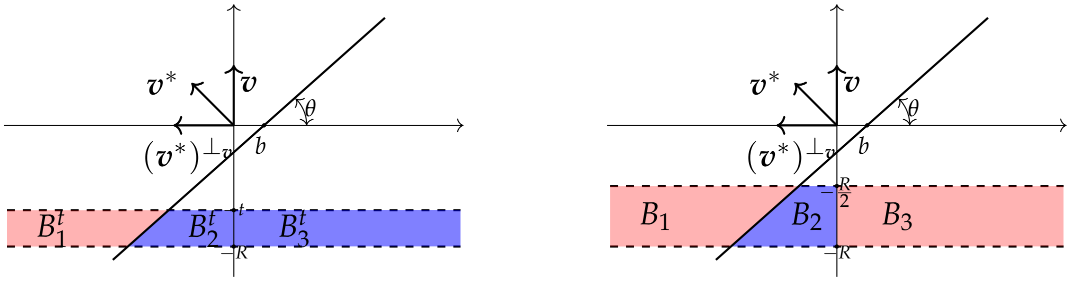


Figure 4.4: In the subspace $\boldsymbol{w}^{\perp}$, the certifying function is simply an indicator $\mathbb{1}\{-R \leq \boldsymbol{v} \cdot \boldsymbol{x} \leq -t_0\}$, for some $t_0 > 0$. See also Equation (4.4). In the left figure we plot the regions B_1, B_2, B_3 defined in the definition of I_2 in the proof of Lemma 4.14. In the right figure we plot the the regions B_1^t, B_2^t, B_3^t used in the definition of I_1^t in the proof of Lemma 4.14. The blue regions have negative contribution to the value of I_1^t (resp. I_2), while the red regions have positive contribution.

Algorithm 5 Computing a Certificate Given Initialization

```
procedure ComputeCertificate($((L,R,\beta),(A,\alpha),\delta,\boldsymbol{v}_0,\widehat{D})$)
Input: Empirical distribution $\widehat{D}$ of a $(2,L,R,\beta)$-well-behaved distribution that satisfies the $(\alpha,A)$-Tsybakov noise condition, initialization vector $\boldsymbol{v}_0$, confidence probability $\delta$.
Output: A certifying vector $\boldsymbol{v}$ and positive scalars $t_1, t_2$ that satisfy (4.4).
    $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}_0$
    $T \leftarrow \mathrm{poly}(1/L, 1/R, A)^{1/\alpha} \cdot \mathrm{poly}(1/b, 1/\beta)$
    $\lambda \leftarrow \frac{1}{\beta^3}\mathrm{poly}(L,R,1/A)^{1/\alpha}$; $c \leftarrow \frac{b}{R\beta}\mathrm{poly}(L,R,1/A)^{1/\alpha}$
    for $t = 1,\dots,T$ do
        $B^{t'} = \{\boldsymbol{x} : -R \leq \boldsymbol{v}^{(t-1)} \cdot \boldsymbol{x} \leq -t'\}$
        if there exists $t_0 \in (R/2, R]$ such that $\mathbf{E}_{(\boldsymbol{x},y)\sim\widehat{D}}[\mathbb{1}_{B^{t_0}}(\boldsymbol{x})\, y] \leq -c$
            return($\boldsymbol{v}^{(t-1)}, R, t_0$)
        $\hat{\boldsymbol{g}}^{(t)} \leftarrow \mathbf{E}_{(\boldsymbol{x},y)\sim\widehat{D}}\left[\mathbb{1}_{B^{R/2}}(\boldsymbol{x})\, y\, \mathrm{proj}_{(\boldsymbol{v}^{(t-1)})^{\perp}}(\boldsymbol{x})\right]$
        $\boldsymbol{v}^{(t)} \leftarrow \frac{\boldsymbol{v}^{(t-1)}+\lambda\hat{\boldsymbol{g}}^{(t)}}{\left\|\boldsymbol{v}^{(t-1)}+\lambda\hat{\boldsymbol{g}}^{(t)}\right\|_2}$
```

We are now ready to state and prove our win-win structural result:

Lemma 4.14 (Win-Win Result)**.** *Let D be a $(2, L, R, \beta)$-well-behaved distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{v}^* \cdot \boldsymbol{x} + b)$, and $\boldsymbol{v} \in \mathbb{R}^d$ be a unit vector with $\boldsymbol{v} \cdot \boldsymbol{v}^* \geq 4b/R$. Consider the band $B^t = \{\boldsymbol{x} : -R \leq \boldsymbol{v} \cdot \boldsymbol{x} \leq -t\}$ for $t \in [R/2, R]$ and define $\boldsymbol{g} = \mathbf{E}_{(\boldsymbol{x},y)\sim D}[\mathbb{1}_{B^{R/2}}(\boldsymbol{x})\, y\, \mathrm{proj}_{\boldsymbol{v}^\perp}(\boldsymbol{x})]$. For some $c = (RL/A)^{O(1/\alpha)}$, one of the following statements is satisfied:*

1. *There exists $t_0 \in (R/2, R]$, such that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}\left[\mathbb{1}_{B^{t_0}}(\boldsymbol{x})\, y\right] \leq -c^2 \frac{b}{R\beta}$.*
2. *It holds $\boldsymbol{g} \cdot \boldsymbol{v}^* \geq c^2 \frac{\pi b}{4\beta}$.*

Moreover, the first condition always holds if $\theta(\boldsymbol{v}, \boldsymbol{v}^*) \leq b\, c/\beta$.

Proof. Since $\boldsymbol{v}$ and $\boldsymbol{v}^*$ span a 2-dimensional subspace, we can assume without loss of generality that $\boldsymbol{v} = \boldsymbol{e}_2$ and $\boldsymbol{v}^* = (-\sin\theta, \cos\theta)$. Our analysis will consider the following regions: $B_1^t = \{\boldsymbol{x} \in B^t : f(\boldsymbol{x}) = +1\}$, $B_2^t = \left\{\boldsymbol{x} \in B^t : f(\boldsymbol{x}) = -1 \text{ and } \mathrm{proj}_{\boldsymbol{v}^\perp}\boldsymbol{x} \cdot \boldsymbol{v}^* \geq 0\right\}$, and $B_3^t = \left\{\boldsymbol{x} \in B^t : f(\boldsymbol{x}) = -1 \text{ and } \mathrm{proj}_{\boldsymbol{v}^\perp}\boldsymbol{x} \cdot \boldsymbol{v}^* < 0\right\}$. See Figures 4.3, 4.4 for an illustration.

For notation convenience, we will also denote $(\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}} = \mathrm{proj}_{\boldsymbol{v}^\perp}(\boldsymbol{v}^*) / \left\|\mathrm{proj}_{\boldsymbol{v}^\perp}(\boldsymbol{v}^*)\right\|_2$ and $\zeta(\boldsymbol{x}) = 1 - 2\eta(\boldsymbol{x})$.

Given the above notation, we can rewrite the two quantities appearing in Items 1, 2 of Lemma 4.14 as follows:

$$I_1^t = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[\mathbb{1}_{B^t}(\boldsymbol{x})y] = \underbrace{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[(\mathbb{1}_{B_1^t}(\boldsymbol{x}) - \mathbb{1}_{B_2^t}(\boldsymbol{x}))\zeta(\boldsymbol{x})\right]}_{I_{1,1}^t} - \underbrace{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\mathbb{1}_{B_3^t}(\boldsymbol{x})\zeta(\boldsymbol{x})\right]}_{I_{1,2}^t}, \tag{4.6}$$

$$\begin{aligned} I_2 = \boldsymbol{g} \cdot (\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}} &= \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[\mathbb{1}_{B^{R/2}}(\boldsymbol{x})y\boldsymbol{x}] \cdot (\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}} \\ &= \underbrace{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[(\mathbb{1}_{B_1^{R/2}}(\boldsymbol{x}) - \mathbb{1}_{B_2^{R/2}}(\boldsymbol{x}))\zeta(\boldsymbol{x})|\boldsymbol{x}_1|\right]}_{I_{2,1}} + \underbrace{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\mathbb{1}_{B_3^{R/2}}(\boldsymbol{x})\zeta(\boldsymbol{x})|\boldsymbol{x}_1|\right]}_{I_{2,2}}. \end{aligned} \tag{4.7}$$

Since $\boldsymbol{v}^* = (-\sin\theta, \cos\theta)$, the quantity $\boldsymbol{g} \cdot \boldsymbol{v}^*$ (that appears in Item 2 of Lemma 4.14) is equal to $\sin(\theta)I_2$. We work with the normalized $(\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}}$ in order to simplify notation.

Before we go into the details of the proof, we give a high-level description of the main steps with pointers to the relevant claims. Note that the quantity I_1^t corresponds to the value of the certifying function (in the subspace $\boldsymbol{w}^\perp$) when we use $\boldsymbol{v}$ as certifying vector and $t_1 = -R, t_2 = t$ as thresholds. See Equation (4.4). When I_1^t is small (see Item 1 of the lemma), we have a certifying function. On the other hand, $\sin(\theta) I_2$ corresponds to the inner product of the update $\boldsymbol{g}$ and the optimal vector $\boldsymbol{v}^*$. Item 2 of the lemma states that this quantity is large, which means that if we update according to $\boldsymbol{g}$ we shall improve the correlation with $\boldsymbol{v}^*$.

Heuristic Argument. Since the formal proof is somewhat technical, we start with a useful (but inaccurate) heuristic argument. If we ignore the presence of $|\boldsymbol{x}_1|$ in $I_{2,1}$ and $I_{2,2}$, we see from Figure 4.3 that if the contribution of region $B_2^{R/2}$ is sufficiently large compared to the positive contribution of $B_1^{R/2}$ (red region in Figure 4.3), then I_1 will be negative in total. That is, Item 1 is true. On the other hand, if the contribution of $B_2^{R/2}$ is not very large, then when we add the contribution of B_3 (red region in Figure 4.4) overall, I_2 will be positive and Item 2 now holds. Notice that in this setting we could take the threshold t in the definition of I_1^t to simply be $R/2$, i.e., use the entire band in our certificate.

Unfortunately, in the actual proof, we need to deal with the term $|\boldsymbol{x}_1|$ in the expectations of I_2 that makes the previous argument invalid. Using the Mean Value Theorem (Fact 4.18), we show that there exists a threshold $t \in [-R, -R/2]$ that makes I_1^t sufficiently negative. This is done in Claim 4.17.

We can now proceed with the formal proof. We will require several technical claims. First, we bound $I_{1,2}^{R/2}$ and $I_{2,2}$ from below using the fact that our distribution is well-behaved. We require the following claim in order to show that the expressions in Item 1 (resp. Item 2) of our lemma are not simply negative (resp. positive), but have a non-trivial gap instead. The proof of the claim relies on two important observations. First, the fact that the distribution is well-behaved means that the contribution of region B_3 would be sufficiently large if we ignore the noise function $\zeta(\boldsymbol{x})$ in the expectations. Second, we use the fact that the Tsybakov noise rate $\zeta(\boldsymbol{x}) = 1 - 2\eta(\boldsymbol{x})$ cannot reduce the contribution of a region by a lot.

Claim 4.15. *We have that $I_{1,2}^{R/2}$ and $I_{2,2}$ are bounded from below by some $c = (RL/A)^{O(1/\alpha)}$.*

The proof of Claim 4.15 can be found in Appendix C.

Now we show that if the angle between the optimal vector and the current one is small, then $I_1^{R/2}$ is negative. In particular, the first condition always holds if $\theta(\boldsymbol{v}, \boldsymbol{v}^*) \leq bc/(4\beta)$.

Claim 4.16. *If $\theta(\boldsymbol{v}, \boldsymbol{v}^*) \leq bc/(4\beta)$, then $I_1^{R/2} \leq -c/4$.*

The proof of Claim 4.16 can be found in Appendix C.

Our next claim shows that when Item 2 does not hold, then Item 1 always does. Having proved Claim 4.16, we may also assume that $\theta(\boldsymbol{v}, \boldsymbol{v}^*) \geq bc/(4\beta)$. Observe that, in this case, if $I_2 \geq c/2$, we have

$$I_2 \geq c/2 = c \sin\theta/(2\sin\theta) \geq \pi c^2 b/(4\beta \sin\theta)\,,$$

where we used the fact that $\sin(\theta) \geq 2\theta/\pi$ for all $\theta \in [0, \pi/2]$ and the fact that $\theta \geq b\,c/(4\beta)$. Therefore, to complete the proof, we need to show the following claim proving that when $I_2 \leq c/2$, Item 1 of the lemma is always true.

Claim 4.17. *If $\theta = \theta(\boldsymbol{v}, \boldsymbol{v}^*) \geq bc/(4\beta)$ and $I_2 \leq c/2$, there exists $t_0 \in (-R, -R/2]$ such that $I_1^{t_0} \leq -bc^2/(16R\beta)$.*

Proof. Given the lower bounds on $I_{2,2}$ and $I_{1,2}^{R}$, we distinguish two cases. Assume that $I_2 \leq c/2$. This implies, from Claim 4.15, that $I_{2,1} \leq -c/2$. We show that in this case there exists a t_0 such that $I_{1,1}^{t_0} \leq -bc^2/(16R\beta)$. To show this, we are going to use the following variant of the standard Mean Value Theorem (MVT) for integrals.

Fact 4.18 (Second Integral MVT). *Let $G : \mathbb{R} \mapsto \mathbb{R}_+$ be a non-negative, non-increasing, continuous function. There exists $s \in (a, b]$ such that $\int_a^b G(t)F(t)\mathrm{d}t = G(a)\int_a^s F(t)\mathrm{d}t$.*

Let $\xi(\boldsymbol{x}_2) = \boldsymbol{x}_2/\tan\theta + b/\sin\theta$ be the first coordinate of a point $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ that lies on the halfspace defined by f, where $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{v}^* \cdot \boldsymbol{x} + b)$ (see Figure 4.4).

We have

$$I_{1,1}^{t} = \int_{-R}^{-t} \left(\int_{-\infty}^{\xi(x_2)} \zeta(x_1, x_2)\gamma(x_1, x_2)\mathrm{d}x_1 - \int_{\xi(x_2)}^{0} \zeta(x_1, x_2)\gamma(x_1, x_2)\mathrm{d}x_1 \right) \mathrm{d}x_2 = \int_{-R}^{-t} g(x_2)\mathrm{d}x_2 \,,$$

where $g(x_2) = \int_{-\infty}^{\xi(x_2)} \zeta(x_1, x_2)\gamma(x_1, x_2)\mathrm{d}x_1 - \int_{\xi(x_2)}^{0} \zeta(x_1, x_2)\gamma(x_1, x_2)\mathrm{d}x_1$. Moreover,

$$\begin{aligned} I_{2,1} &= \int_{-R}^{-R/2} \left(\int_{-\infty}^{\xi(x_2)} \zeta(x_1, x_2)\gamma(x_1, x_2)|x_1|\mathrm{d}x_1 - \int_{\xi(x_2)}^{0} \zeta(x_1, x_2)\gamma(x_1, x_2)|x_1|\mathrm{d}x_1 \right) \mathrm{d}x_2 \\ &\geq \int_{-R}^{-R/2} |\xi(x_2)|g(x_2)\mathrm{d}x_2 = |\xi(-R)| \int_{-R}^{-t_0} g(x_2)\mathrm{d}x_2 = |\xi(-R)| I_{1,1}^{t_0}, \end{aligned}$$

for some $t_0 \in (-R, -R/2]$. Observe that the inequality above follows by replacing $|x_1|$ with its lower bound $|\xi(x_2)|$ in the first integral and by its upper bound $|\xi(x_2)|$ in the second.

We now observe that $|\xi(x_2)| = -x_2/\tan\theta - b/\sin\theta$, where to remove the absolute value we used the assumption that $\cos\theta \geq 4b/R$. Therefore, $|\xi(x_2)|$ is a decreasing and non-negative function of x_2. Using the Mean Value Theorem, Fact 4.18, we obtain

$$I_{2,1} \geq \int_{-R}^{-R/2} |\xi(x_2)|g(x_2)\mathrm{d}x_2 = |\xi(-R)| \int_{-R}^{-t_0} g(x_2)\mathrm{d}x_2 = |\xi(-R)| I_{1,1}^{t_0} \,. \tag{4.8}$$

Thus,

$$I_{1,1}^{t_0} \leq I_{2,1}/|\xi(-R)| \leq -c\sin\theta/(2R) \leq -bc^2/(16R\beta) \,,$$

where we used that $\theta \geq bc/(4\beta)$. This completes the proof of Claim 4.17. □

Putting together the above claims, Lemma 4.14 follows. □

In the next lemma, we show that if Item 2 of Lemma 4.14 is satisfied, then an update step decreases the angle between the current vector v and the optimal vector v^*.

Lemma 4.19 (Correlation Improvement). *For unit vectors $\boldsymbol{v}^*, \boldsymbol{v} \in \mathbb{R}^d$, let $\hat{\boldsymbol{g}} \in \mathbb{R}^d$ such that $\hat{\boldsymbol{g}} \cdot \boldsymbol{v}^* \geq \frac{c}{\beta}$, $\hat{\boldsymbol{g}} \cdot \boldsymbol{v} = 0$, and $\|\hat{\boldsymbol{g}}\|_2 \leq \beta$, with $c > 0$ and $\beta \geq 1$. Then, for $\boldsymbol{v}' = \frac{\boldsymbol{v}+\lambda\hat{\boldsymbol{g}}}{\|\boldsymbol{v}+\lambda\hat{\boldsymbol{g}}\|_2}$, with $\lambda = \frac{c}{2\beta^3}$, we have that $\boldsymbol{v}' \cdot \boldsymbol{v}^* \geq \boldsymbol{v} \cdot \boldsymbol{v}^* + \lambda^2\beta^2/2$.*

Proof. We will show that $\boldsymbol{v}' \cdot \boldsymbol{v}^* = \cos\theta' \geq \cos\theta + \lambda^2\beta^2$, where $\cos\theta = \boldsymbol{v} \cdot \boldsymbol{v}^*$. We have that

$$\|\boldsymbol{v} + \lambda\hat{\boldsymbol{g}}\|_2 = \sqrt{1 + \lambda^2 \|\hat{\boldsymbol{g}}\|_2^2 + 2\lambda\hat{\boldsymbol{g}} \cdot \boldsymbol{v}} \leq 1 + \lambda^2 \|\hat{\boldsymbol{g}}\|_2^2 \ , \tag{4.9}$$

where we used that $\sqrt{1+a} \leq 1 + a/2$. Using the update rule, we have

$$\boldsymbol{v}' \cdot \boldsymbol{v}^* = \boldsymbol{v}' \cdot (\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}} \sin\theta + \boldsymbol{v}' \cdot \boldsymbol{v} \cos\theta = \frac{\lambda\hat{\boldsymbol{g}} \cdot (\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}}}{\|\boldsymbol{v} + \lambda\hat{\boldsymbol{g}}\|_2} \sin\theta + \frac{\boldsymbol{v} + \lambda\hat{\boldsymbol{g}} \cdot \boldsymbol{v}}{\|\boldsymbol{v} + \lambda\hat{\boldsymbol{g}}\|_2} \cos\theta \ .$$

Now using Equation (4.9), we get

$$\boldsymbol{v}' \cdot \boldsymbol{v}^* \geq \frac{\lambda\hat{\boldsymbol{g}} \cdot (\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}}}{1 + \lambda^2 \|\hat{\boldsymbol{g}}\|_2^2} \sin\theta + \frac{\cos\theta}{1 + \lambda^2 \|\hat{\boldsymbol{g}}\|_2^2} = \cos\theta + \frac{\lambda\hat{\boldsymbol{g}} \cdot (\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}}}{1 + \lambda^2 \|\hat{\boldsymbol{g}}\|_2^2} \sin\theta + \frac{-\lambda^2 \|\hat{\boldsymbol{g}}\|_2^2 \cos\theta}{1 + \lambda^2 \|\hat{\boldsymbol{g}}\|_2^2} \ .$$

Then, using that $\hat{\boldsymbol{g}} \cdot \boldsymbol{v}^* = \hat{\boldsymbol{g}} \cdot (\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}} \sin\theta$, we have that $\hat{\boldsymbol{g}} \cdot (\boldsymbol{v}^*)^{\perp_{\boldsymbol{v}}} \geq \frac{c}{\beta \sin\theta}$, thus

$$\boldsymbol{v}' \cdot \boldsymbol{v}^* \geq \cos\theta + \frac{\lambda c/\beta - \lambda^2 \|\hat{\boldsymbol{g}}\|_2^2}{1 + \lambda^2 \|\hat{\boldsymbol{g}}\|_2^2} \geq \cos\theta + \frac{\lambda c/\beta - \lambda^2\beta^2}{1 + \lambda^2 \|\hat{\boldsymbol{g}}\|_2^2} = \cos\theta + \frac{1}{2}\frac{\lambda c/\beta}{1 + \lambda^2 \|\hat{\boldsymbol{g}}\|_2^2} \ ,$$

where in the first inequality we used that $\|\hat{\boldsymbol{g}}\|_2 \leq \beta$ and in the second that for $\lambda = c/(2\beta^3)$ it holds $c/\beta - \lambda\beta^2 \geq c/(2\beta)$. Finally, we have that

$$\cos\theta' = \boldsymbol{v}' \cdot \boldsymbol{v}^* \geq \cos\theta + \frac{1}{2}\frac{\lambda c/\beta}{1 + \lambda^2(9\beta^2)} \geq \cos\theta + \frac{1}{4}\lambda c/\beta = \cos\theta + \frac{1}{2}\lambda^2\beta^2 \ .$$

This completes the proof. □

To analyze the sample complexity of Algorithm 5, we require the following simple lemma, which bounds the sample complexity of estimating the update function and testing the current candidate certificate. The simple proof can be found in Appendix C.

Lemma 4.20 (Estimating $\boldsymbol{g}$)**.** *Let D be a $(2, L, R, \beta)$-well-behaved distribution. Given $N = O((d\beta^2/\epsilon^2)\log(d/\delta))$ i.i.d samples $(\boldsymbol{x}^{(i)}, y^{(i)}))$ from D, the estimator*

$$\hat{\boldsymbol{g}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{B^{R/2}}\left(\boldsymbol{x}^{(i)}\right) y^{(i)}\boldsymbol{x}^{(i)}$$

satisfies the following with probability at least $1-\delta$:

- *$\|\hat{\boldsymbol{g}} - \boldsymbol{g}\|_2 \le \epsilon$, where $\boldsymbol{g} = \mathbf{E}_{(\boldsymbol{x},y)\sim D}[\mathbb{1}_{B^{R/2}}(\boldsymbol{x})\, y\, \boldsymbol{x}]$, and*
- *$\|\hat{\boldsymbol{g}}\|_2 \le e\beta + \epsilon$.*

Before we proceed with the proof of Proposition 4.13, we show that we can efficiently check for the certificate in Line 9 of Algorithm 5 with high probability.

Lemma 4.21. *Let $\widehat{D}_N$ be the empirical distribution obtained from D with $N = O(\log(1/\delta)/\epsilon^2)$ samples. Then, with probability $1-\delta$, for every $t \in \mathbb{R}_+$, $|\, \mathbf{E}_{(\boldsymbol{x},y)\sim D}\left[\mathbb{1}_{B^t}(\boldsymbol{x})\, y\right] - \mathbf{E}_{(\boldsymbol{x},y)\sim \widehat{D}_N}\left[\mathbb{1}_{B^t}(\boldsymbol{x})\, y\right] | \le \epsilon$.*

The proof of Lemma 4.21 can be found in Appendix C. We are now ready to prove Proposition 4.13.

Proof of Proposition 4.13. Consider the k-th iteration of Algorithm 5. Let $\boldsymbol{g}^{(k)} = \mathbf{E}_{(\boldsymbol{x},y)\sim D}[\mathbb{1}_{B_k^{R/2}}(\boldsymbol{x})y\boldsymbol{x}]$, where $B_k^{R/2}(\boldsymbol{x}) = \{\boldsymbol{x} : -R \le \boldsymbol{x}\cdot\boldsymbol{v}^{(k)} \le -R/2\}$ and $G := \sqrt{b}(RL/A)^{O(1/\alpha)}$. Moreover, let $\hat{\boldsymbol{g}}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{B_k^{R/2}}\left(\boldsymbol{x}^{(i)}\right) y^{(i)}\boldsymbol{x}^{(i)}$ and note that from Lemma 4.20 we have that given $N = O\left(d\beta^2/G^4\log(1/(LR))\log(dT/\delta)\right)$ samples, for every iteration k, it holds that $\left\|\hat{\boldsymbol{g}}^{(k)} - \boldsymbol{g}^{(k)}\right\|_2 \le G^2/(16\beta)$ and $\left\|\hat{\boldsymbol{g}}^{(k)}\right\|_2 \le e\beta + G^2/(16\beta) \le 3\beta$, with probability $1-\delta/T$.

We first show that if Condition 1 of Lemma 4.14 is satisfied, then Algorithm 5 terminates at Line 10 returning a certifying vector. The only issue is that we have access to the empirical distribution $\widehat{D}_N$ instead of D. From Lemma 4.21, we have that the empirical expectation of Line 9 is sufficiently close to the true expectation that appears in Condition 1 of Lemma 4.14, thus it is going to find it.

We now analyze the case when Condition 1 of Lemma 4.14 is not true. From Lemma 4.14, we immediately get that since Condition 1 is not satisfied, Condition 2 is true. Then, using the update rule $\boldsymbol{v}^{(k+1)} = \frac{\boldsymbol{v}^{(k)}+\lambda\tilde{\boldsymbol{g}}^{(k)}}{\left\|\boldsymbol{v}^{(k)}+\lambda\tilde{\boldsymbol{g}}^{(k)}\right\|_2}$ with $\lambda = G^2/(64\beta^3)$, where $\tilde{\boldsymbol{g}}^{(k)} = \mathrm{proj}_{(\boldsymbol{v}^{(k)})^{\perp}}\hat{\boldsymbol{g}}^{(k)}$ (here $\tilde{\boldsymbol{g}}^{(k)}$ is the $\hat{\boldsymbol{g}}^{(k)}$ with the component on the direction $\boldsymbol{v}^{(k)}$ removed). Note that this procedure only decreases the norm of $\tilde{\boldsymbol{g}}$ (by the Pythagorean theorem). Then, from Lemma 4.19, we have $\boldsymbol{v}^{(k+1)} \cdot \boldsymbol{v}^* \geq \boldsymbol{v}^{(k)} \cdot \boldsymbol{v}^* + G^4/\beta^4$.

The update rule is repeated for at most $O(\beta^4/G^4)$ iterations. From Lemma 4.14, we have that a certificate exists if the angle with the optimal vector is sufficiently small. Putting everything together, our total sample complexity is $N = \tilde{O}\left(\frac{d\beta^4}{b^2G^4}\right)\log(1/\delta)$. It is also clear that the runtime is $\mathrm{poly}(N, d)$, which completes the proof. □

Proof of Theorem 4.5

To prove Theorem 4.5, we will use the iterative algorithm developed in Proposition 4.13 initialized with a uniformly random unit vector $\boldsymbol{v}_0$. It is easy to show that such a random vector will have non-trivial correlation with $\boldsymbol{v}^*$.

Fact 4.22 (see, e.g., Remark 3.2.5 of Vershynin (2018a))**.** *Let $\boldsymbol{v}$ be a unit vector in $\mathbb{R}^d$. For a random unit vector $\boldsymbol{u} \in \mathbb{R}^d$, with constant probability, it holds $|\boldsymbol{v} \cdot \boldsymbol{u}| = \Omega(1/\sqrt{d})$.*

We now present the proof of Theorem 4.5 putting together the machinery developed in the previous subsections.

Proof of Theorem 4.5. As explained in Section 4.3, we are looking for a certificate function $T_{\boldsymbol{w}}(\boldsymbol{x})$ of the form given in Equation (4.3). As argued in Section 4.3, the search for such a certificate function can be simplified by projecting the samples to a $(d-1)$-dimensional subspace via the perspective projection.

From Proposition 4.8, choosing $\rho = O(\theta/\sqrt{d})$, there is a $c = (LR)^{O(1)}$ such that the resulting distribution $D_B^{\pi_{\boldsymbol{w}}}$ is $(2, c\theta/\sqrt{d}, \sqrt{d}/\theta, \beta\sqrt{d}/(c\theta)\log(\sqrt{d}/\theta))$-well-behaved and satisfies the $(\alpha, Ad^{1/2}/(c\theta))$-Tsybakov noise condition.

From Fact 4.22, a random unit vector $\boldsymbol{v} \in \mathbb{R}^{d-1}$ with constant probability satisfies $\boldsymbol{v} \cdot (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} = \Omega(1/\sqrt{d})$. We call this event $\mathcal{E}$.

From Proposition 4.13, conditioning on the event $\mathcal{E}$ and using $\frac{\beta^4}{b^2}\left(\frac{A}{RL}\right)^{O(1/\alpha)} \log(1/\delta)$ samples, with probability $1-\delta$, we get a $(\boldsymbol{v}', R, t_0)$ such that

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D_B^{\pi_{\boldsymbol{w}}}} \left[\mathbb{1}[-R \leq \boldsymbol{v}' \cdot \boldsymbol{x} \leq -t_0]\, y\right] \leq -\left(\theta LR/(Ad)\right)^{O(1/\alpha)}/\beta\,.$$

By inverting the transformation (Lemma 4.9), we get that

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D} [T_{\boldsymbol{w}}(\boldsymbol{x})\boldsymbol{x} \cdot \boldsymbol{w} y] \leq -\left(\theta LR/(Ad)\right)^{O(1/\alpha)}/\beta\,.$$

Overall, we conclude that with constant probability Algorithm 5 returns a valid certificate. Repeating the process $k = O(\log(1/\delta))$ times, we can boost the probability to $1-\delta$. The total number of samples for finding and testing these candidate certificates until we find a correct one with probability at least $1-\delta$ is $N = \left(\frac{d\,A}{\theta LR}\right)^{O(1/\alpha)} \log(1/\delta)$. It is also clear that the runtime is $\mathrm{poly}(N, d)$, which completes the proof. □

4.4 More Efficient Certificate for Log-Concave Distributions

In this section, we present a more efficient certificate algorithm for the important special case of isotropic log-concave distributions. To achieve this, we use Algorithm 5 from the previous section starting from a significantly better initialization vector. To obtain such an initialization, we leverage the structure of log-concave distributions. The main result of this section is the following theorem.

Theorem 4.23 (Certificate for Log-concave Distributions)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$ and is such that $D_{\boldsymbol{x}}$ is isotropic log-concave. Let $\boldsymbol{w}$ be* *a unit vector that satisfies $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \geq \theta$, where $\theta \in (0, \pi]$. There is an algorithm that, given as input $\boldsymbol{w}$, θ, and $N = \mathrm{poly}(d) \cdot \left(\frac{A}{\theta}\right)^{O(1/\alpha^2)} \log(1/\delta)$ samples from D, it runs in $\mathrm{poly}(d, N)$ time, and with probability at least $1 - \delta$ returns a certifying function $T_{\boldsymbol{w}} : \mathbb{R}^d \mapsto \mathbb{R}_+$ such that*

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D} [T_{\boldsymbol{w}}(\boldsymbol{x})\, y \boldsymbol{w} \cdot \boldsymbol{x}] \leq -\left(\frac{\theta}{A}\right)^{O(1/\alpha^2)} . \tag{4.10}$$

In other words, we give an algorithm whose sample complexity and running time as a function of d is a fixed degree polynomial, independent of the noise parameters.

To establish Theorem 4.23, we apply Algorithm 5 starting from a better initialization vector. The main technical contribution of this section is an efficient algorithm to obtain such a vector for log-concave marginals.

Theorem 4.24 (Efficient Initialization for Log-Concave Distributions)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$ and is such that $D_{\boldsymbol{x}}$ is isotropic log-concave. There exists an algorithm that, given an $\epsilon > 0$, a unit vector $\boldsymbol{w}$ such that $\|\boldsymbol{w}^* - \boldsymbol{w}\|_2 = \Theta(\epsilon)$, and $N = \mathrm{poly}(d) \cdot (A/(\alpha\epsilon))^{O(1/\alpha)}$ samples from D, it runs in $\mathrm{poly}(d, N)$ time, and with constant probability returns a unit vector $\boldsymbol{v}$ such that $\boldsymbol{v} \cdot (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \geq (\alpha\epsilon/A)^{O(1/\alpha)}$, where $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ is the component of $\boldsymbol{w}^*$ perpendicular to $\boldsymbol{w}$.*

Intuition and Roadmap of the Proof

Here we sketch the proof of Theorem 4.24 and point to the relevant lemmas in the formal argument (Section 4.4). Given a weight vector $\boldsymbol{w}$ of unit length, our goal is to find a unit vector $\boldsymbol{v}$ that has non-trivial correlation with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$, i.e., such that $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{v}$ is roughly $\epsilon^{1/\alpha}$, where $\boldsymbol{w}^*$ is the optimal halfspace.

Our first step is to condition on a thin band around the current candidate $\boldsymbol{w}$ (similarly to Section 4.3, see Figure 4.1). When the size of the band approaches 0, we get an instance whose separating hyperplane is perpendicular to $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ and has much larger Tsybakov noise rate. After that, we would like (similarly to Section 4.3) to project the points on the subspace $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. Instead of having a zero length band, we will instead take a very thin band. We have already seen in Section 4.3 that we can apply a perspective transformation in order to project the points on $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ and obtain an instance that satisfies the Tsybakov noise condition (with somewhat worse parameters). Unfortunately, for the current setting of log-concave distributions, we cannot use the perspective projection, as it *does not preserve the log-concavity* of the underlying distribution. On the other hand, we know that log-concavity is preserved when we condition on convex sets (such as the thin band we consider here) and when we perform orthogonal projections.

As we have seen (see Figure 4.2), an orthogonal projection will create a “fuzzy" region with arbitrary sign. However, we can control the probability of this “fuzzy" region by taking a sufficiently thin random band. In particular, instead of Tsybakov noise, we will end up with the following noise condition: For some small $\xi > 0$, with probability 2/3 the noise $\eta(\boldsymbol{x})$ is bounded above by $1/2 - \xi$, and with probability roughly $\xi^{\Theta(1)}$ we have $\eta(\boldsymbol{x}) > 1/2$ (this corresponds to the probability of the “fuzzy" region). For the proof of this statement and detailed discussion on how the random band results in this above noise guarantee, see Lemma 4.33.

From this point on, we will be working in the subspace $\boldsymbol{w}^{\perp}$ and assume that the distribution satisfies the aforementioned noise condition. As we have discussed, the marginal distribution on the examples remains log-concave and it is not hard to make its covariance be close to the identity. However, conditioning on the thin slice may result in a distribution with large mean, even though originally the distribution was centered. This is a non-trivial technical issue. We cannot simply translate the distribution to be origin-centered, as this would result in a potentially very biased optimal halfspace. Our proof crucially relies on the assumption of having a distribution that is *nearly* centered and at the same time for the optimal halfspace to have *small bias*. We overcome this obstacle in Step 1 below.

Our approach is as follows:

1. First, we show that there is an efficient rejection sampling procedure that

preserves log-concavity and gives us a distribution that is nearly isotropic (see Definition 4.25). For the algorithm and its detailed proof of correctness, see Algorithm 7 and Lemma 4.36.

2. Then we show the following statement: Under the following assumptions

   (i) the $\boldsymbol{x}$-marginal is nearly isotropic,

   (ii) the optimal halfspace has sufficiently small bias, and

   (iiii) the noise $\eta(\boldsymbol{x})$ is bounded away from $1/2$ with constant probability,

   we can compute in polynomial time a vector $\boldsymbol{v}$ with good correlation to the target $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$. This is established in Proposition 4.30.

We start by describing our algorithm to transform the distribution to nearly isotropic position (Step 1 above). We avoid translating the samples by reweighting the distribution using rejection sampling. To achieve this, we find an approximate stationary point of the non-convex objective $F(\boldsymbol{r}) = \|\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\boldsymbol{x}\max(1, \exp(-\boldsymbol{r}\cdot\boldsymbol{x})]\|_2^2$. Notice that, since this is a non-convex objective as a function of $\boldsymbol{r}$, we can only use (projected) SGD to efficiently find a stationary point. In particular, we show that a γ-stationary point $\boldsymbol{r}$ of $F(\boldsymbol{r})$ will make the above norm of the expectation roughly $O(\gamma)$ (Claim 4.37). Therefore, in time $\mathrm{poly}(d/\gamma)$, we find a reweighting of the initial distribution whose mean is close to $\mathbf{0}$. Given this point $\boldsymbol{r}$, we then perform rejection sampling: We draw $\boldsymbol{x}$ from the initial distribution D and accept it with probability $\max(1, \exp(-\boldsymbol{r}\cdot\boldsymbol{x}))$, i.e., we “shrink" the distribution along the direction $\boldsymbol{r}$.

We now explain how to handle the setting that the distribution is approximately log-concave (Step 2 above). After we make our distribution nearly isotropic, we compute the degree-2 Chow parameters of the distribution, i.e., the vector $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[y\boldsymbol{x}]$ and the matrix $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[y(\boldsymbol{x}\boldsymbol{x}^\intercal - \boldsymbol{I})]$. We show that there exists a degree-2 polynomial $p((\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}\cdot\boldsymbol{x})$ that correlates non-trivially with the labels y (Lemma 4.31). This means that $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ correlates reasonably with the degree-2 Chow parameters. In particular, $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$ has a non-trivial projection on the subspace V spanned by the degree-1 Chow parameters (this is a single vector) and the eigenvectors of the degree-2 Chow matrix with large eigenvalues. Our plan is to return a random unit vector of the subspace V. However, in order for this random vector to have non-trivial correlation with $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$, we also need to show that the dimension of V is not very large.

The last part of our argument shows that V has reasonably small dimension. To prove this, we first show that the dimension of V can be bounded above by the variance of the projection of D onto V, D^{proj_V}, $\mathbf{Var}_{\boldsymbol{x}\sim D^{\mathrm{proj}_V}}[\|\boldsymbol{x}\|_2^2]$. Then we make essential use of a recent "thin-shell" result about log-concave measures that bounds from above $\mathbf{Var}_{\boldsymbol{x}\sim D^{\mathrm{proj}_V}}[\|\boldsymbol{x}\|_2^2]$, see Lemma 4.28 and Lemma 4.32.

Proof of Theorem 4.24

The proof of Theorem 4.24 requires a number of intermediate results. As already mentioned, our initialization algorithm works by restricting D to a narrow band perpendicular to $\boldsymbol{w}$. Unfortunately, this restriction will be log-concave but will no longer be isotropic, even in the directions perpendicular to $\boldsymbol{w}$. However, it will be close in the following sense.

Definition 4.25 ((α,β)-isotropic distribution)**.** *We say that a distribution D is (α,β)-isotropic, if for every unit vector $\boldsymbol{u}\in\mathbb{R}^d$, it holds $|\mathbf{E}_{\boldsymbol{x}\sim D}[\boldsymbol{x}\cdot\boldsymbol{u}]|\leq\alpha$ and $1/\beta\leq\mathbf{E}_{\boldsymbol{x}\sim D}[\boldsymbol{x}\cdot\boldsymbol{u}^2]\leq\beta$.*

Useful Technical Tools. We will require the following standard anti-concentration result for low-degree multivariate polynomials under log-concave distributions.

Lemma 4.26 (Theorem 8 of Carbery and Wright (2001))**.** *Let D be a log-concave distribution on $\mathbb{R}^d$ and $p:\mathbb{R}^d\mapsto R$ be a polynomial of degree at most n. Then there is an absolute constant $C>0$ such that for any $0<q<\infty$ and $t\in\mathbb{R}_+$, it holds $\mathbf{Pr}_{\boldsymbol{x}\sim D}[|p(\boldsymbol{x})|\leq t]\leq Cqt^{1/n}\,\mathbf{E}_{\boldsymbol{x}\sim D}[|p(\boldsymbol{x})|^{q/n}]^{1/q}$.*

The following statement is well-known. (It follows for example by combining Theorem 5.14 of Lovász and Vempala (2007) and Lemma 7 of Klivans et al. (2009b).)

Fact 4.27. *Let z be an isotropic log-concave distribution on $\mathbb{R}^d$ and let $\gamma(\cdot)$ be its density function. There exists a constant $c_d > 0$ such that:*

1. *For any z with $\|z\|_2 \leq c_d$, we have that $\gamma(z) \geq c_d$.*
2. *For any z, we have that $\gamma(z) \leq 1/c_d \exp(-1/c_d \|z\|_2)$.*

Our proof makes essential use of the following "thin-shell" estimate bounding the variance of the norm of any isotropic log-concave random vector.

Lemma 4.28 (Corollary 13 of Lee and Vempala (2017))**.** *Let D be any isotropic log-concave distribution on $\mathbb{R}^d$. We have that $\mathbf{Var}_{x \sim D}[\|x\|_2^2] \leq d^{3/2}$.*

In particular, it is important for our analysis that the above bound is sub-quadratic in d.

Finally, we will require the following simple lemma bounding the sample complexity of approximating the degree-2 Chow parameters of a halfspace under isotropic log-concave distributions.

Lemma 4.29. *Let D be an isotropic log-concave distribution on $\mathbb{R}^d$ and $\widehat{D}_N$ be the empirical distribution obtained from D with $N = \mathrm{poly}(d/\epsilon)$ samples. Then, with high constant probability, we have $\left\|\mathbf{E}_{(x,y)\sim D}[yx] - \mathbf{E}_{(x,y)\sim \widehat{D}_N}[yx]\right\|_2 \leq \epsilon$ and*

$$\left\| \underset{(x,y)\sim D}{\mathbf{E}}[y(xx^\intercal - I)] - \underset{(x,y)\sim \widehat{D}_N}{\mathbf{E}}[y(xx^\intercal - I)] \right\|_F \leq \epsilon.$$

The proof of this lemma can be found in Appendix C.1.

We now have the necessary tools to proceed with our proof. We start by showing how we can find a vector v with non-trivial correlation with $(w^*)^{\perp_w}$ if the marginal distribution is (approximately) isotropic. Since in general this will not hold, we will then need to reduce to the isotropic case.

Proposition 4.30. *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ such that D_x is (α, β)-isotropic log-concave. Let $f(x) = \mathrm{sign}(v^* \cdot x - \theta)$ be such that $\mathbf{Pr}_{(x,y)\sim D}[y \neq f(x) | x] = \eta(x)$,* *where for some $\xi > 0$ we have that $\mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\eta(\boldsymbol{x}) < 1/2 - \xi] \geq 2/3$ and $\mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\eta(\boldsymbol{x}) > 1/2] \leq \xi'$, where ξ' is a constant degree polynomial in ξ[1]. Then, as long as $|\alpha| + |\theta|$ is less than a sufficiently small constant multiple of $1/(\log(1/\xi))$, there exists an algorithm with sample complexity and runtime $\mathrm{poly}(d/\xi)$ that with constant probability returns a unit vector $\boldsymbol{v} \in \mathbb{R}^d$ such that $\boldsymbol{v} \cdot \boldsymbol{v}^* > \mathrm{poly}(\xi)$.*

Proof. For clarity of the analysis, we begin by presenting our algorithm for the case that $D_{\boldsymbol{x}}$ is exactly isotropic log-concave. We then show how the algorithm and its analysis can be modified for the approximate log-concave setting.

Our algorithm is fairly simple. We compute high-precision estimates $\boldsymbol{T}_1'$ and $\boldsymbol{T}_2'$ of the vector $\boldsymbol{T}_1 := \mathbf{E}_{(\boldsymbol{x},y)\sim D}[y\boldsymbol{x}]$ and the matrix $\boldsymbol{T}_2 := \mathbf{E}_{(\boldsymbol{x},y)\sim D}[y(\boldsymbol{x}\boldsymbol{x}^\intercal - \boldsymbol{I})]$ respectively. This can be easily done by taking $\mathrm{poly}(d/\epsilon)$ samples from D and using the empirical estimates (see Lemma 4.29). We then define V to be the subspace spanned by $\boldsymbol{T}_1$ and the eigenvectors of $\boldsymbol{T}_2$ whose eigenvalue has absolute value at least 2ζ, for ζ some sufficiently large constant power of ξ. The algorithm returns a uniform random unit vector $\boldsymbol{v}$ from V.

It is clear that the above algorithm has polynomial sample complexity and runtime. We need to show that with constant probability it holds that $\boldsymbol{v} \cdot \boldsymbol{v}^* > \mathrm{poly}(\xi)$. The desired statement will follow by establishing the following two claims:

1. The size of the projection of $\boldsymbol{v}^*$ onto V is at least $\mathrm{poly}(\xi)$.

2. The dimension of V is at most $\mathrm{poly}(1/\xi)$.

The desired result then follows by noting that the median value of $|\boldsymbol{v}^* \cdot \boldsymbol{v}|$ is on the order of $\left\|\mathrm{proj}_V(\boldsymbol{v}^*)\right\|_2 / \sqrt{d(V)}$, and observing that the sign of the inner product is independent of its size.

To establish the first claim, we prove the following lemma for isotropic log-concave distributions.

[1] It is not difficult to verify that $\xi' = \Theta(\xi^3)$ suffices.

Lemma 4.31. *Let $D_{\boldsymbol{x}}$ be isotropic log-concave. There exists a degree-2 polynomial $p : \mathbb{R} \to \mathbb{R}$ such that $\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[p(\boldsymbol{v}^* \cdot \boldsymbol{x})] = 0$, $\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[p(\boldsymbol{v}^* \cdot \boldsymbol{x})^2] = 1$, and $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[y\, p(\boldsymbol{v}^* \cdot \boldsymbol{x})] = \Omega(\xi)$.*

Proof. We consider the polynomial

$$q(x) = (x - \theta)(x + 1/\theta) = x^2 + (1/\theta - \theta)x - 1$$

and we set $p(x) = q(x)/\sqrt{\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[q(\boldsymbol{v}^* \cdot \boldsymbol{x})^2]}$. It is easy to see that $\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[p(\boldsymbol{v}^* \cdot \boldsymbol{x})] = 0$ and $\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[p(\boldsymbol{v}^* \cdot \boldsymbol{x})^2] = 1$. To show that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[yp(\boldsymbol{v}^* \cdot \boldsymbol{x})] = \Omega(\xi)$, we note that

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[yp(\boldsymbol{v}^* \cdot \boldsymbol{x})] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[(1 - 2\eta(\boldsymbol{x}))f(\boldsymbol{x})p(\boldsymbol{v}^* \cdot \boldsymbol{x})] \,.$$

We observe that if $|\boldsymbol{v}^* \cdot \boldsymbol{x}| \leq 1/|\theta|$, then $\mathrm{sign}(p(\boldsymbol{v}^* \cdot \boldsymbol{x})) = f(\boldsymbol{x})$, where $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{v}^* \cdot \boldsymbol{x} - \theta)$. Thus, unless $|\boldsymbol{v}^* \cdot \boldsymbol{x}| > 1/|\theta|$ or $\eta(\boldsymbol{x}) > 1/2$ (which happens with probability at most ξ', a sufficiently high power of ξ), we have that $(1 - 2\eta(\boldsymbol{x}))f(\boldsymbol{x})p(\boldsymbol{v}^* \cdot \boldsymbol{x}) \geq 0$ except with probability at most ξ'.

Let $I(\boldsymbol{x})$ denote the indicator of the event $(1 - 2\eta(\boldsymbol{x}))f(\boldsymbol{x})p(\boldsymbol{v}^* \cdot \boldsymbol{x}) < 0$. We have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y\, p(\boldsymbol{v}^* \cdot \boldsymbol{x})] &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|(1 - 2\eta(\boldsymbol{x}))p(\boldsymbol{v}^* \cdot \boldsymbol{x})|] - 2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|(1 - 2\eta(\boldsymbol{x}))p(\boldsymbol{v}^* \cdot \boldsymbol{x})|I(\boldsymbol{x})] \\
&\geq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|(1 - 2\eta(\boldsymbol{x}))p(\boldsymbol{v}^* \cdot \boldsymbol{x})|] - 2\sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[I^2(\boldsymbol{x})] \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[p(\boldsymbol{v}^* \cdot \boldsymbol{x})^2]} \\
&\geq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|(1 - 2\eta(\boldsymbol{x}))p(\boldsymbol{v}^* \cdot \boldsymbol{x})|] - 2\sqrt{\xi'}.
\end{aligned}$$

Recall that by assumption there is at least a $2/3$ probability that $(1 - 2\eta(\boldsymbol{x})) \geq \xi$.

By anti-concentration of Gaussian polynomials, Lemma 4.26, applied for $q = 4$ and $n = 2$, we have that $\mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|p(\boldsymbol{v}^* \cdot \boldsymbol{x})| \leq t] = O(\sqrt{t})$. Thus, for small enough t, we have that $|p(\boldsymbol{v}^* \cdot \boldsymbol{x})| = \Omega(1)$ with probability at least $2/3$. Therefore, with probability at least $1/3$ both statements hold. Since $|1 - 2\eta(\boldsymbol{x})||p(\boldsymbol{v}^* \cdot \boldsymbol{x})| \geq 0$ for all $\boldsymbol{x}$, we have that $\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|1 - 2\eta(\boldsymbol{x})||p(\boldsymbol{v}^* \cdot \boldsymbol{x})|] = \Omega(\xi)$. This completes our proof. □

Given Lemma 4.31, it is not hard to see that $p(\boldsymbol{v}^* \cdot \boldsymbol{x}) = a(\boldsymbol{v}^* \cdot \boldsymbol{x}) + b((\boldsymbol{v}^* \cdot \boldsymbol{x})^2 - 1)$ for some real numbers a and b with $|a| + |b| = \Theta(1)$. We note that there is another way to compute $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[yp(\boldsymbol{v}^* \cdot \boldsymbol{x})]$ relating it to $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$. In particular, we can write

$$\begin{aligned}\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y\,p(\boldsymbol{v}^* \cdot \boldsymbol{x})] &= a \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y(\boldsymbol{v}^* \cdot \boldsymbol{x})] + b \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y((\boldsymbol{v}^* \cdot \boldsymbol{x})^2 - 1)] \\ &= a\boldsymbol{v}^* \cdot \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y\boldsymbol{x}] + b \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y((\boldsymbol{v}^*)^\intercal(\boldsymbol{x}\boldsymbol{x}^\intercal - \boldsymbol{I})\boldsymbol{v}^*)] \\ &= a\boldsymbol{v}^* \cdot \boldsymbol{T}_1 + b(\boldsymbol{v}^*)^\intercal \boldsymbol{T}_2 \boldsymbol{v}^*.\end{aligned}$$

Thus, Lemma 4.31 implies that either $|\boldsymbol{v}^* \cdot \boldsymbol{T}_1| = \Omega(\zeta)$ or $|(\boldsymbol{v}^*)^\intercal \boldsymbol{T}_2 \boldsymbol{v}^*| = \Omega(\zeta)$.

Assuming that $\boldsymbol{T}_1'$ and $\boldsymbol{T}_2'$ estimate $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ to error less than this quantity, i.e., $O(\zeta)$, the above implies that either $\boldsymbol{v}^* \cdot \boldsymbol{T}_1' = \Omega(\zeta)$ or $(\boldsymbol{v}^*)^\intercal \boldsymbol{T}_2' \boldsymbol{v}^* = \Omega(\zeta)$. In the former case, we have that $\left\|\mathrm{proj}_V(\boldsymbol{v}^*)\right\|_2 \geq |\boldsymbol{v}^* \cdot \boldsymbol{T}_1| = \Omega(\zeta)$. In the latter case, we note that since V contains the span of all eigenvectors of $\boldsymbol{T}_2'$ with eigenvalue having absolute value at least ζ, it holds that $|(\boldsymbol{v}^*)^\intercal \boldsymbol{T}_2' \boldsymbol{v}^*| \leq \zeta + \|\boldsymbol{T}_2'\|_2 \left\|\mathrm{proj}_V(\boldsymbol{v}^*)\right\|_2$. This will imply that in this case as well we have that $\left\|\mathrm{proj}_V(\boldsymbol{v}^*)\right\|_2 = \Omega(\zeta)$, if $\|\boldsymbol{T}_2'\|_2$ is $O(1)$. To show this, we note that for any unit vector $\boldsymbol{v}$, we have

$$\begin{aligned}\boldsymbol{v}^\intercal \boldsymbol{T}_2 \boldsymbol{v} &= \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y(\boldsymbol{v}^\intercal(\boldsymbol{x}\boldsymbol{x}^\intercal - \boldsymbol{I})\boldsymbol{v})] = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y(\boldsymbol{v} \cdot \boldsymbol{x}^2 - 1)] \\ &\leq \sqrt{\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y^2] \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[(\boldsymbol{v} \cdot \boldsymbol{x}^2 - 1)^2]} = O(1)\ .\end{aligned}$$

This completes the proof that the projection of $\boldsymbol{v}^*$ onto V has size at least $\mathrm{poly}(\zeta)$.

It remains to show that the dimension of V is at most $\mathrm{poly}(\zeta)$. We prove the following lemma:

Lemma 4.32. *We have that $d(V) = O(\zeta^{-4})$.*

Proof. Let V_+ denote the subspace spanned by the eigenvectors of $\boldsymbol{T}_2'$ with eigenvalue at least ζ. Let V_- denote the subspace spanned by eigenvectors of eigenvalue at most $-\zeta$. Clearly $d(V) \leq d(V_+) + d(V_-) + 1$. We will show that $d(V_+) = O(\zeta^{-4})$ and the bound on $d(V_-)$ will follow symmetrically.

Let $m = d(V_+)$ and let $\boldsymbol{P}$ be the projection matrix that maps a vector $\boldsymbol{z}$ onto V_+. Since $\boldsymbol{T}_2'$ is sufficiently close to $\boldsymbol{T}_2$, the restriction of $\boldsymbol{T}_2'$ to V_+ will have all of its eigenvalues at least $\zeta/2$. Therefore, it holds that

$$\begin{aligned} m\zeta/2 \leq \operatorname{tr}(\boldsymbol{P}\boldsymbol{T}_2) &= \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y \operatorname{tr}(\boldsymbol{P}(\boldsymbol{x}\boldsymbol{x}^\intercal - \boldsymbol{I}))] = \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y(\|\boldsymbol{P}\boldsymbol{x}\|_2^2 - m)] \\ &\leq \sqrt{\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y^2] \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[(\|\boldsymbol{P}\boldsymbol{x}\|_2^2 - m)^2]} = \sqrt{\mathbf{Var}[\|\boldsymbol{P}\boldsymbol{x}\|_2^2]} \;. \end{aligned}$$

In other words, we have that

$$m^2\zeta^2 \leq 4\mathbf{Var}[\|\boldsymbol{P}\boldsymbol{x}\|_2^2] \;.$$

To conclude the proof, observe that $\boldsymbol{P}\boldsymbol{x}$ is a log-concave distribution in m dimensions, since projections preserve log-concavity. From Lemma 4.28, we have that $\mathbf{Var}[\|\boldsymbol{P}\boldsymbol{x}\|_2^2] = O(m^{3/2})$ and together with the above, we obtain that $m = O(\zeta^{-4})$. This completes our proof. □

Thus far, we have shown the desired claim if the distribution is in isotropic position, $\theta = O(1/\log(1/\xi))$, and we have access to sufficiently accurate approximations $\boldsymbol{T}_1', \boldsymbol{T}_2'$ to the degree-2 Chow parameters with accuracy $\zeta/2$. To handle the case that the distribution D is (α,β)-isotropic, we can let $\boldsymbol{z} \sim D_{\boldsymbol{z}}'$, where $\boldsymbol{z} = \mathbf{Cov}[\boldsymbol{x}]^{-1/2}(\boldsymbol{x} - \mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\boldsymbol{x}])$, be an isotropic log-concave distribution. We need to show that if we have good approximations of $\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\boldsymbol{x}]$ and $\mathbf{Cov}[\boldsymbol{x}]$, we can compute $O(\zeta)$-approximations to $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ for $\boldsymbol{z}$ (i.e., $\mathbf{E}_{(\boldsymbol{z},y)\sim D'}[y\boldsymbol{z}]$ and $\mathbf{E}_{(\boldsymbol{z},y)\sim D'}[y(\boldsymbol{z}\boldsymbol{z}^\intercal - \boldsymbol{I})]$). By taking $\mathrm{poly}(d/\zeta)$ samples, we can compute $\widehat{\boldsymbol{m}}$ and $\widehat{\boldsymbol{M}}$ such that $\|\widehat{\boldsymbol{m}} - \mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\boldsymbol{x}]\|_2 \leq \zeta/16$ and $\left\|\widehat{\boldsymbol{M}} - \mathbf{Cov}[\boldsymbol{x}]\right\|_2 \leq \zeta/16$. Let $\widehat{\boldsymbol{z}} = \widehat{\boldsymbol{M}}^{-1/2}(\boldsymbol{x} - \widehat{\boldsymbol{m}})$. Then we have that $\|\widehat{\boldsymbol{z}} - \boldsymbol{z}\|_2 \leq \zeta/4$. Thus, we obtain that

$$\left\|\mathop{\mathbf{E}}_{(\boldsymbol{z},y)\sim D'}[y\boldsymbol{z}] - \mathop{\mathbf{E}}_{(\boldsymbol{z},y)\sim D'}[y\widehat{\boldsymbol{z}}]\right\|_2 \leq \mathop{\mathbf{E}}_{(\boldsymbol{z},y)\sim D'}[\|\boldsymbol{z} - \widehat{\boldsymbol{z}}\|_2] \leq \zeta/4$$

and similarly that

$$\left\| \mathop{\mathbf{E}}_{(z,y)\sim D'}[y(\boldsymbol{z}\boldsymbol{z}^{\intercal} - \boldsymbol{I})] - \mathop{\mathbf{E}}_{(z,y)\sim D'}[y(\widehat{\boldsymbol{z}}\widehat{\boldsymbol{z}}^{\intercal} - \boldsymbol{I})] \right\|_2 \leq \zeta/4 \, .$$

By approximating the degree-2 Chow parameters $\boldsymbol{T}_1, \boldsymbol{T}_2$ to accuracy $\zeta/4$, we obtain overall error $\zeta/2$.

We note that $(\boldsymbol{z}, y)$ satisfies our assumptions for the function

$$f'(\boldsymbol{x}) = \operatorname{sign}\left((\boldsymbol{v}^*)^{\intercal} \mathbf{Cov}[\boldsymbol{x}]^{1/2} \boldsymbol{x} - (\theta - \langle \boldsymbol{v}^*, \mathop{\mathbf{E}}_{x\sim D_x}[\boldsymbol{x}] \rangle) \right) \, .$$

From our assumptions, we have that $|\theta - \boldsymbol{v}^* \cdot \mathbf{E}_{x\sim D_x}[\boldsymbol{x}]| = O(1/\log(1/\xi))$. Using the aforementioned algorithm for $\boldsymbol{z}$, this allows us to compute a $\boldsymbol{v}$ so that with constant probability $\boldsymbol{v}^{\intercal}\mathbf{Cov}[\boldsymbol{x}]^{1/2}\boldsymbol{v}^* \geq \operatorname{poly}(\xi)$, or $\mathbf{Cov}[\boldsymbol{x}]^{1/2}\boldsymbol{v} \cdot \boldsymbol{v}^* \geq \operatorname{poly}(\xi)$. This completes the proof. □

Thus far, we have dealt with the case that the mean of our log-concave distribution is sufficiently close to zero. As already mentioned, this property will not hold in general after projection. The following important lemma shows that by conditioning on a random thin band before projecting onto $\boldsymbol{w}^{\perp}$, we obtain a log-concave distribution whose mean has small distance from the origin. Moreover, we show that the noise condition of the instance after we perform this transformation satisfies the assumptions of Proposition 4.30. We note that this is the step that crucially relies on picking a *random* thin band.

Lemma 4.33 (Properties of Transformed Instance)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$ and is such that $D_{\boldsymbol{x}}$ is isotropic log-concave. Fix $\epsilon > 0$ and unit vector $\boldsymbol{w}$ such that $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \Theta(\epsilon)$. Let s be a sufficiently small multiple of ϵ^2. Set $\xi = (\Theta(s/A))^{1/\alpha}$ and $s' = \Theta(\xi^3 \, s \, \epsilon)$. Pick x_0 uniformly at random from $[s, 2s]$ and define the random band $B_{x_0} = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} \cdot \boldsymbol{w} \in [x_0, x_0 + s']\}$.*

[2] We need s to be smaller than the absolute constant of Fact 4.27 for dimension $d = 2$.

Define the distribution $D^\perp = D_{B_{x_0}}^{\mathrm{proj}_{w^\perp}}$*, the classifier* $f^\perp(\boldsymbol{x}^\perp) = \mathrm{sign}(x_0/\tan\theta + \boldsymbol{x}^\perp \cdot (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}})$*, and the noise function*

$$\eta^\perp(\boldsymbol{x}^\perp) = \mathop{\mathbf{Pr}}_{(\boldsymbol{z},y)\sim D^\perp}[y \neq f^\perp(\boldsymbol{z}) | \boldsymbol{z} = \boldsymbol{x}^\perp] \,.$$

Then $D^\perp$ *is an* $(O(1), O(1))$*-isotropic log-concave distribution and, with probability at least* 99%*, satisfies the following noise condition:*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}^\perp\sim D_{\boldsymbol{x}}^\perp}[\eta^\perp(\boldsymbol{x}^\perp) \leq 1/2 - \xi] \geq 2/3 \quad \textit{and} \quad \mathop{\mathbf{Pr}}_{\boldsymbol{x}^\perp\sim D_{\boldsymbol{x}}^\perp}[\eta^\perp(\boldsymbol{x}^\perp) \geq 1/2] \leq \xi^3 \,.$$

Proof. We first calculate how far the distribution $D_{B_{x_0}}^{\mathrm{proj}_{w^\perp}}$ is from being isotropic. Since our final goal is to have a distribution whose mean is arbitrarily close to $\mathbf{0}$, we need to bound the distance from $\mathbf{0}$ of the mean of the distribution obtained after we condition on B and project onto $\boldsymbol{w}^\perp$. The following claim shows that the mean and covariance of $D_{B_{x_0}}^{\mathrm{proj}_{w^\perp}}$ differ from these of the initial distribution D only by constant factors (additive for the mean and multiplicative for the covariance).

Claim 4.34. $D_{B_{x_0}}^{\mathrm{proj}_{w^\perp}}$ *is* $(O(1), O(1))$*-isotropic.*

The proof of Claim 4.34 relies on Fact 4.27 and is given in Appendix C.1.

It remains to prove how the noise condition changes via the transformation. In our argument, we are going to repeatedly use the following anti-concentration, and anti-anti-concentration properties of log-concave distributions that follow directly from Fact 4.27. In particular, for every interval $[a, b]$, we have that:

1. $\mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\boldsymbol{x} \cdot \boldsymbol{v} \in [a, b]] = O(b - a)$ (anti-concentration).

2. If $|a|, |b|$ are smaller than some absolute constant (see Fact 4.27), then it also holds that $\mathbf{Pr}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\boldsymbol{x} \cdot \boldsymbol{v} \in [a, b]] = \Omega(b - a)$ (anti-anti-concentration).

Using the condition $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \Theta(\epsilon)$, we can assume that $\boldsymbol{w}^* = \lambda_1 \boldsymbol{w} + \lambda_2 (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$, where $\lambda_1 = \cos\theta$ and $\lambda_2 = \sin\theta$. It holds $|\lambda_1| = 1 - \Theta(\epsilon)$ and $\lambda_2 = \Theta(\epsilon)$. Next we set $\boldsymbol{x} = (\boldsymbol{x}_{\boldsymbol{w}}, \boldsymbol{x}^{\perp})$, where $\boldsymbol{x}_{\boldsymbol{w}} = \boldsymbol{w} \cdot \boldsymbol{x}$ and $\boldsymbol{x}^{\perp}$ is the projection of $\boldsymbol{x}$ on the subspace $\boldsymbol{w}^{\perp}$.

For some $\zeta \in (0,1)$, set $\xi = (\zeta s/A)^{1/\alpha}/2$. In what follows, we shall see that ζ is some absolute constant, i.e., that $\xi = (\Theta(s/A))^{1/\alpha}$. Recall that the orthogonal projection on $\boldsymbol{w}^{\perp}$ creates a "fuzzy" region, i.e., a region where $\eta^{\perp}(\boldsymbol{x}^{\perp}) \geq 1/2$, see Figure 4.2. We first show that we can control the probability that we get points inside this "fuzzy" region. More, formally we will show that

$$\mathop{\mathbf{Pr}}_{(\boldsymbol{x}^{\perp}, y) \sim D^{\perp}} [\eta^{\perp}(\boldsymbol{x}^{\perp}) \geq 1/2] \leq \xi^3 \,. \tag{4.11}$$

Notice that in this part of the proof the randomness of x_0 is not important and we are able to establish a stronger claim that holds for every band B_{x_0}. Conditioned on $\boldsymbol{x} \in B_{x_0}$, i.e., $\boldsymbol{x}_{\boldsymbol{w}} \in [x_0, x_0 + s']$, it holds that

$$\boldsymbol{w}^* \cdot \boldsymbol{x} = \lambda_1 \boldsymbol{x}_{\boldsymbol{w}} + \lambda_2 (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x}^{\perp} = \lambda_1 x_0 + \lambda_2 (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x}^{\perp} + \rho s' \,,$$

for some $\rho \in [-1, 1]$ (recall that $|\lambda_1| \leq 1$). Notice that when $|\lambda_1 x_0 + \lambda_2 (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x}^{\perp}| > s'$, $f^{\perp}(\boldsymbol{x}^{\perp})$ is equal to the sign of $\boldsymbol{w}^* \cdot \boldsymbol{x}$ (recall that $\lambda_2 > 0$), and therefore we are outside of the fuzzy region, see Figure 4.2. Thus, we need to bound the probability of the event $|\lambda_1 x_0 + \lambda_2 (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x}^{\perp}| \leq s'$, or equivalently $(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x}^{\perp} \in [-\lambda_1 x_0 - s', -\lambda_1 x_0 + s'] =: I_{x_0}^{s'}$. We have that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}^{\perp} \sim D_{\boldsymbol{x}}^{\perp}} \left[(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x}^{\perp} \in I_{x_0}^{s'} \right] = \frac{\mathbf{Pr}_{\boldsymbol{x} \sim D_{\boldsymbol{x}}} \left[(\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x} \in I_{x_0}^{s'} \right]}{\mathbf{Pr}_{\boldsymbol{x} \sim D_{\boldsymbol{x}}} [\boldsymbol{x} \in B_{x_0}]} = O(s'/(\lambda_2 s)) \leq \xi^3 \,,$$

where to bound the numerator we used the anti-concentration property of D, Property 1, for the interval $I_{x_0}^{s'}$ of length s', and to bound the denominator we used the anti-anti-concentration, Property 2. The last inequality holds because we have that $\lambda_2 = \Theta(\epsilon)$ and also, from the assumptions of the lemma, we have $s' = \Theta(\xi^3 s \epsilon)$. This proves (4.11).

Now we deal with the case where $|\lambda_1 x_0 + \lambda_2 (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x}^{\perp}| \leq s'$, i.e., we are in the non-fuzzy region of Figure 4.2. This is where the randomness of x_0 helps us control the probability that the noise is close to $1/2$. Recall that,

$$\eta^{\perp}(\boldsymbol{x}^{\perp}) = \mathop{\mathbf{Pr}}_{(\boldsymbol{x}^{\perp},y)\sim D^{\perp}}\left[y \neq \mathrm{sign}((\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \cdot \boldsymbol{x}^{\perp} + x_0)\right] = \int_{x_0}^{x_0+s'} \eta(\boldsymbol{x}_{\boldsymbol{w}}, \boldsymbol{x}^{\perp})\gamma(\boldsymbol{x}_{\boldsymbol{w}}|\boldsymbol{x}^{\perp})\mathrm{d}\boldsymbol{x}_{\boldsymbol{w}}\ ,$$

where $\gamma(\boldsymbol{x}_{\boldsymbol{w}}|\boldsymbol{x}^{\perp})$ is the density of $D_{B_{x_0}}$ conditioned on $\boldsymbol{x}^{\perp}$, that is $\gamma(\boldsymbol{x}_{\boldsymbol{w}}|\boldsymbol{x}^{\perp}) = \gamma(\boldsymbol{x}_{\boldsymbol{w}}, \boldsymbol{x}^{\perp}) / \int \gamma(\boldsymbol{x}_{\boldsymbol{w}}, \boldsymbol{x}^{\perp})\mathrm{d}\boldsymbol{x}_{\boldsymbol{w}}$, and γ is the density of the $\boldsymbol{x}$-marginal of $D_{B_{x_0}}$. Note that, from Lemma 4.11, it follows that $\mathbf{Pr}[\eta(\boldsymbol{x}) \geq 1/2 - t \mid \boldsymbol{x}_{\boldsymbol{w}} \in [s, 2s + s']] = O(\frac{A}{s}t^{\alpha})$. Therefore, $\mathbf{Pr}[\eta(\boldsymbol{x}) > 1/2 - 2\xi \mid \boldsymbol{x}_{\boldsymbol{w}} \in [s, 2s + s']] \leq \zeta$, and it remains to prove that $\mathbf{Pr}[\eta^{\perp}(\boldsymbol{x}^{\perp}) > 1/2 - \xi]$ is at most a small constant multiple of ζ with high constant probability.

To prove this, let $M(\boldsymbol{x})$ be the indicator of the event $\eta(\boldsymbol{x}) > 1/2 - 2\xi$ and consider the random variable $Y = \int_{x_0}^{x_0+s'} M(\boldsymbol{x}_{\boldsymbol{w}}, \boldsymbol{x}^{\perp})\gamma(\boldsymbol{x}_{\boldsymbol{w}}|\boldsymbol{x}^{\perp})\mathrm{d}\boldsymbol{x}_{\boldsymbol{w}}$. Observe that the randomness of Y is over the randomly chosen x_0 and $\boldsymbol{x}^{\perp}$. We will first show that the probability that the noise function $\eta^{\perp}(\boldsymbol{x}^{\perp})$ exceeds $1/2 - \xi$ can be bounded above by the probability that the random variable Y exceeds $1/2$, that is

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}^{\perp}\sim D_{\boldsymbol{x}}^{\perp}, x_0}[\eta^{\perp}(\boldsymbol{x}^{\perp}) > 1/2 - \xi] \leq \mathop{\mathbf{Pr}}_{\boldsymbol{x}^{\perp}\sim D_{\boldsymbol{x}}^{\perp}, x_0}[Y \geq 1/2]\,. \tag{4.12}$$

In fact, we show a stronger statement than Equation (4.12) that holds for any fixed $x_0 \in [s, 2s]$. To see this, let $\eta'(\boldsymbol{x}) = 1/2 - 2\xi(1 - M(\boldsymbol{x}))$ and notice that $\eta'(\boldsymbol{x}) \geq \eta(\boldsymbol{x})$ for every $\boldsymbol{x}$. Then, it holds

$$\begin{aligned}1/2 - \xi < \eta^{\perp}(\boldsymbol{x}^{\perp}) &= \int_{x_0}^{x_0+s'} \eta(\boldsymbol{x}_{\boldsymbol{w}}, \boldsymbol{x}^{\perp})\gamma(\boldsymbol{x}_{\boldsymbol{w}}|\boldsymbol{x}^{\perp})\mathrm{d}\boldsymbol{x}_{\boldsymbol{w}} \leq \int_{x_0}^{x_0+s'} \eta'(\boldsymbol{x}_{\boldsymbol{w}}, \boldsymbol{x}^{\perp})\gamma(\boldsymbol{x}_{\boldsymbol{w}}|\boldsymbol{x}^{\perp})\mathrm{d}\boldsymbol{x}_{\boldsymbol{w}} \\ &= 1/2 - 2\xi + 2\xi \int_{x_0}^{x_0+s'} M(\boldsymbol{x}_{\boldsymbol{w}}, \boldsymbol{x}^{\perp})\gamma(\boldsymbol{x}_{\boldsymbol{w}}|\boldsymbol{x}^{\perp})\mathrm{d}\boldsymbol{x}_{\boldsymbol{w}}\ ,\end{aligned}$$

which is equivalent to $\int_{x_0}^{x_0+s'} M(\boldsymbol{x}_{\boldsymbol{w}}, \boldsymbol{x}^{\perp})\gamma(\boldsymbol{x}_{\boldsymbol{w}}|\boldsymbol{x}^{\perp})\mathrm{d}\boldsymbol{x}_{\boldsymbol{w}} = Y \geq 1/2$.

Our next step is to bound from above the probability of the event $Y \geq 1/2$. For convenience, let $\phi(\boldsymbol{x})$ be the density of the initial isotropic log-concave marginal $D_{\boldsymbol{x}}$.

Thus, we have $\gamma(\boldsymbol{x_w}, \boldsymbol{x}^\perp) = \phi(\boldsymbol{x_w}, \boldsymbol{x}^\perp) / \mathbf{Pr}_D[B_{x_0}]$. Moreover, set $Q = \min_{x_0 \in [s,2s]} \mathbf{Pr}_D[B_{x_0}]$ and recall that from Properties 1, 2 we have that for any $x_0 \in [s, 2s]$ it holds $\mathbf{Pr}_D[B_{x_0}] = \Theta(s')$, and thus $Q = \Theta(s')$. We can bound from above the expectation of Y, i.e.,

$$\begin{aligned}
\mathbf{E}[Y] &= \int_s^{2s} \frac{1}{s} \int_{x_0}^{x_0+s'} \int_{\boldsymbol{w}^\perp} M(\boldsymbol{x_w}, \boldsymbol{x}^\perp) \frac{\phi(\boldsymbol{x_w}, \boldsymbol{x}^\perp)}{\mathbf{Pr}_D[B_{x_0}]} \, \mathrm{d}\boldsymbol{x}^\perp \, \mathrm{d}\boldsymbol{x_w} \, \mathrm{d}x_0 \\
&\le \frac{1}{sQ} \int_s^{2s} \int_{x_0}^{x_0+s'} \int_{\boldsymbol{w}^\perp} M(\boldsymbol{x_w}, \boldsymbol{x}^\perp) \phi(\boldsymbol{x_w}, \boldsymbol{x}^\perp) \, \mathrm{d}\boldsymbol{x}^\perp \, \mathrm{d}\boldsymbol{x_w} \, \mathrm{d}x_0 \\
&\le \frac{s'}{sQ} \int_s^{2s+s'} \int_{\boldsymbol{w}^\perp} M(\boldsymbol{x_w}, \boldsymbol{x}^\perp) \phi(\boldsymbol{x_w}, \boldsymbol{x}^\perp) \mathrm{d}\boldsymbol{x}^\perp \, \mathrm{d}\boldsymbol{x_w} \\
&\le \frac{s' \, \mathbf{Pr}_D[\boldsymbol{x_w} \in [s, 2s+s']]}{sQ} \mathbf{Pr}[\eta(\boldsymbol{x}) > 1/2 - 2\xi | \boldsymbol{x_w} \in [s, 2s+s']] \le \frac{s'}{Q} \zeta = O(\zeta) \; ,
\end{aligned}$$

where to get the third inequality we used the fact that for any non-negative function $g(t)$ it holds

$$\int_s^{2s} \int_u^{u+s'} g(t) \mathrm{d}t \mathrm{d}u = \int_0^{s'} \int_{s+u}^{2s+u} g(t) \mathrm{d}t \mathrm{d}u \le \int_0^{s'} \int_s^{2s+s'} g(t) \mathrm{d}t \mathrm{d}u = s' \int_s^{2s+s'} g(t) \mathrm{d}t \, .$$

The final inequality follows from Properties 1 and 2. By Markov's inequality, we obtain $\mathbf{Pr}[Y \ge 1/2] = O(\zeta)$. Therefore, combining this bound with Equation (4.12), we obtain the probability that $\eta(\boldsymbol{x}^\perp) > 1/2 - \xi$ is at most

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}^\perp \sim D_{\boldsymbol{x}}^\perp, x_0} [\eta^\perp(\boldsymbol{x}^\perp) > 1/2 - \xi] \le \mathop{\mathbf{Pr}}_{\boldsymbol{x}^\perp \sim D_{\boldsymbol{x}}^\perp, x_0} [Y \ge 1/2] = O(\zeta) \; .$$

So, choosing ζ to be a sufficiently small absolute constant, we get that $\mathbf{Pr}_{(\boldsymbol{x}^\perp, y) \sim D^\perp}[\eta^\perp(\boldsymbol{x}^\perp) \ge 1/2] \le \xi^3$ and $\mathbf{Pr}_{\boldsymbol{x}^\perp}[\eta^\perp(\boldsymbol{x}^\perp) > 1/2 - \xi] \le 1/3$ with probability at least 99%. This completes the proof. □

We next show how to efficiently decrease the mean of a nearly identity covariance log-concave distribution and make it arbitrary close to zero. We achieve this by further conditioning. In particular, we show that we can efficiently find a reweighting of the conditional distribution on $\boldsymbol{x}^\perp$ such that it is approximately mean zero isotropic. The high-level idea to achieve this is, for some vector $\boldsymbol{r}$, to run rejection sampling, where $\boldsymbol{x}$ is kept with probability $\min(1, \exp(-\langle \boldsymbol{r}, \boldsymbol{x} \rangle))$. The problem is then to find $\boldsymbol{r}$. We do this by finding an approximate stationary point of an appropriately defined constrained non-convex optimization problem.

We will use the following standard fact about the convergence of projected stochastic gradient descent (PSGD) to stationary points of smooth non-convex functions. Consider the constrained optimization setting of minimizing a (differentiable) function F in the set $\mathcal{X}$. In this setting, a point $\boldsymbol{x}$ is called ϵ-stationary, $\epsilon > 0$, if for all $\boldsymbol{u} \in \mathcal{X}$ it holds $\nabla F(\boldsymbol{x}) \cdot \boldsymbol{u} - \boldsymbol{x} \geq -\epsilon \left\| \boldsymbol{u} - \boldsymbol{x} \right\|_2$. Note that if $\boldsymbol{x} \in \mathrm{int}(\mathcal{X})$, i.e., $\boldsymbol{x}$ is not on the boundary of $\mathcal{X}$, this inequality is equivalent to $\|\nabla F(\boldsymbol{x})\|_2 \leq \epsilon$.

Fact 4.35 (see, e.g., Ghadimi et al. (2016), Corollary 4 and Equations (4.23) and (4.25))**.** *Let D be a distribution supported on $\mathbb{R}^d$. Let $F : \mathcal{X} \mapsto \mathbb{R}$ be an L-smooth differentiable function on a compact convex set $\mathcal{X} \subset \mathbb{R}^d$ with diameter D. Let $\boldsymbol{g} : \mathcal{X} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ be such that $\mathbf{E}_{\boldsymbol{x} \sim D}[\boldsymbol{g}(\boldsymbol{r}, \boldsymbol{x})] = \nabla F(\boldsymbol{r})$ and $\mathbf{E}_{\boldsymbol{x} \sim D}[\|\boldsymbol{g}(\boldsymbol{r}, \boldsymbol{x})\|_2^2] \leq \sigma^2$, for some $\sigma > 0$. Then randomized projected SGD uses $T = O(\sigma^3 D^2 L^2 / \epsilon^2)$ samples from D, runs $\mathrm{poly}(T, d)$ time, and returns a point $\boldsymbol{r}'$ such that with probability at least $2/3$, $\boldsymbol{r}'$ is an ϵ-stationary point of F.*

We show the following:

Lemma 4.36. *Let D be an isotropic log-concave distribution on $\mathbb{R}^d$. Let $\boldsymbol{w} \in \mathbb{R}^d$ be a unit vector and let $B = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{w} \cdot \boldsymbol{x} \in [a, b]\}$ for $a, b > 0$ smaller than some universal absolute constant. There exists an algorithm that, given $\gamma > 0$ and $\mathrm{poly}(d/\gamma)$ independent samples from $D_B^{\mathrm{proj}_{\boldsymbol{w}^\perp}}$, runs in sample polynomial time, and returns a vector $\boldsymbol{r}$ such that if $\boldsymbol{z}$ is obtained from $D_B^{\mathrm{proj}_{\boldsymbol{w}^\perp}}$ by rejection sampling, where a sample $\boldsymbol{x}$ is accepted with probability $\min(1, e^{-\boldsymbol{r} \cdot \boldsymbol{x}})$, then:*

- *A sample is rejected with probability p, where $p \in (0, 1)$ is an absolute constant.*
- *The distribution of $\boldsymbol{z}$ is $(\gamma, O(1))$-isotropic log-concave.*

Proof. For notational convenience, let $D' = D_B^{\mathrm{proj}_{w^\perp}}$. First, we note that for $\boldsymbol{u}$ any unit vector perpendicular to $\boldsymbol{w}$ and any $\boldsymbol{r}$ perpendicular to $\boldsymbol{w}$, we can apply Fact 4.27 to the projection of $\boldsymbol{x}$ onto the subspace spanned by $\boldsymbol{u}, \boldsymbol{w}$ and $\boldsymbol{r}$.

We denote by c the constant c_3 from Fact 4.27. As a result, we have that the distribution of $\boldsymbol{u} \cdot \boldsymbol{x}'$ will have constant probability density in a neighborhood of 0 and will have exponential tails. Furthermore, this will still hold after rejection sampling with probability $\min(1, e^{-\boldsymbol{r}\cdot\boldsymbol{x}'})$. This implies that no matter what $\boldsymbol{r}$ is chosen, $\boldsymbol{z}$ will be approximately isotropic. Moreover, $\boldsymbol{z}$ will be log-concave automatically, because the rejection sampling multiplies the pdf by a log-concave function. Furthermore, the probability of a sample being accepted will be at least $\mathbf{Pr}_{\boldsymbol{x}'\sim D'}[\boldsymbol{r} \cdot \boldsymbol{x}' \leq 0]$, which is at least c^4.

It remains to prove the second condition of the lemma. We let R be a sufficiently large constant and apply projected SGD to find an approximate stationary point of the non-convex function $F(\boldsymbol{r}) := \|\boldsymbol{g}(\boldsymbol{r})\|_2^2$, where $\boldsymbol{g}(\boldsymbol{r}) := \mathbf{E}_{\boldsymbol{x}'\sim D'}[\boldsymbol{x}' \min(1, \exp(-\boldsymbol{r} \cdot \boldsymbol{x}'))]$, in the feasible set $\{\boldsymbol{r} \in \mathbb{R}^d : \|\boldsymbol{r}\|_2 \leq R\}$. Note that $\boldsymbol{g}(\boldsymbol{r})$ is the mean of the distribution of $\boldsymbol{z}$.

We will need the following claim about the approximate stationary points of $F(\boldsymbol{r})$.

Claim 4.37. *Any interior point of the feasible region, i.e., a point $\boldsymbol{r}$ such that $\|\boldsymbol{r}\|_2 < R$, has $\|\nabla F(\boldsymbol{r})\|_2 = \Omega(\|g(\boldsymbol{r})\|_2)$. Moreover, F has no stationary points on the boundary, i.e., on the set $\{\boldsymbol{r} \in \mathbb{R}^d : \|\boldsymbol{r}\|_2 = R\}$.*

Proof. We show that the Jacobian matrix of $\boldsymbol{g}$ is negative definite. In particular, for any vector $\boldsymbol{u} \neq \mathbf{0}$, we have

$$\begin{aligned}\boldsymbol{u} \cdot \mathrm{Jac}(\boldsymbol{g}(\boldsymbol{r}))\boldsymbol{u} &= \boldsymbol{u} \cdot - \mathop{\mathbf{E}}_{\boldsymbol{x}' \sim D'} \left[\boldsymbol{x}'(\boldsymbol{x}')^{\intercal} \mathbb{1}\{\boldsymbol{r} \cdot \boldsymbol{x}' \geq 0\} \exp(-\boldsymbol{r} \cdot \boldsymbol{x}')\right] \boldsymbol{u} \\ &= - \mathop{\mathbf{E}}_{\boldsymbol{x}' \sim D'} [\mathbb{1}\{\boldsymbol{r} \cdot \boldsymbol{x}' \geq 0\} \exp(-\boldsymbol{r} \cdot \boldsymbol{x}')(\boldsymbol{u} \cdot \boldsymbol{x}')^2] \\ &= - \mathop{\mathbf{E}}_{\boldsymbol{x}' \sim D'} \left[\mathbb{1}\{\boldsymbol{r} \cdot \boldsymbol{x}' \geq 0\} \exp(-\boldsymbol{r} \cdot \boldsymbol{x}')(\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} \cdot \boldsymbol{x}')^2\right] \|\boldsymbol{u}\|_2^2 \\ &\leq -\frac{c^2}{4} e^{-c} \mathop{\mathbf{E}}_{\boldsymbol{x}' \sim D'} \left[\mathbb{1}\{c \geq \boldsymbol{r} \cdot \boldsymbol{x}' \geq 0\} \mathbb{1}\left\{\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} \cdot \boldsymbol{x}' \geq c/2\right\}\right] \|\boldsymbol{u}\|_2^2 \\ &\leq -\frac{c^5}{24} e^{-c} \|\boldsymbol{u}\|_2^2 \,,\end{aligned}$$

where we used Fact 4.27 which gives $\boldsymbol{u} \cdot \mathrm{Jac}(\boldsymbol{g}(\boldsymbol{r}))\boldsymbol{u} = -O(\|\boldsymbol{u}\|_2^2)$. Observe that the gradient of F at $\boldsymbol{r}$ is $\nabla F(\boldsymbol{r}) = 2\mathrm{Jac}(\boldsymbol{g}(\boldsymbol{r}))\boldsymbol{g}(\boldsymbol{r})$, where $\mathrm{Jac}(\boldsymbol{g}(\boldsymbol{r}))$ is the Jacobian of $\boldsymbol{g}$ at point $\boldsymbol{r}$, thus $\|\nabla F(\boldsymbol{r})\|_2 \geq \boldsymbol{u} \cdot \mathrm{Jac}(\boldsymbol{g}(\boldsymbol{r}))\boldsymbol{g}(\boldsymbol{r}) / \|\boldsymbol{u}\|_2$ for any vector $\boldsymbol{u}$. Setting $\boldsymbol{u} = \boldsymbol{g}(\boldsymbol{r})$, we have that $\|\nabla F(\boldsymbol{r})\|_2 = \Omega(\|g(\boldsymbol{r})\|_2)$.

It remains to prove that there is no stationary point on the boundary. That is, for a point $\boldsymbol{r}$ with $\|\boldsymbol{r}\|_2 = R$, the gradient of F at $\boldsymbol{r}$ is a negative multiple of $\boldsymbol{r}$. It is easy to see that, using Fact 4.27 for R at least a sufficiently large constant, $\boldsymbol{g}(\boldsymbol{r}) \cdot \boldsymbol{r} < 0$. So, if the gradient of F at $\boldsymbol{r}$ is a negative multiple of $\boldsymbol{r}$, we have that

$$0 < \boldsymbol{g}(\boldsymbol{r}) \cdot \nabla F(\boldsymbol{r}) = 2\boldsymbol{g}(\boldsymbol{r}) \cdot \mathrm{Jac}(\boldsymbol{g}(\boldsymbol{r}))\boldsymbol{g}(\boldsymbol{r}) < 0 \,,$$

which is a contradiction. □

As a result, an internal stationary point of F must have $\|\boldsymbol{g}(\boldsymbol{r})\|_2$ close to 0, which would imply that the conditional distribution $\boldsymbol{z}$ with that $\boldsymbol{r}$ has mean less than γ. In the following claim, we prove that $F(\boldsymbol{r})$ is smooth with respect the Euclidean norm. See Appendix C.1 for the proof.

Claim 4.38. *The function $F(\boldsymbol{r})$ is L-smooth, for some $L = \mathrm{poly}(d)$.*

Thus, by Fact 4.35, running Stochastic Gradient Descent for $T = \mathrm{poly}(d/\gamma)$, we obtain a γ-stationary point, assuming we have an unbiased estimator for the gradient of F. Note that by taking two independent samples $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ from D' and setting $\hat{\boldsymbol{g}}(\boldsymbol{r}, \boldsymbol{x}) = \boldsymbol{x} \min(1, \exp(-\boldsymbol{r} \cdot \boldsymbol{x}))$, the quantity $2\mathrm{Jac}(\hat{\boldsymbol{g}}(\boldsymbol{r}, \boldsymbol{x}^{(1)}))\hat{\boldsymbol{g}}(\boldsymbol{r}, \boldsymbol{x}^{(2)})$ is an unbiased estimator for $\nabla F(\boldsymbol{r})$. This completes our proof. □

Algorithm 6 Computing a Good Initialization Vector

```
**procedure** WarmStart($(A, \alpha), \epsilon, \boldsymbol{w}, D$)
**Input:** Samples from an $O(\gamma, O(1))$-isotropic log-concave distribution that satisfies the $(\alpha, A)$-Tsybakov noise condition, and a unit vector $\boldsymbol{w}$ such that $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \Theta(\epsilon)$.
**Output:** A vector $\boldsymbol{v}$ such that $\boldsymbol{v} \cdot (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \geq (\alpha\epsilon/A)^{O(1/\alpha)}$.

    $s \leftarrow \Theta(\alpha\epsilon/\log(A\log(A)/(\alpha\epsilon)))$, $\xi \leftarrow (\Theta(A/s))^{1/\alpha}$, $s' = \Theta(\xi^3 s\epsilon)$
    $N \leftarrow \mathrm{poly}(d) \cdot (A/(\alpha\epsilon))^{O(1/\alpha)}$
    Let $x_0$ be a uniform random number on $[s, 2s]$.
    Let $D'$ denote $D$ conditioned on $\boldsymbol{w} \cdot \boldsymbol{x} \in [x_0, x_0 + s']$ and projected onto $\boldsymbol{w}^{\perp}$.
    $\widehat{D} \leftarrow$ MakeIsotropic($D', 1/\log(1/\xi), N$)
    $\bar{\boldsymbol{x}} \leftarrow \mathbf{E}_{\boldsymbol{x}\sim\widehat{D}_{\boldsymbol{x}}}[\boldsymbol{x}]$; $\bar{\boldsymbol{X}} \leftarrow \mathbf{E}_{\boldsymbol{x}\sim\widehat{D}_{\boldsymbol{x}}}[\boldsymbol{x}\boldsymbol{x}^{\intercal}]$
    Normalize all samples in $\widehat{D}$ with $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{X}}$
    $\boldsymbol{T}_1' \leftarrow \mathbf{E}_{(\boldsymbol{x},y)\sim\widehat{D}}[y\boldsymbol{x}]$ and $\boldsymbol{T}_2' \leftarrow \mathbf{E}_{(\boldsymbol{x},y)\sim\widehat{D}}[y(\boldsymbol{x}\boldsymbol{x}^{\intercal} - \boldsymbol{I})]$
    Let $V$ be the subspace spanned by $\boldsymbol{T}_1'$ and the eigenvectors of $\boldsymbol{T}_2'$ whose eigenvalues have absolute value at least $\xi$.
    **return** a random vector in $V$.
```

Algorithm 7 Putting the Distribution in Nearly-Isotropic Position

```
procedure MakeIsotropic($D_B^{\mathrm{proj}_{w^\perp}}$, $\gamma$, $N$)
Input: Samples from the log-concave distribution $D_B^{\mathrm{proj}_{w^\perp}}$, i.e., the log-concave distribution $D$ conditioned on a band $B = \{\boldsymbol{x} : \boldsymbol{w} \cdot \boldsymbol{x} \in [a, b]\}$ and then projected onto $\boldsymbol{w}^\perp$.
Output: $N$ i.i.d. samples from a $(\gamma, O(1))$-isotropic log-concave distribution obtained from $D_B^{\mathrm{proj}_{w^\perp}}$ by rejection sampling.

    Let $F(\boldsymbol{r}) = \left\| \mathbf{E}_{\boldsymbol{x} \sim D_B^{\mathrm{proj}_{w^\perp}}} [\boldsymbol{x} \min(1, \exp(-\boldsymbol{r} \cdot \boldsymbol{x}))] \right\|_2^2$
    Runs SGD on $F$ to obtain a $\gamma$-stationary point $\boldsymbol{r}'$.    ▷ Takes poly$(d/\gamma)$ time.
    $S \leftarrow \emptyset$
    while $|S| \le N$
        Draw sample $(\boldsymbol{x}, y)$ from $D_B^{\mathrm{proj}_{w^\perp}}$.
        $S \leftarrow S \cup \{(\boldsymbol{x}, y)\}$ with probability $\min(1, \exp(-\boldsymbol{r}' \cdot \boldsymbol{x}))$.
    Let $D_S$ be the uniform distribution from $S$.
return the sample $D_S$.
```

We are now ready to prove Theorem 4.24.

Proof of Theorem 4.24. Using the condition $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \Theta(\epsilon)$, we can assume that $\boldsymbol{w}^* = \lambda_1 \boldsymbol{w} + \lambda_2 (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}}$, where $|\lambda_1| = 1 - \Theta(\epsilon)$, and $\lambda_2 = \Theta(\epsilon)$. If $\boldsymbol{w} \cdot \boldsymbol{w}^* = 0$, then we can directly apply Proposition 4.30 to obtain a vector with non-trivial correlation. For the general case, we show how we can construct a distribution that satisfies the conditions of Proposition 4.30.

Let s be a sufficiently small multiple of $\alpha\epsilon / \log(A \log(A)/(\alpha\epsilon))$, $\xi = (\Theta(s/A))^{1/\alpha}$, and let $s' = \xi^3 s\epsilon$. Finally, let x_0 be a uniform random number in $[s, 2s]$. Consider the conditional distribution on the random band $B_{x_0} = \{\boldsymbol{w} \cdot \boldsymbol{x} \in [x_0, x_0 + s']$ and projected onto $\boldsymbol{w}^\perp$, i.e., $D_{B_{x_0}}^{\mathrm{proj}_{w^\perp}} := D^\perp$.

Set $\boldsymbol{x}^\perp = \mathrm{proj}_{\boldsymbol{w}^\perp} \boldsymbol{x}$, $f^\perp(\boldsymbol{x}^\perp) = \mathrm{sign}\left(\boldsymbol{x}^\perp \cdot (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} + \frac{\lambda_1 x_0}{\lambda_2} \right)$, and

$$\eta^\perp(\boldsymbol{x}^\perp) = \mathop{\mathbf{Pr}}_{(\boldsymbol{x}^\perp, y) \sim D^\perp} [y \neq f^\perp(\boldsymbol{z}) | \boldsymbol{z} = \boldsymbol{x}^\perp].$$

Using Lemma 4.33, we get that $D^{\perp}$ is $(O(1), O(1))$-isotropic and with high probability it holds $\mathbf{Pr}_{\boldsymbol{x}^{\perp} \sim D_{\boldsymbol{x}}^{\perp}}[\eta^{\perp}(\boldsymbol{x}^{\perp}) \leq 1/2 - \xi] \geq 2/3$ and $\mathbf{Pr}_{\boldsymbol{x}^{\perp} \sim D_{\boldsymbol{x}}^{\perp}}[\eta^{\perp}(\boldsymbol{x}^{\perp}) \geq 1/2] \leq \xi^3$.

At this point, we have that $D^{\perp}$ is approximately isotropic, but may be relatively far from mean 0 (the mean can be at constant distance from the origin, whereas we need it to be roughly $1/\log(1/\xi)$). To overcome this issue, we apply Lemma 4.36. We define $\bar{D}$ to be the distribution of $\boldsymbol{z}$ that is produced according to Lemma 4.36 with γ a small multiple of $1/\log(1/\xi)$, and consider the distribution on $\boldsymbol{z}$ and y. Notice that y is a noisy version of $f^{\perp}(\boldsymbol{x})$ (with noise rate $\eta^{\perp}(\boldsymbol{x}^{\perp})$), because rejection sampling does not increase the noise rate. Moreover, the mean of $\boldsymbol{z} \cdot (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} + \frac{\lambda_1 x_0}{\lambda_2}$ is at most $\gamma + O(s/\epsilon)$, which is a sufficiently small multiple of $1/\log(1/\xi)$. This means that we can apply Proposition 4.30 to the distribution on $(\boldsymbol{z}, y)$, yielding our final result. □

Proof of Theorem 4.23

Using Theorem 4.24, we can prove Theorem 4.23. The proof is similar to the proof of Theorem 4.5, but we additionally need to guess how far the current guess $\boldsymbol{w}$ is from $\boldsymbol{w}^*$.

Proof of Theorem 4.23. First, we guess a value ϵ such that $\|\boldsymbol{w} - \boldsymbol{w}^*\|_2 = \Theta(\epsilon)$, where $\epsilon = \Omega(\theta)$. From Proposition 4.8, for $\rho = O(\theta(\alpha\epsilon/A)^{O(1/\alpha)})$, the distribution $D_B^{\pi_{\boldsymbol{w}}}$ is $(2, \Omega(\rho), 1/\rho, O(1/\rho \log(1/\rho))$-well-behaved and satisfies the $(\alpha, O(A/\rho))$-Tsybakov noise condition, where we used (from Fact 4.27) that the values L, R are absolute constants. Using Theorem 4.24, a random unit vector $\boldsymbol{v} \in \mathbb{R}^d$ with constant probability δ_1 satisfies $\boldsymbol{v} \cdot (\boldsymbol{w}^*)^{\perp_{\boldsymbol{w}}} \geq (\alpha\epsilon/A)^{O(1/\alpha)}$. We call this event $\mathcal{E}$. Conditioning on the event $\mathcal{E}$, from Proposition 4.13, using $\frac{\beta^4}{\theta^2}\left(\frac{A}{\theta\alpha}\right)^{O(1/\alpha^2)} \log(1/\delta)$ samples, with probability $1 - \delta$, we get a $(\boldsymbol{v}', R, t_0)$ such that

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y) \sim D_B^{\pi_{\boldsymbol{w}}}}\left[\mathbb{1}[-R \leq \boldsymbol{v}' \cdot \boldsymbol{x} \leq -t_0] y\right] \leq -\left(\theta\alpha/A\right)^{O(1/\alpha^2)}/\beta\,.$$

Using Lemma 4.9, we get that

$$\mathop{\mathbf{E}}_{(x,y)\sim D}[T_{w}(x)x\cdot wy]\leq -\left(\theta\alpha/A\right)^{O(1/\alpha^2)}/\beta\ .$$

Conditioning on the event $\mathcal{E}^c$, where $\mathcal{E}^c$ is the complement of $\mathcal{E}$, Algorithm 5 either returns a certificate or returns nothing. Thus, by taking $k = O(\log(1/\delta))$ random vectors, we get that the probability that event $\mathcal{E}^c$ happens is at most $(1-\delta_1)^k \leq e^{-\delta_1 k}$. Thus, by taking $O(\log 1/\delta)$ random vectors and running Algorithm 5 with confidence $\delta/\log(1/\delta)$, we get a certificate with probability $1-2\delta$. Moreover, the number of samples needed to construct the empirical distribution is $\left(\frac{A}{\theta\alpha}\right)^{O(1/\alpha^2)}\log(1/\delta)$. Finally, to guess the value of ϵ, it suffices to run the algorithm for the values $\theta, 2\theta, \ldots, 1$ which will increase the complexity by a $\log(1/\theta)$ factor. This completes the proof of Theorem 4.23. □

4.5 Learning a Near-Optimal Halfspace via Online Convex Optimization

In this section we present a black-box approach that uses our certificate algorithms from the previous sections to learn halfspaces in the presence of Tsybakov noise. In more detail, we provide a generic result showing that one can apply a certificate oracle in a black-box manner combined with online gradient descent to learn the unknown halfspace. We note that an essentially identical approach, with slightly different formalism, was given in Diakonikolas et al. (2021b).

Using the aforementioned approach, we establish the two main algorithmic results of this paper.

Theorem 4.39 (Learning Tsybakov Halfspaces under Well-Behaved Distributions). *Let D be a $(3, L, R, U, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(x) = \mathrm{sign}(w^* \cdot x)$. There exists an algorithm that draws $N = \beta^4 \left(\frac{dUA}{RL\epsilon}\right)^{O(1/\alpha)} \log(1/\delta)$ sam-*

ples from D, runs in $\mathrm{poly}(N, d)$ *time, and computes a vector* $\widehat{\boldsymbol{w}}$ *such that, with probability* $1 - \delta$*, we have that* $\mathrm{err}_{0-1}^{D_x}(h_{\widehat{\boldsymbol{w}}}, f) \leq \epsilon$*.*

For the important special case of log-concave distributions on examples, we give a more efficient learning algorithm.

Theorem 4.40 (Learning Tsybakov Halfspaces under Log-concave Distributions)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$ and is such that $D_{\boldsymbol{x}}$ is isotropic log-concave. There exists an algorithm that draws $N = \mathrm{poly}(d) \cdot \left(\frac{A}{\epsilon}\right)^{O(1/\alpha^2)} \log(1/\delta)$ samples from D, runs in $\mathrm{poly}(N, d)$ time, and computes a vector $\widehat{\boldsymbol{w}}$ such that, with probability $1 - \delta$, we have that $\mathrm{err}_{0-1}^{D_x}(h_{\widehat{\boldsymbol{w}}}, f) \leq \epsilon$.*

To formally describe the approach of this section, we require the notion of a *certificate oracle*. A certificate oracle is an algorithm that, given a candidate weight vector $\boldsymbol{w}$ and an accuracy parameter $\rho > 0$, it returns a certifying function $T(\boldsymbol{x})$. Recall that a certifying function is a non-negative function that satisfies $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x}) y \boldsymbol{x} \cdot \boldsymbol{w}] \leq -\rho$ for some $\rho > 0$. We have already described how to efficiently implement such an oracle in Section 4.3.

Definition 4.41 (Certificate Oracle)**.** *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$. For a decreasing function $\rho(\cdot) : \mathbb{R}_+ \mapsto \mathbb{R}_+$, we define $\mathcal{C}(\boldsymbol{w}, \theta, \delta)$ to be the following ρ-certificate oracle: For any unit vector $\boldsymbol{w}$ and $\theta > 0$, if $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \geq \theta$, then a call to $\mathcal{C}(\boldsymbol{w}, \theta, \delta)$, with probability at least $1 - \delta$, returns a function $T(\boldsymbol{x})$, with $\|T\|_\infty \leq 1$ such that*

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x}) y \boldsymbol{x} \cdot \boldsymbol{w}] \leq -\rho(\theta)\,,$$

and with probability at most δ returns "FAIL".

Remark 4.42. *We note that the above oracle provides a "one-sided" guarantee in the following sense. When the candidate vector $\boldsymbol{w}$ satisfies $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \geq \theta$, the oracle is required to return a certifying function T with high probability. But it may also return* *such a function when $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \leq \theta$. In other words, the oracle is not required to output "FAIL" with high probability when $\boldsymbol{w}$ is nearly parallel to $\boldsymbol{w}^*$. We show that an one-sided oracle of non-optimality suffices for our purposes.*

Remark 4.43. *By Fact 3.3, the optimal halfspace $\boldsymbol{w}^*$ satisfies $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\, y\boldsymbol{x} \cdot \boldsymbol{w}^*] \geq 0$ for any non-negative function T. Therefore, as $\boldsymbol{w}$ approaches $\boldsymbol{w}^*$, we have that*

$$\lim_{\theta(\boldsymbol{w},\boldsymbol{w}^*)\to 0} \inf_{T:\|T\|_\infty \leq 1} \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\, y\boldsymbol{x} \cdot \boldsymbol{w}] = 0,$$

where $\|T\|_\infty$ is the ℓ_∞ norm for functions, i.e., $\|T\|_\infty = \sup_{\boldsymbol{x}\in\mathbb{R}^d} |T(\boldsymbol{x})|$. That is, $\lim_{\theta\to 0} \rho(\theta) = 0$ and it is natural that the non-negative function $\rho(\theta)$ is a decreasing function of the (lower bound on the) angle between $\boldsymbol{w}$ and $\boldsymbol{w}^*$. Intuitively, the closer $\boldsymbol{w}$ is to $\boldsymbol{w}^*$, the harder it is to find a certifying function T that makes $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[T(\boldsymbol{x})\, y\boldsymbol{x} \cdot \boldsymbol{w}]$ sufficiently negative. Moreover, if our goal is to estimate the vector $\boldsymbol{w}^*$ within angle ϵ, we can always give the oracle this worst-case target angle, i.e., $\theta = \epsilon$. Finally, notice that when the distribution D is isotropic, we have $\rho(\theta) \leq 1$, as follows from $\|T\|_\infty \leq 1$ and the Cauchy-Schwarz inequality.

Given a certificate oracle, the following result shows we can efficiently approximate the optimal halfspace using projected online gradient descent.

Proposition 4.44 (Certificate-Based Optimization)**.** *Let D be a $(3, L, R, \beta)$-well-behaved isotropic distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition with respect to an unknown halfspace $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^* \cdot \boldsymbol{x})$, and let $\mathcal{C}$ be a ρ-certificate oracle. There exists an algorithm that makes at most $T = \frac{1}{\rho^2(\epsilon)} \frac{1}{\alpha} \left(\frac{A}{R\,L}\right)^{O(1/\alpha)}$ calls to $\mathcal{C}(\cdot)$, draws $N = d\frac{T\beta^2}{\rho^2(\epsilon)} \log\left(\frac{dT}{\delta\rho(\epsilon)}\right)$ samples from D, runs in time $\mathrm{poly}(T, N, d)$, and computes a weight vector $\widehat{\boldsymbol{w}}$ such that with probability $1 - \delta$ we have that $\theta(\widehat{\boldsymbol{w}}, \boldsymbol{w}^*) \leq \epsilon$.*

The algorithm establishing Proposition 4.44 is given in pseudocode in Algorithm 8. In the remaining part of this section, we provide a proof sketch of Proposition 4.44. The full argument is given in Appendix C.2.

Proof Sketch. The main idea of the algorithm is to provide a sequence of adaptively chosen convex loss functions to an Online Convex Optimization algorithm, for example Online Gradient Descent (OGD). In more detail, we construct these loss functions using our certificate oracle $\mathcal{C}$. At round t, we call the certificate oracle to obtain a certifying function $T(\boldsymbol{x})$ and set

$$\ell_t(\boldsymbol{w}) = - \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D} [(T(\boldsymbol{x}) + \lambda) y \boldsymbol{x}] \cdot \boldsymbol{w} \,,$$

where $\lambda > 0$ acts similarly to a regularizer. The term $\lambda \, \mathbf{E}_{(\boldsymbol{x},y)\sim D}[y\boldsymbol{x}] \cdot \boldsymbol{w}$ prevents the trivial vector $\boldsymbol{w} = \mathbf{0}$ from being a valid solution (in the sense of one that minimizes regret, see also the full proof in Appendix C.2).

The crucial property of the above sequence of loss functions is that they are positive and bounded away from 0 when $\boldsymbol{w}$ is far from $\boldsymbol{w}^*$. Their value will always be greater than (roughly) $\rho(\epsilon)$, given the guarantee of our certificate oracle from Definition 4.41 for $\theta = \epsilon$ and assuming that the regularizer λ is sufficiently small.

We then provide this convex loss function to the OGD algorithm that updates the guess according to the gradient of $\ell_t(\boldsymbol{w})$. Our analysis follows from the regret guarantee of OGD. Since we provide convex (and in particular linear) loss functions to OGD, we know the average regret will converge to 0 as $T \to \infty$ with a convergence rate roughly $O(1/\sqrt{T})$. This means that the oracle can only succeed in returning certifying functions for a bounded number of rounds, since every time the oracle succeeds, OGD suffers loss of at least $\rho(\epsilon)$. Therefore, after roughly $1/\rho(\epsilon)^2$ rounds the regret will be so small that for at least one round the certificate oracle must have failed. Our algorithm then stops and returns the halfspace of that iteration. Even though our certificate is “one-sided", we know that the probability that it failed with $\theta(\boldsymbol{w}, \boldsymbol{w}^*)$ being larger than ϵ is very small, which implies that we have indeed found a vector $\boldsymbol{w}$ very close to $\boldsymbol{w}^*$. □

Algorithm 8 Learning Halfspaces with Tsybakov Noise using a ρ-certificate oracle $\mathcal{C}$

```
procedure ALG($\epsilon, \delta, D, \mathcal{C}$)                    ▷ $\epsilon$: accuracy, $\delta$: confidence
Input: $D$ is a $(3, L, R, \beta)$-well-behaved distribution that satisfies the $(\alpha, A)$-Tsybakov noise condition, and $\mathcal{C}$ is a $\rho$-certificate oracle.
Output: A vector $\widehat{w}$ such that $\mathrm{err}_{0-1}^{D_x}(h_{\widehat{w}}, f) \le \epsilon$ with probability at least $1 - \delta$.
    $w^{(0)} \leftarrow e_1$
    $T \leftarrow \frac{1}{\rho(\epsilon)^2 \alpha} \left(\frac{A}{R\,L}\right)^{O(1/\alpha)}$
    Draw $N = \tilde{O}\left(d \cdot \frac{T\beta^2}{\rho^2(\epsilon)} \log\left(\frac{1}{\delta}\right)\right)$ samples from $D$ to form the empirical distribution $\widehat{D}$
    for $t = 1, \ldots, T$ do
        $\eta_t \leftarrow 1/(\sqrt{t} + \rho(\epsilon))$
        if $w^{(t-1)} = \mathbf{0}$ then
            Set $\hat{\ell}_t(w) \leftarrow w \cdot -\mathbf{E}_{(x,y)\sim\widehat{D}}\left[\frac{\rho(\epsilon)}{2} y x\right]$
            $w^{(t)} \leftarrow \Pi_{\mathcal{B}}\left(w^{(t-1)} - \eta_t \nabla_w \hat{\ell}_t\left(w^{(t-1)}\right)\right)$
        else
            ANS $\leftarrow \mathcal{C}(w^{(t-1)} / \left\|w^{(t-1)}\right\|_2, \epsilon, \delta/T)$
            if ANS = FAIL then
                return $w^{(t-1)}$
        $T_{w^{(t)}}(x) \leftarrow$ ANS
        Set $\hat{\ell}_t(w) \leftarrow w \cdot -\mathbf{E}_{(x,y)\sim\widehat{D}}\left[\left(T_{w^{(t)}}(x) + \frac{\rho(\epsilon)}{2}\right) y x\right]$
        $w^{(t)} \leftarrow \Pi_{\mathcal{B}}\left(w^{(t-1)} - \eta_t \nabla_w \hat{\ell}_t\left(w^{(t-1)}\right)\right)$      ▷ $\mathcal{B} = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$
```

Given Proposition 4.44, it is straightforward to prove our main results. Here we give the proof for the case of log-concave densities and provide a similar argument for well-behaved distributions in Appendix C.2.

Proof of Theorem 4.40. First, we require a ρ-certificate oracle for log-concave distributions. The algorithm of Theorem 4.23 returns a function T_w such that

$$\mathop{\mathbf{E}}_{(x,y)\sim D}\left[T_w(x) y w \cdot x\right] \le -\left(\theta/A\right)^{O(1/\alpha^2)} .$$

From the definition of $T_{\boldsymbol{w}}$ (i.e., Equation (4.3)), it is clear that $\|T_{\boldsymbol{w}}\|_\infty \leq \frac{1}{\min_{\boldsymbol{x}\in B}|\boldsymbol{w}\cdot\boldsymbol{x}|} \leq \left(\frac{\log A}{\alpha\theta}\right)^{O(1/\alpha)}$, where B is the band from Equation (4.3). Note that the function $T_{\boldsymbol{w}}/\,\|T_{\boldsymbol{w}}\|_\infty$ satisfies the conditions of the ρ-certificate oracle. Thus, by scaling the output of the algorithm of Theorem 4.23, we obtain a $(\theta\alpha/A)^{O(1/\alpha^2)}$-certificate oracle. From Proposition 4.44, this gives us an algorithm that returns a vector $\widehat{\boldsymbol{w}}$ such that $\theta(\widehat{\boldsymbol{w}}, \boldsymbol{w}^*) \leq \frac{\epsilon}{\log^2(1/\epsilon)}$ with probability $1-\delta$. Using the fact that for log-concave distributions $\mathrm{err}_{0-1}^{D_{\boldsymbol{x}}}(h_{\widehat{\boldsymbol{w}}}, f) \leq O\left(\log^2(1/\epsilon)\theta(\widehat{\boldsymbol{w}}, \boldsymbol{w}^*)\right) + \epsilon$ (Claim C.12) the result follows. □

5 NOISY LABEL RANKING

5.1 Formal Statement of the Results

Our main contributions are the first efficient algorithms for learning LSFs with bounded noise with respect to Kendall's Tau distance and top-r disagreement loss.

Learning in Kendall's Tau Distance. The most standard metric in rankings Shalev-Shwartz and Ben-David (2014a) is Kendall's Tau (KT) distance which, for two rankings $\pi, \tau \in \mathbb{S}_k$, measures the fraction of pairs (i, j) on which they disagree. That is, $\Delta_{\mathrm{KT}}(\pi, \tau) = \sum_{i \prec_\pi j} \mathbb{1}\{i \succ_\tau j\} / \binom{k}{2}$. Our first result is an efficient learning algorithm that, given samples from an η-noisy linear label ranking distribution $\mathcal{D}$, computes a parameter matrix $\boldsymbol{W}$ that ranks the alternatives almost optimally with respect to the KT distance from the ground-truth ranking $\sigma_{\boldsymbol{W}^\star}(\cdot)$.

Theorem 5.1 (Learning LSFs in KT Distance). *Fix $\eta \in [0, 1/2)$ and $\epsilon, \delta \in (0, 1)$. Let $\mathcal{D}$ be an η-noisy linear label ranking distribution satisfying the assumptions of Definition 1.10 with ground-truth LSF $\sigma_{\boldsymbol{W}^\star}(\cdot)$. There exists an algorithm that draws $N = \widetilde{O}\left(\frac{d}{\epsilon(1-2\eta)^6} \log(k/\delta)\right)$ samples from $\mathcal{D}$, runs in sample-polynomial time, and computes a matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ such that, with probability at least $1 - \delta$,*

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d} [\Delta_{\mathrm{KT}}(\sigma_{\boldsymbol{W}}(\boldsymbol{x}), \sigma_{\boldsymbol{W}^\star}(\boldsymbol{x}))] \leq \epsilon \,.$$

Theorem 5.1 gives the first efficient algorithm with provable guarantees for the supervised problem of learning noisy linear rankings. We remark that the sample complexity of our learning algorithm is qualitatively optimal (up to logarithmic factors) since, for $k = 2$, our problem subsumes learning a linear classifier with Massart noise [1] for which $\Omega(d/\epsilon)$ are known to be information theoretically necessary Massart and Nédélec (2006). Moreover, our learning algorithm is *proper* in the sense that it computes a linear sorting function $\sigma_{\boldsymbol{W}}(\cdot)$. As opposed to improper

[1] Notice that in this case Kendall's Tau distance is simply the standard 0-1 binary loss.

learners (see also Section 5.2), a proper learning algorithm gives us a compact representation (storing $\boldsymbol{W}$ requires $O(kd)$ memory) of the sorting function that allows us to efficiently compute (with runtime $O(kd + k \log k)$) the ranking corresponding to a fresh datapoint $\boldsymbol{x} \in \mathbb{R}^d$.

Learning in top-*r* Disagreement. We next present our learning algorithm for the top-r metric formally defined as $\Delta_{\text{top}-r}(\pi, \tau) = \mathbb{1}\{\pi_{1..r} \neq \tau_{1..r}\}$, where by $\pi_{1..r}$ we denote the ordering on the first r elements of the permutation π. The top-r metric is a disagreement metric in the sense that it takes binary values and for $r = 1$ captures the standard (multiclass) top-1 classification loss. We remark that, in contrast with the top-r classification loss, which only requires the predicted label to be in the top-r predictions of the model, the top-r ranking metric that we consider here requires that the model puts *the same elements in the same order* as the ground truth in the top-r positions. The top-r ranking is well-motivated as, for example, in ad targeting (discussed in Section 5.1) we want to be accurate on the top-r ad categories for a user so that we can diversify the content that they receive.

Theorem 5.2 (Learning LSFs in top-r Disagreement)**.** *Fix $\eta \in [0, 1/2)$, $r \in [k]$ and $\epsilon, \delta \in (0, 1)$. Let $\mathcal{D}$ be an η-noisy linear label ranking distribution satisfying the assumptions of Definition 1.10 with ground-truth LSF $\sigma_{\boldsymbol{W}^\star}(\cdot)$. There exists an algorithm that draws $N = \widetilde{O}\left(\frac{drk}{\epsilon(1-2\eta)^6} \log(1/\delta)\right)$ samples from $\mathcal{D}$, runs in sample-polynomial time and computes a matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ such that, with probability at least $1 - \delta$,*

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\Delta_{\text{top}-r}(\sigma_{\boldsymbol{W}}(\boldsymbol{x}), \sigma_{\boldsymbol{W}^\star}(\boldsymbol{x}))] \leq \epsilon \,.$$

As a direct corollary of our result, we obtain a proper algorithm for learning the top-1 element with respect to the standard 0-1 loss that uses $\widetilde{O}(kd)$ samples. In fact, for small values of r, i.e., $r = O(1)$, our sample complexity is essentially tight. It is known that $\Theta(kd)$ samples are information theoretically necessary Natarajan (1989) for top-1 classification. [2] For the case $r = k$, i.e., when we want to learn

[2] Strictly speaking, those lower bounds do not directly apply in our setting because our labels

the whole ranking with respect to the 0-1 loss, our sample complexity is $O(k^2 d)$. However, using arguments similar to Daniely et al. (2011), one can show that in fact $O(dk)$ ranking samples are sufficient in order to learn the whole ranking with respect to the 0-1 loss. In this case, it is unclear whether a better sample complexity can be achieved with an efficient algorithm and we leave this as an interesting open question for future work.

5.2 Overview of Techniques

Learning in Kendall's Tau distance. Our proper learning algorithm consists of two steps: an improper learning algorithm that decomposes the ranking problem to $O(k^2)$ binary linear classification problems and a convex (second order conic) program that "compresses" the k^2 linear classifiers to obtain a $k \times d$ matrix $\boldsymbol{W}$. Our improper learning algorithm splits the ranking learning problem into $O(k^2)$ binary, d-dimensional linear classification problems with Massart noise. In particular, for every pair of elements $i, j \in [k]$, each binary classification task asks whether element i is ranked higher than element j in the ground-truth permutation $\sigma_{\boldsymbol{W}^\star}(\boldsymbol{x})$. As we already discussed, we have that, under the Gaussian distribution, there exist efficient Massart learning algorithms Balcan and Zhang (2017b); Mangoubi and Vishnoi (2019b); Diakonikolas et al. (2020e); Zhang et al. (2020a); Zhang and Li (2021) that can recover linear classifiers $\mathrm{sign}(\boldsymbol{v}_{ij} \cdot \boldsymbol{x})$ that correctly order the pair i, j for all $\boldsymbol{x}$ apart from a region of $O(\epsilon)$-Gaussian mass. However, we still need to aggregate the results of the *approximate* binary classifiers in order to obtain a ranking of the k alternatives for each $\boldsymbol{x}$. We first show that we can design a "voting scheme" that combines the results of the binary classifiers using an efficient constant factor approximation algorithm for the Minimum Feedback Arc Set (MFAS) problem Ailon et al. (2008). This gives us an efficient but improper algorithm for learning LSFs in Kendall's Tau distance. In order to obtain a proper learning algorithm, we

are whole rankings instead of just the top classes but, in the Appendix D.4, we show that we can adapt the lower bound technique of Daniely et al. (2011) to obtain the same sample complexity lower bound for our ranking setting.

further "compress" the $O(k^2)$ approximate linear classifiers with normal vectors $\boldsymbol{v}_{ij}$ and obtain a matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ with the property that the difference of every two rows $\boldsymbol{W}_i - \boldsymbol{W}_j$ is $O(\epsilon)$-close to the vector $\boldsymbol{v}_{ij}$. More precisely, we show that, given the linear classifiers $\boldsymbol{v}_{ij} \in \mathbb{R}^d$, we can efficiently compute a matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ such that the following angle distance with $\boldsymbol{W}^\star$ is small:

$$d_{\text{angle}}(\boldsymbol{W}, \boldsymbol{W}^\star) \triangleq \max_{i,j} \theta(\boldsymbol{W}_i - \boldsymbol{W}_j, \boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \le O(\epsilon) \,. \tag{5.1}$$

It is not hard to show that, as long as the above angle metric is at most $O(\epsilon)$, then (in expectation over the standard Gaussian) Kendall's Tau distance between the LSFs is also $O(\epsilon)$. A key technical difficulty that we face in this reduction is bounding the "condition number" of the convex (second order conic) program that finds the matrix $\boldsymbol{W}$ given the vectors $\boldsymbol{v}_{ij}$, see Claim 5.6. Finally, we remark that the proper learning algorithm of Theorem 5.1 results in a compact and efficient sorting function that requires: (i) storing $O(k)$ weight vectors as opposed to the initial $O(k^2)$ vectors of the improper learner; and (ii) evaluating k inner products with $\boldsymbol{x}$ to find its ranking (instead of $O(k^2)$).

Learning in top-*r* Disagreement. We next turn our attention to the more challenging top-r ranking disagreement metric. In particular, suppose that we are interested in recovering only the top element of the ranking. One approach would be to directly use the improper learning algorithm for this task and ask for KT distance of order roughly ϵ/k^2. The resulting hypothesis would produce good predictions for the top element but the required sample complexity would be $O(dk^2)$. While it seems that training $O(k^2)$ d-dimensional binary classifiers inherently requires $O(dk^2)$ samples, we show that, using the proper KT distance learning algorithm of Theorem 5.1, we can also obtain improved sample complexity results for the top-r metric. Our main technical contribution here is a novel estimate of the top-r disagreement in terms of the angle metric. In general, one can show that the top-r disagreement is at most $O(k^2)$ $d_{\text{angle}}(\boldsymbol{W}, \boldsymbol{W}^\star)$. We significantly sharpen this estimate by showing the following lemma.

Lemma 5.3 (Top-r Disagreement via Parameter Distance)**.** *Consider two matrices $\boldsymbol{W}, \boldsymbol{W}^\star \in \mathbb{R}^{k \times d}$ and let $\mathcal{N}_d$ be the standard Gaussian in d dimensions. We have that*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\sigma_{1..r}(\boldsymbol{W}\boldsymbol{x}) \neq \sigma_{1..r}(\boldsymbol{W}^\star \boldsymbol{x})] \leq \widetilde{O}(kr)\, d_{\mathrm{angle}}(\boldsymbol{W}, \boldsymbol{W}^\star)\,.$$

We remark that Lemma 5.3 is a general geometric tool that we believe will be useful in other distribution-specific multiclass learning settings. The proof of Lemma 5.3 mainly relies on geometric Gaussian surface area computations that we believe are of independent interest. For the details, we refer the reader to Section 5.5. An interesting question with a convex-geometric flavor is whether the sharp bound of Lemma 5.3 also holds under the more general class of isotropic log-concave distributions.

5.3 Notation and Preliminaries

General Notation. We use $\widetilde{O}(\cdot)$ to omit poly-logarithmic factors. A learning algorithm has sample-polynomial runtime if it runs in time polynomial in the size of the description of the input training set. We denote vectors by boldface $\boldsymbol{x}$ (with elements x_i) and matrices with $\boldsymbol{W}$, where we let $\boldsymbol{W}_i \in \mathbb{R}^d$ denote the i-th row of $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ and W_{ij} its elements. We denote $\boldsymbol{a} \cdot \boldsymbol{b}$ the inner product of two vectors and $\theta(\boldsymbol{a}, \boldsymbol{b})$ their angle. Let $\mathcal{N}_d$ denote the d-dimensional standard normal and $\Gamma(\cdot)$ the Gaussian surface area.

Rankings. We let $\mathrm{argsort}_{i \in [k]} \boldsymbol{v}$ denote the ranking of $[k]$ in decreasing order according to the values of $\boldsymbol{v}$. For a ranking π, we let $\pi(i)$ denote the position of the i-th element. If $\pi = \pi(\boldsymbol{x})$, we may also write $\pi(\boldsymbol{x})(i)$ to denote the position of i. We often refer to the elements of a ranking as *alternatives*. For a ranking σ, we let $\sigma_{1..r}$ denote the top-r part of σ. When $\sigma = \sigma(\boldsymbol{x})$, we may also write $\sigma_{1..r}(\boldsymbol{x})$ and $\sigma_\ell(\boldsymbol{x})$ will be the alternative at the ℓ-th position. We let Δ_{KT} denote the (normalized) KT distance, i.e., $\Delta_{\mathrm{KT}}(\pi, \tau) = \sum_{i \prec_\pi j} \mathbb{1}\{i \succ_\tau j\} / \binom{k}{2}$ for $\pi, \tau \in \mathbb{S}_k$.

5.4 Learning in KT distance: Theorem 5.1

In this section, we present the main tools required to obtain our proper learning algorithm of Theorem 5.1. Our proper algorithm adopts a two-step approach: it first invokes an efficient *improper* algorithm which, instead of a linear sorting function (i.e., a matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$), outputs a list of $O(k^2)$ linear classifiers. We then design a novel convex program in order to find the matrix $\boldsymbol{W}$ satisfying the guarantees of Theorem 5.1. Let us begin with the improper learner for LSFs with bounded noise with respect to the KT distance, whose description can be found in Algorithm 9.

Improper Learning Algorithm

Algorithm 9 Non-proper Learning Algorithm `ImproperLSF`

```
Input: Training set $T = \{(\boldsymbol{x}^t, \pi^t)\}_{t \in [N]}, \epsilon, \delta \in (0,1), \eta \in [0, 1/2)$
Output: Sorting function $h : \mathbb{R}^d \to \mathbb{S}_k$

For any $1 \leq i < j \leq k$, create $T_{ij} = \{(\boldsymbol{x}^t, \mathrm{sign}(\pi^t(i) - \pi^t(j)))\}$
For any $1 \leq i < j \leq k$, compute $\boldsymbol{v}_{ij} = \texttt{MassartLTF}(T_{ij}, \frac{\epsilon}{4}, \frac{\delta}{10k^2}, \eta)$ ▷ See Appendix D.1
Ranking Phase: Given $\boldsymbol{x} \in \mathbb{R}^d$:
    (a) Construct directed graph $G$ with $V(G) = [k]$ and edges $e_{i \to j}$ only if $\boldsymbol{v}_{ij} \cdot \boldsymbol{x} > 0 \ \forall i \neq j$
    (b) Output $h(\boldsymbol{x}) = \texttt{MFAS}(G)$                ▷ See Appendix D.1
```

Let us assume that the target function is $\sigma^\star(\boldsymbol{x}) = \sigma_{\boldsymbol{W}^\star}(\boldsymbol{x}) = \mathrm{argsort}(\boldsymbol{W}^\star \boldsymbol{x})$ for some $\boldsymbol{W}^\star \in \mathbb{R}^{k \times d}$.

Step 1: Binary decomposition and Noise Structure. For each drawn example $(\boldsymbol{x}, \pi)$ from the η-noisy linear label ranking distribution $\mathcal{D}$ (see Definition 1.10), we create $\binom{k}{2}$ binary examples $(\boldsymbol{x}, y_{ij})$ with $y_{ij} = \mathrm{sign}(\pi(i) - \pi(j))$ for any $1 \leq i < j \leq$

k. We have that

$$\begin{aligned}&\mathop{\mathbf{Pr}}_{(\boldsymbol{x},\pi)\sim\mathcal{D}}\left[y_{ij}\cdot\mathrm{sign}((\boldsymbol{W}_i^\star-\boldsymbol{W}_j^\star)\cdot\boldsymbol{x})<0\mid\boldsymbol{x}\right]\\&=\mathop{\mathbf{Pr}}_{\pi\sim\mathcal{M}(\sigma^\star(\boldsymbol{x}))}\left[\pi(i)<\pi(j)\mid\boldsymbol{W}_i^\star\cdot\boldsymbol{x}<\boldsymbol{W}_j^\star\cdot\boldsymbol{x}\right].\end{aligned}$$

Since $\mathcal{M}(\sigma^\star(\boldsymbol{x}))$ is an η-noisy ranking distribution (see Definition 1.9), we get that the above quantity is at most $\eta < 1/2$. Therefore, each sample $(\boldsymbol{x}, y_{ij})$ can be viewed as a sample from a distribution $\mathcal{D}_{ij}$ with Gaussian $\boldsymbol{x}$-marginal, optimal linear classifier $\mathrm{sign}((\boldsymbol{W}_i^\star-\boldsymbol{W}_j^\star)\cdot\boldsymbol{x})$, and Massart noise η. Hence, we have reduced the task of learning noisy LSFs to a number of $\binom{k}{2}$ sub-problems concerning the learnability of halfspaces in the presence of bounded (Massart) noise.

Step 2: Solving Binary Sub-problems. We can now apply the algorithm `MassartLTF` for LTFs with Massart noise under standard Gaussian marginals Zhang et al. (2020a) (for details, see Appendix D.1): for all the pairs of alternatives $1 \leq i < j \leq k$ with accuracy parameter ϵ', confidence $\delta' = O(\delta/k^2)$, and a total number of $N = \widetilde{\Omega}\left(\frac{d}{\epsilon'(1-2\eta)^6}\log(k/\delta)\right)$ i.i.d. samples from $\mathcal{D}$, we can obtain a collection of linear classifiers with normal vectors $\boldsymbol{v}_{ij}$ for any $i < j$. We remark that each one of these halfspaces $\boldsymbol{v}_{ij}$ achieves ϵ disagreement with the ground-truth halfspaces $\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star$ with high probability, i.e.,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[\mathrm{sign}(\boldsymbol{v}_{ij}\cdot\boldsymbol{x})\neq\mathrm{sign}((\boldsymbol{W}_i^\star-\boldsymbol{W}_j^\star)\cdot\boldsymbol{x})]\leq\epsilon'.$$

Step 3: Ranking Phase. We now have to aggregate the linear classifiers and compute a single sorting function $h : \mathbb{R}^d \to \mathbb{S}_k$. Given an example $\boldsymbol{x}$, we create the tournament graph G with k nodes that contains a directed edge $e_{i\to j}$ if $\boldsymbol{v}_{ij}\cdot\boldsymbol{x} > 0$. If G is acyclic, we output the induced permutation; otherwise, the graph contains cycles which should be eliminated. In order to output a ranking, we remove cycles from G with an efficient, 3-approximation algorithm for MFAS Ailon et al. (2008); Van Zuylen and Williamson (2009). Hence, the output $h(\boldsymbol{x})$ and the true target $\sigma^\star(\boldsymbol{x})$ will have $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}_d}[\Delta_{\mathrm{KT}}(h(\boldsymbol{x}),\sigma^\star(\boldsymbol{x}))] \leq \epsilon' + 3\epsilon' = 4\epsilon'$. This last equation indicates why a constant factor approximation algorithm suffices for our purposes – we can always pick $\epsilon' = \epsilon/4$ and complete the proof. For details, see Appendix D.1.

Proper Learning Algorithm: Theorem 5.1

Having obtained the improper learning algorithm, we can now describe our proper Algorithm 10. Initially, the algorithm starts similarly with the improper learner and obtains a collection of binary linear classifiers. The crucial idea is the next step: the design of an appropriate convex program which will efficiently give the matrix $\boldsymbol{W}$. We proceed with the details. For the proof, see Appendix D.1.

Algorithm 10 Proper Learning Algorithm `ProperLSF`

```
Input: Training set $T = \{(\boldsymbol{x}^t, \pi^t)\}_{t\in[N]}, \epsilon, \delta \in (0,1), \eta \in [0,1/2)$
Output: Linear Sorting function $h : \mathbb{R}^d \to \mathbb{S}_k$, i.e., $h(\cdot) = \sigma_{\boldsymbol{W}}(\cdot)$ for some matrix $\boldsymbol{W} \in \mathbb{R}^{k\times d}$

Compute $(\boldsymbol{v}_{ij})_{1\leq i<j\leq k}$ = ImproperLSF$(T, \epsilon, \delta, \eta)$          ▷ See Algorithm 9
Setup the CP 5.2 and compute $\boldsymbol{W}$ = Ellipsoid(CP)          ▷ See Appendix D.1
Ranking Phase:    Given $\boldsymbol{x} \in \mathbb{R}^d$, output $h(\boldsymbol{x}) = \mathrm{argsort}(\boldsymbol{W}\boldsymbol{x})$
```

Step 1: Calling Non-proper Learners. As a first step, the algorithm calls Algorithm 9 with parameters ϵ, δ and $\eta \in [0,1/2)$ and obtains a list of linear classifiers with normal vectors $\boldsymbol{v}_{ij}$ for $i < j$. Without loss of generality, assume that $\|\boldsymbol{v}_{ij}\|_2 = 1$.

Step 2: Designing and Solving the CP 5.2. Our main goal is to find a matrix $\boldsymbol{W}$ whose LSF is close to the true target in KT distance. We show the following lemma that connects the KT distance between two LSFs with the angle metric $d_{\mathrm{angle}}(\cdot,\cdot)$ defined in Eq. (5.1). The proof can be found in the Appendix D.1.

Lemma 5.4. *For $\boldsymbol{W}, \boldsymbol{W}^\star \in \mathbb{R}^{k\times d}$, it holds* $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}_d}[\Delta_{\mathrm{KT}}(\sigma_{\boldsymbol{W}}(\boldsymbol{x}), \sigma_{\boldsymbol{W}^\star}(\boldsymbol{x}))] \leq d_{\mathrm{angle}}(\boldsymbol{W}, \boldsymbol{W}^\star)$.

The above lemma states that, for our purposes, it suffices to control the d_{angle} metric between the guess $\boldsymbol{W}$ and the true matrix $\boldsymbol{W}^\star$. It turns out that, given the binary classifiers $\boldsymbol{v}_{ij}$, we can design a convex program whose solution will satisfy this property. Thinking of the binary classifier $\boldsymbol{v}_{ij}$ as a proxy for $\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star$, we want each difference $\boldsymbol{W}_i - \boldsymbol{W}_j$ to have small angle with $\boldsymbol{v}_{ij}$ or equivalently to have large correlation with it, i.e., $(\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{v}_{ij} \approx \|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2$. To enforce this condition, we can therefore use the second order conic constraint $(\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{v}_{ij} \geq (1-\phi)\|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2$. We formulate the following convex program 5.2 with variable the matrix $\boldsymbol{W}$:

$$\begin{aligned} &\text{Find} && \boldsymbol{W} \in \mathbb{R}^{k \times d}, \;\; \|\boldsymbol{W}\|_F \leq 1, \\ &\text{such that} && (\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{v}_{ij} \geq (1-\phi) \cdot \|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2 \;\; \text{for any } 1 \leq i < j \leq k, \end{aligned} \tag{5.2}$$

for some $\phi \in (0,1)$ to be decided. Intuitively, since any $\boldsymbol{v}_{ij}$ has good correlation with $\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star$ (by the guarantees of the improper learning algorithm) and the CP 5.2 requires that its solution $\boldsymbol{W}$ similarly correlates well with $\boldsymbol{v}_{ij}$, we expect that $d_{\text{angle}}(\boldsymbol{W}, \boldsymbol{W}^\star)$ will be small. We show that:

Claim 5.5. *The convex program 5.2 is feasible and any solution $\boldsymbol{W}$ of 5.2 satisfies $d_{\text{angle}}(\boldsymbol{W}, \boldsymbol{W}^\star) \leq \epsilon$.*

To see this, note that any solution of CP 5.2 is a matrix $\boldsymbol{W}$ whose angle metric (see Eq. (5.1)) with the true matrix is small by an application of the triangle inequality between the angles of $(\boldsymbol{v}_{ij}, \boldsymbol{W}_i - \boldsymbol{W}_j)$ and $(\boldsymbol{v}_{ij}, \boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star)$ for any $i \neq j$. We next have to deal with the feasibility of CP 5.2. Our goal is to determine the value of ϕ that makes the CP 5.2 feasible. For the pair $1 \leq i < j \leq k$, the guess $\boldsymbol{v}_{ij}$ and the true normal vector $\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star$ satisfy, with high probability,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{D}_x}[\operatorname{sign}(\boldsymbol{v}_{ij} \cdot \boldsymbol{x}) \neq \operatorname{sign}((\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \cdot \boldsymbol{x})] \leq \epsilon\,. \tag{5.3}$$

Under the Gaussian distribution (which is rotationally symmetric), it is well known that the angle $\theta(\boldsymbol{u},\boldsymbol{v})$ between two vectors $\boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^d$ is equal to $\pi \cdot \mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}_d}[\mathrm{sign}(\boldsymbol{u} \cdot \boldsymbol{x}) \neq \mathrm{sign}(\boldsymbol{v} \cdot \boldsymbol{x})]$. Hence, using Eq. (5.3), we get that the angle between the guess $\boldsymbol{v}_{ij}$ and the true normal vector $\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star$ is $\theta(\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star, \boldsymbol{v}_{ij}) \leq c\epsilon$. For sufficiently small ϵ, this bound implies that the cosine of the above angle is of order $1 - (c\epsilon)^2$ and so the following inequality will hold (since $\boldsymbol{v}_{ij}$ is unit):

$$(\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \cdot \boldsymbol{v}_{ij} \geq (1 - 2(c\epsilon)^2) \cdot \|\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star\|_2 \,.$$

Hence, by setting $\phi = 2(c\epsilon)^2$, the convex program 5.2 with variables $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ will be feasible; since $\|\boldsymbol{W}^\star\|_F \leq 1$ comes without loss of generality, $\boldsymbol{W}^\star$ will be a solution with probability $1 - \delta$.

Next, we have to control the volume of the feasible region. This is crucial in order to apply the ellipsoid algorithm (for details, see in Appendix D.1) and, hence, solve the convex program. We show the following claim (see Appendix D.1 for the proof):

Claim 5.6. *There exists $\rho \geq 2^{-\mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))}$ so that the feasible set of CP 5.2 with $\phi = O(\epsilon^2)$ contains a ball (with respect to the Frobenius norm) of radius ρ.*

Critically, the runtime of the ellipsoid algorithm is *logarithmic* in $1/\rho$. So, the ellipsoid runs in time polynomial in the parameters of the problem and outputs the desired matrix $\boldsymbol{W}$.

5.5 Learning in top-*r* Disagreement: Theorem 5.2

In this section we show that the proper learning algorithm of Section 5.4 learns noisy LSFs in the top-r disagreement metric. We have seen that, with $\widetilde{O}(d\log(k)/\epsilon)$ samples, Algorithm 10 of Section 5.4 computes a matrix $\boldsymbol{W}$ such that $d_{\mathrm{angle}}(\boldsymbol{W}, \boldsymbol{W}^\star) \leq \epsilon$, see Claim 5.5. Let us be more specific. Lemma 5.4 relates the expected KT distance with the angle metric of the two matrices (see also Equation (5.1)). Our Algorithm 10 essentially gives an upper bound on this angle metric. When we shift our objective and our goal is to control the top-r disagreement, we can still apply

Algorithm 10 which essentially controls the angle metric. The crucial ingredient that is missing is the relation between the loss we have to control, i.e., the expected top-r disagreement and the angle metric of Equation 5.1. This relation is presented right after and essentially says that the expected top-r disagreement is at most $O(kr)$ times this angle metric. Hence, in order to get top-r disagreement of order ϵ, it suffices to apply our Algorithm 10 with $\epsilon' = O(\epsilon/(kr))$.

We continue with our main contribution which is the following lemma that connects the top-r disagreement metric with the geometric distance $d_{\text{angle}}(\cdot,\cdot)$, recall Lemma 5.3. To keep this sketch simple we shall present a sketch of the proof of Lemma 5.3 for the special case of top-1 classification, which we restate below. The proof of the top-1 case can be found at the Appendix D.2. The detailed proof of the general case $(r > 1)$ can be found in the Appendix D.3.

Lemma 5.7 (Top-1 Disagreement Loss via $d_{\text{angle}}(\cdot,\cdot)$)**.** *Consider two matrices $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{k\times d}$ and let $\mathcal{N}_d$ be the standard Gaussian in d dimensions. We have that*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[\sigma_1(\boldsymbol{Ux}) \neq \sigma_1(\boldsymbol{Vx})] \leq O\left(k\sqrt{\log k}\right)\, d_{\text{angle}}(\boldsymbol{U},\boldsymbol{V})\,.$$

We observe that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[\sigma_1(\boldsymbol{Ux}) \neq \sigma_1(\boldsymbol{Vx})] = \sum_{i\in[k]} \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[\sigma_1(\boldsymbol{Ux}) = i, \sigma_1(\boldsymbol{Vx}) \neq i]\,. \tag{5.4}$$

We denote by $\mathcal{C}_{\boldsymbol{U}}^{(i)} \triangleq \mathbb{1}\{\boldsymbol{x} : \sigma_1(\boldsymbol{Ux}) = i\} = \prod_{j\neq i} \mathbf{1}\{(\boldsymbol{U}_i - \boldsymbol{U}_j)\cdot \boldsymbol{x} \geq 0\}$, i.e., this is the set where the ranking corresponding to $\boldsymbol{U}$ picks i as the top element. Note that $\mathcal{C}_{\boldsymbol{U}}^{(i)}$ is the indicator of a homogeneous polyhedral cone since it can be written as the intersection of homogeneous halfspaces. Using these cones we can rewrite the top-1 disagreement of Eq. (5.4) as

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[\sigma_1(\boldsymbol{Ux}) \neq \sigma_1(\boldsymbol{Vx})] = \sum_{i\in[k]} \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[C_{\boldsymbol{U}}^{(i)}(\boldsymbol{x}) = 1, C_{\boldsymbol{V}}^{(i)}(\boldsymbol{x}) = 0]\,. \tag{5.5}$$

Hence, our task is to control the mass of the disagreement region of two cones. The next Lemma 5.8 achieves this task and, combined with Eq. (5.5) directly gives the conclusion of Lemma 5.7.

Next we work with two general homogeneous polyhedral cones with set indicators C_1, C_2:

Lemma 5.8 (Cone Disagreement). *Let $C_1, C_2 : \mathbb{R}^d \mapsto \{0,1\}$ be homogeneous polyhedral cones defined by the k unit vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ respectively. For some universal constant $c > 0$, it holds that $\mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{N}_d}[C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})] \leq c\sqrt{\log k}\ \max_{i \in [k]} \theta(\boldsymbol{v}_i, \boldsymbol{u}_i)$.*

Roadmap of the Proof of Lemma 5.8: Assume that we rotate one face of the polyhedral cone C_1 by a very small angle θ to obtain the perturbed cone C_2. At a high-level, we expect the probability of the disagreement region between the new cone C_2 and C_1 to be roughly (this is an underestimation) equal to the size of the perturbation θ times the (Gaussian) surface area of the face of the convex cone that we perturbed. The Gaussian Surface Area (GSA) of a convex set $A \subset \mathbb{R}^d$, is defined as $\Gamma(A) \triangleq \int_{\partial A} \phi_d(\boldsymbol{x}) d\mu(\boldsymbol{x})$, where $d\mu(\boldsymbol{x})$ is the standard surface measure in $\mathbb{R}^d$ and $\phi_d(\boldsymbol{x}) = (2\pi)^{-d/2} \cdot \exp(-\|\boldsymbol{x}\|_2^2/2)$. In fact, in Claim 5.10 below, we show that the probability of the disagreement between C_1 and C_2 is roughly $O(\theta)\Gamma(F_1)\sqrt{\log(1/\Gamma(F_1)+1)}$, where F_1 is the face of cone C_1 that we rotated. Now, when we perturb all the faces by small angles (all perturbations are at most θ), we can show (via a sequence of triangle inequalities) that the total probability of the disagreement region is bounded above by the perturbation size θ times the sum of the Gaussian surface area of every face (times a logarithmic blow-up factor):

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})] \leq O(\theta) \sum_{i=1}^{k} \Gamma(F_i) \sqrt{\log(1/\Gamma(F_i)+1)}\,.$$

Surprisingly, for homogeneous convex cones, the above sum cannot grow very fast with k. In fact, we show that it can be at most $O(\sqrt{\log k})$. To prove this, we crucially rely on the following convex geometry result showing that the Gaussian surface area of a homogeneous convex cone is $O(1)$ regardless of the number of its faces k.

Lemma 5.9 (Nazarov (2003b))**.** *Let C be a homogeneous polyhedral cone with k faces $F_1, \dots, F_k$. Then C has Gaussian surface area $\Gamma(C) = \sum_{i=1}^{k} \Gamma(F_i) \leq 1$.*

Using an inequality similar to the fact that the maximum entropy of a discrete distribution on k elements is at most $\log k$, and, since, from Lemma 5.9, it holds that $\sum_{i=1}^{k} \Gamma(F_i) \leq 1$, we can show that $\sum_{i=1}^{k} \Gamma(F_i)\sqrt{\log(1/\Gamma(F_i)+1)} = O(\sqrt{\log k})$. Therefore, with the above lemma we conclude that, if the maximum angle perturbation that we perform on C_1 is θ, then the probability of the disagreement region is $O(\theta)$. We next give the formal proof resulting in the upper bound of $O(\sqrt{\log k}\,\theta)$ for the disagreement.

Single Face Perturbation Bound: Claim 5.10: We will use the following notation for the positive orthant indicator $R(\boldsymbol{z}) = \prod_{i=1}^{k} \mathbb{1}\{\boldsymbol{z}_i \geq 0\}$. Notice that the homogeneous polyhedral cone C_1 can be written as $C_1(\boldsymbol{x}) = R(\boldsymbol{V}\boldsymbol{x}) = R(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x})$. Claim 5.10 below shows that the disagreement of two cones that differ on a single normal vector is bounded by above by the Gaussian surface area of a particular face F_1 times a logarithmic blow-up factor $\sqrt{\log(1/\Gamma(F_1)+1)}$.

Claim 5.10. *Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k \in \mathbb{R}^d$ and $\boldsymbol{r} \in \mathbb{R}^d$ with $\theta(\boldsymbol{v}_1, \boldsymbol{r}) \leq \theta$ for some sufficiently small $\theta \in (0, \pi/2)$. Let F_1 be the face with $\boldsymbol{v}_1 \cdot \boldsymbol{x} = 0$ of the cone $R(\boldsymbol{V}\boldsymbol{x})$ and $c > 0$ be some universal constant. Then,*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}\left[R(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \neq R(\boldsymbol{r} \cdot \boldsymbol{x}, \boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x})\right]$$

$$\leq c \cdot \theta \cdot \Gamma(F_1)\sqrt{\log\left(\frac{1}{\Gamma(F_1)} + 1\right)}.$$

Proof Sketch of Claim 5.10. Since the constraints $\boldsymbol{v}_2 \cdot \boldsymbol{x} \geq 0, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x} \geq 0$ are common in the two cones, we have that $R(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \neq R(\boldsymbol{r} \cdot \boldsymbol{x}, \boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x})$ only when the first "halfspaces" disagree, i.e., when $(\boldsymbol{v}_1 \cdot \boldsymbol{x})(\boldsymbol{r} \cdot \boldsymbol{x}) < 0$. Thus, we have that the LHS probability of Claim 5.10 is equal to

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_d}\left[R(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \cdot \mathbb{1}\{(\boldsymbol{v}_1 \cdot \boldsymbol{x})(\boldsymbol{r} \cdot \boldsymbol{x}) < 0\}\right]. \tag{5.6}$$

This expectation contains two terms: the term $R(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots \boldsymbol{v}_k \cdot \boldsymbol{x})$ that contains the last $k-1$ common constrains of the two cones and the region where the first two halfspaces disagree, i.e., the set $\{\boldsymbol{x} : (\boldsymbol{v}_1 \cdot \boldsymbol{x})(\boldsymbol{r} \cdot \boldsymbol{x}) < 0\}$. In order to upper bound this integral in terms of the angle θ, we observe that (for θ sufficiently small) it is not hard to show (see Appendix B) that the disagreement region, which is itself a (non-convex) cone, is a subset of the region $\{\boldsymbol{x} : |\boldsymbol{v}_1 \cdot \boldsymbol{x}| \leq 2\theta |\boldsymbol{q} \cdot \boldsymbol{x}|\}$, where $\boldsymbol{q}$ the normalized projection of $\boldsymbol{r}$ onto the orthogonal complement of $\boldsymbol{v}_1$, i.e., $\boldsymbol{q} = \mathrm{proj}_{\boldsymbol{v}_1^\perp} \boldsymbol{r} / \|\mathrm{proj}_{\boldsymbol{v}_1^\perp} \boldsymbol{r}\|_2$. Therefore, we have that the integral of Eq. (5.6) is at most

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d} \left[R(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \, \mathbb{1}\{|\boldsymbol{v}_1 \cdot \boldsymbol{x}| \leq 2\theta |\boldsymbol{q} \cdot \boldsymbol{x}|\} \right] .$$

This is where the definition of the Gaussian surface area appears. In fact, we have to compute the derivative of the above expression (which is a function of θ) with respect to θ and evaluate it at $\theta = 0$. The idea behind this computation is that we can upper bound probability mass of the cone disagreement, i.e., the term $\mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{N}_d} [R(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \neq R(\boldsymbol{r} \cdot \boldsymbol{x}, \boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x})]$ by its derivative with respect to θ (evaluated at 0) times θ by introducing $o(\theta)$ error. Hence, it suffices to upper bound the value of this derivative at 0, which is:

$$2 \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d} \left[R(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \, |\boldsymbol{q} \cdot \boldsymbol{x}| \, \delta(|\boldsymbol{v}_1 \cdot \boldsymbol{x}|) \right] ,$$

where δ is the Dirac delta function. Notice that, if we did not have the term $|\boldsymbol{q} \cdot \boldsymbol{x}|$, the above expression would be exactly equal to two times the Gaussian surface area of the face with $\boldsymbol{v}_1 \cdot \boldsymbol{x} = 0$, i.e., it would be equal to $2\Gamma(F_1)$. We now show that this extra term of $|\boldsymbol{q} \cdot \boldsymbol{x}|$ can only increase the above surface integral by at most a logarithmic factor. For some ξ to be decided, we have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_d}[R(\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x})\ |\boldsymbol{q}\cdot\boldsymbol{x}|\ \delta(|\boldsymbol{v}_1\cdot\boldsymbol{x}|)] = \int_{\boldsymbol{x}\in F_1}\phi_d(\boldsymbol{x})|\boldsymbol{q}\cdot\boldsymbol{x}|d\mu(\boldsymbol{x})\\
&\le \int_{\boldsymbol{x}\in F_1}\phi_d(\boldsymbol{x})|\boldsymbol{q}\cdot\boldsymbol{x}|\mathbb{1}\{|\boldsymbol{q}\cdot\boldsymbol{x}|\le\xi\}d\mu(\boldsymbol{x}) + \int_{\boldsymbol{x}\in F_1}\phi_d(\boldsymbol{x})|\boldsymbol{q}\cdot\boldsymbol{x}|\mathbb{1}\{|\boldsymbol{q}\cdot\boldsymbol{x}|\ge\xi\}d\mu(\boldsymbol{x})\\
&\le \xi\int_{\boldsymbol{x}\in F_1}\phi_d(\boldsymbol{x})d\mu(\boldsymbol{x}) + \int_{\boldsymbol{x}\in F_1}\phi_d(\boldsymbol{x})|\boldsymbol{q}\cdot\boldsymbol{x}|\mathbb{1}\{|\boldsymbol{q}\cdot\boldsymbol{x}|\ge\xi\}d\mu(\boldsymbol{x})\,,
\end{aligned}$$

where $d\mu(\boldsymbol{x})$ is the standard surface measure in $\mathbb{R}^d$. The first integral above is exactly equal to the Gaussian surface area of the face F_1. To bound from above the second term we can use the next claim showing that not a lot of mass of the face F_1 can concentrate on the region where $|\boldsymbol{q}\cdot\boldsymbol{x}|$ is very large. Its proof relies on standard Gaussian concentration arguments, and is provided in Appendix D.2.

Claim 5.11. *It holds that* $\int_{\boldsymbol{x}\in F_1}\phi_d(\boldsymbol{x})|\boldsymbol{q}\cdot\boldsymbol{x}|\mathbb{1}\{|\boldsymbol{q}\cdot\boldsymbol{x}|\ge\xi\}d\mu(\boldsymbol{x}) \le O(\exp(-\xi^2/2))\,.$

Using the above result, we get that

$$\begin{aligned}
&\frac{d}{d\theta}\Big(\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_d}[R(\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x})\ \mathbb{1}\{|\boldsymbol{v}_1\cdot\boldsymbol{x}|\le 2\theta|\boldsymbol{q}\cdot\boldsymbol{x}|\}]\Big)\Big|_{\theta=0}\\
&\le O(\xi)\ \Gamma(F_1) + O(\exp(-\xi^2/2))\,.
\end{aligned}$$

By picking $\xi = \Theta(\sqrt{\log(1+1/\Gamma(F_1))})$, the result follows since, up to introducing $o(\theta)$ error, we can bound the term $\mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}_d}[R(\boldsymbol{v}_1\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x}) \neq R(\boldsymbol{r}\cdot\boldsymbol{x},\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x})]$ by its derivative with respect to θ, evaluated at 0, times θ. □

Further Related Work

Robust Supervised Learning. We start with a summary of prior work on PAC learning with Massart noise. The Massart noise model was formally defined in Massart and Nédélec (2006) but similar variants had been defined by Vapnik, Sloan and Rivest Vapnik (1982); Sloan (1988, 1992); Rivest and Sloan (1994b); Sloan (1996). This model is a strict extension of the Random Classification Noise (RCN) model Angluin and Laird (1988), where the label noise is uniform, i.e., context-independent and is a special case of the agnostic model Haussler (2018); Kearns et al. (1994b), where the label noise is fully adversarial and computational barriers are known to exist Guruswami and Raghavendra (2009); Feldman et al. (2006b); Daniely (2016b); Diakonikolas et al. (2020d); Goel et al. (2020c); Diakonikolas et al. (2021e); Hsu et al. (2022). Our work partially builds upon on the algorithmic task of PAC learning halfspaces with Massart noise Balcan and Haghtalab (2020b). In the distribution-independent setting, known efficient algorithms Diakonikolas et al. (2019a); Chen et al. (2020b); Diakonikolas et al. (2021c) achieve error $\eta + \epsilon$ and the works of Diakonikolas and Kane (2020); Nasser and Tiegel (2022) indicate that this error bound is the best possible in the Statistical Query model Kearns (1998). This lower bound motivates the study of the distribution-specific setting (which is also the case of our work). There is an extensive line of work in this direction: **?**Awasthi et al. (2016b); Yan and Zhang (2017b); Zhang et al. (2017c); Balcan and Zhang (2017b); Mangoubi and Vishnoi (2019b); Diakonikolas et al. (2020e); Zhang et al. (2020a); Zhang and Li (2021) with the currently best algorithms succeeding for all $\eta < 1/2$ with a sample and computational complexity $\text{poly}(d, 1/\epsilon, 1/(1-2\eta))$ under a class of distributions including isotropic log-concave distributions. For details, see Diakonikolas et al. (2021d). In this work we focus on Gaussian marginals but some of our results extend to larger distribution classes.

Label Ranking. Our work lies in the area of Label Ranking, which has received significant attention over the years Shalev-Shwartz (2007); Hüllermeier et al. (2008); Cheng and Hüllermeier (2008); Har-Peled et al. (2003); Fürnkranz et al. (2008); Dekel et al. (2003). There are multiple approaches for tackling this problem (see Vembu and Gärtner (2010), Zhou et al. (2014b)). Some of them are based on probabilistic models Cheng and Hüllermeier (2008); Cheng et al. (2010); Grbovic et al. (2012); Zhou et al. (2014a) or may be tree based, such as decision trees Cheng et al. (2009), entropy based ranking trees and forests Rebelo de Sá et al. (2015); de Sá et al. (2017), bagging techniques Aledo et al. (2017) and random forests Zhou and Qiu (2018). There are also works focusing on supervised clustering Grbovic et al. (2013). Finally, Cheng and Hüllermeier (2008); Cheng et al. (2010,

2009) adopt an instance-based approaches using nearest neighbors approaches. The above results are industrial. From a theoretical perspective, LR has been mainly studied from a statistical learning theory framework Clémençon and Vogel (2020); Clémençon et al. (2018); Korba et al. (2018, 2017). Fotakis et al. (2021b) provide some computational guarantees for the performance of decision trees in the noiseless case and some experimental results on the robustness of random forests to noise. The setting of Djuric et al. (2014) is close to ours but is investigated from an experimental standpoint. We remark that while reducing LR to multiple binary classification tasks has been used in prior literature Hüllermeier et al. (2008); Cheng and Hüllermeier (2012); Fotakis et al. (2021b), standard reductions can not tolerate noise in rankings (nevertheless, from an experimental perspective, e.g., random forests seem robust to noise but lack formal theoretical guarantees). Our reduction crucially relies on the existence of efficient learning algorithms for binary linear classification with Massart noise.

Part II

Learning From Truncated or Coarse Data

6 LEARNING TRUNCATED GAUSSIANS

6.1 Formal Statement of Results

Preliminaries

Notation Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, we define $\boldsymbol{A}^\flat \in \mathbb{R}^{d^2}$ to be the standard vectorization of $\boldsymbol{A}$. Let also $\mathcal{Q}_d$ be the set of all the symmetric $d \times d$ matrices. The *Frobenius norm* of a matrix $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\|_F = \left\|\boldsymbol{A}^\flat\right\|_2$.

Gaussian Distribution. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with the following probability density function

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{6.1}$$

Also, let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S)$ denote the *probability mass of a measurable set* S under this Gaussian measure. We shall also denote by $\mathcal{N}_0$ the standard Gaussian, whether it is single or multidimensional will be clear from the context.

Truncated Gaussian Distribution. Let $S \subseteq \mathbb{R}^d$ be a subset of the d-dimensional Euclidean space, we define the *S-truncated normal distribution* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ the normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ conditioned on taking values in the subset S. The probability density function of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ is the following

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S; \boldsymbol{x}) = \frac{\mathbb{1}_S(\boldsymbol{x})}{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S)} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}). \tag{6.2}$$

We will assume that the covariance matrix Σ is full rank. We can easily detect the case where Σ is not full rank and solve the estimation problem in the linear subspace of samples.

The core complexity measure of Borel sets in $\mathbb{R}^d$ that we use is the notion of Gaussian Surface Area defined below.

Definition 6.1 (Gaussian Surface Area). *For a Borel set $A \subseteq \mathbb{R}^d$, $\delta \geq 0$ let $A_\delta = \{x : \boldsymbol{h}rmdist(x, A) \leq \delta\}$. The Gaussian surface area of A is*

$$\Gamma(A) = \liminf_{\delta \to 0} \frac{\mathcal{N}_0(A_\delta \setminus A)}{\delta}.$$

We define the Gaussian surface area of a family of sets $\mathcal{C}$ to be $\Gamma(\mathcal{C}) = \sup_{C \in \mathcal{C}} \Gamma(C)$.

Problem formulation

Given samples from a truncated Gaussian $\mathcal{N}_S^* \triangleq \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$, our goal is to learn the parameters $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ and recover the set S. We denote by $\alpha^* = \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S)$, the total mass contained in set S by the untruncated Gaussian $\mathcal{N}^* \triangleq \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. Throughout this paper, we assume that we know an absolute constant $\alpha > 0$ such that

$$\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S) = \alpha^* \geq \alpha. \tag{6.3}$$

We first analyze the sample compexity of learning the true Gaussian parameters when the truncation set has bounded VC-dimension. In particular, we show that the overhead over the d^2/ϵ^2 samples (which is the sample compexity of learning the parameters of the Gaussian without truncation) is proportional to the VC dimension of the class.

Theorem 6.2. *Let $\mathcal{S}$ be a family of sets of finite VC dimension, and let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ be a truncated Gaussian distribution such that $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) \geq \alpha$. Given N*

$$N = \mathrm{poly}(1/\alpha)\, \widetilde{O}\left(\frac{d^2}{\epsilon^2} + \frac{\mathrm{VC}(\mathcal{S})}{\epsilon}\right),$$

samples, then, with probability at least 99%, it is possible to identify $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ that satisfy

$$d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})) \leq \epsilon$$

and $\left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})\right\|_2 \leq \epsilon$ *and* $\left\|\boldsymbol{I} - \boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\right\|_F \leq \epsilon$*.*

We now show our algorithmic results under the assumption that the untruncated Gaussian $\mathcal{N}^*$ is known to be in near-isotropic position.

Definition 6.3 (Near-Isotropic Position)**.** *Let* $\boldsymbol{\mu} \in \mathbb{R}^d$*,* $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ *be a positive semidefinite symmetric matrix and* $a, b > 0$*. We say that* $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *is in* (a, b)*-isotropic position if the following hold.*

$$\|\boldsymbol{\mu}\|_2^2 \leq a, \quad \|\boldsymbol{\Sigma} - \boldsymbol{I}\|_F^2 \leq a, \quad (1 - b)\boldsymbol{I} \preceq \boldsymbol{\Sigma} \preceq \frac{1}{1 - b}\boldsymbol{I}$$

We later transform the more interesting case with an unknown mean and an unknown diagonal covariance matrix to the isotropic case.

Theorem 6.4. *Let* $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ *be a d-dimensional Gaussian distribution that is in*

$$(O(\log(1/\alpha^*), 1/16)\text{-}\textit{isotropic position}$$

and consider a set S *such that* $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S) \geq \alpha$*. There exists an algorithm such that for all* $\epsilon > 0$*, the algorithm uses* $n > d^{\mathrm{poly}(1/\alpha)\frac{\Gamma^2(S)}{\epsilon^8}}$ *samples and produces, in* $\mathrm{poly}(n)$ *time, estimates that, with probability at least* 99%*, satisfy* $d_{TV}(\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \leq \epsilon$*.*

We next investigate the sample complexity of the problem of estimating the parameters of a truncated Gaussian using a different approach that does not depend on the VC dimension of the family $\mathcal{S}$ of the truncation sets to be finite. For example, we settle the sample complexity of learning the parameters of a Gaussian distribution truncated at an unknown convex set (recall that the class of convex sets has infinite VC dimension). Our method relies on finding a tuple $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$ of parameters so that the moments of the corresponding truncated Gaussian $\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$ are all close to the moments of the unknown truncated Gaussian distribution, for which we have unbiased estimates using samples. The main question that we need to answer to determine the sample complexity of this problem is how many moments are needed to be matched in order to be sure that our guessed parameters are close to the parameters of the unknown truncated Gaussian. We state now the main result. Its proof is based on Lemma 6.31 and can be found in Appendix E.6.

Theorem 6.5 (Moment Matching)**.** *Let $\mathcal{S}$ be a family of subsets of $\mathbb{R}^d$ of bounded Gaussian surface area $\Gamma(\mathcal{S})$. Moreover, assume that if T is an affine map and $T(\mathcal{S}) = \{T(S) : S \in \mathcal{S}\}$ is the family of the images of the sets of $\mathcal{S}$, then it holds $\Gamma(T(\mathcal{S})) = O(\Gamma(\mathcal{S}))$. For some $S \in \mathcal{S}$, let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ be an unknown truncated Gaussian. $d^{O(\Gamma(\mathcal{S})/\epsilon^4)}$ samples are sufficient to find parameters $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S}$ such that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S), \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})) \leq \epsilon$.*

Finally, we present an information-theoretic lower bound showing that there exists families of truncation sets whose sample complexity depends exponentially on their Gaussian Surface Area.

Theorem 6.6. *There exists a family of sets $\mathcal{S}$ with $\Gamma(\mathcal{S}) = O(d)$ such that any algorithm that draws m samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}, S)$ and computes an estimate $\widetilde{\boldsymbol{\mu}}$ with $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq 1$ must have $m = \Omega(2^{d/2})$.*

Simulations. In addition to the theoretical guarantees of our algorithm, we empirically evaluate its performance using simulated data. We present the results that we get in Figure 6.3, where one can see that even when the truncation set is complex, our algorithm finds an accurate estimation of the mean of the untruncated distribution. Observe that our algorithm succeeds in estimating the true mean of the input distribution despite the fact that the set is unknown and the samples look similar in both cases.

6.2 Identifiability with Bounded VC dimension

In this section we analyze the sample compexity of learning the true Gaussian parameters when the truncation set has bounded VC-dimension. In particular we show that the overhead over the d^2/ϵ^2 samples (which is the sample compexity of learning the parameters of the Gaussian without truncation) is proportional to the VC dimension of the class. For convenience, we restate Theorem 6.2 below.

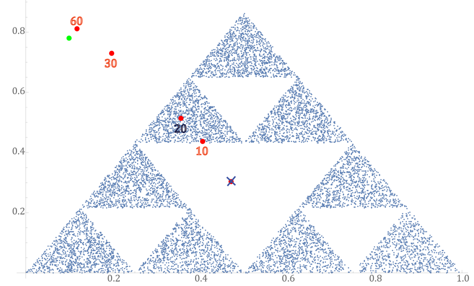


Figure 6.1: Execution of our algorithm for isotropic Gaussian distribution with $\boldsymbol{\mu}^* = (0.1, 0.78)$ and $\boldsymbol{\mu}_S = (0.48, 0.32)$.

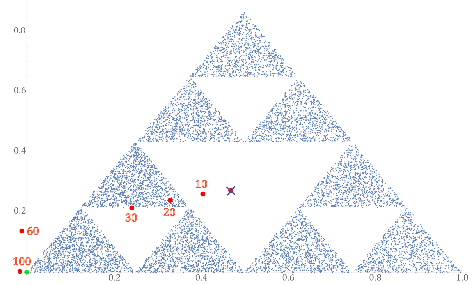


Figure 6.2: Execution of our algorithm for isotropic Gaussian distribution with $\boldsymbol{\mu}^* = (0, 0)$ and $\boldsymbol{\mu}_S = (0.47, 0.27)$.

Figure 6.3: Illustration of the results of our algorithm for an unknown truncation set. The × sign corresponds to the conditional mean of the truncated distribution, while the green point corresponds to the true mean and the red points correspond to the estimated true mean depending on the degree of the Hermite polynomials that are being used by the algorithm.

Theorem 6.7. *Let $\mathcal{S}$ be a family of sets of finite VC dimension, and let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ be a truncated Gaussian distribution such that $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) \geq \alpha$. Given N*

$$N = \mathrm{poly}(1/\alpha)\, \widetilde{O}\left(\frac{d^2}{\epsilon^2} + \frac{\mathrm{VC}(\mathcal{S})}{\epsilon}\right),$$

samples, then, with probability at least 99%, it is possible to identify $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ that satisfy

$$d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})) \leq \epsilon$$

and $\left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}})\right\|_2 \leq \epsilon$ *and* $\left\|\boldsymbol{I}-\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\right\|_F \leq \epsilon$.

Our algorithm works by first learning the truncated distribution within total variation distance ϵ. To do this, we first assume that we know the mean and covariance of the underlying Gaussian by guessing the parameters and accurately learn the underlying set. After drawing $N = \Theta(\frac{\mathrm{VC}(\mathcal{S})\log(1/\epsilon)}{\epsilon})$ samples from the distribution, any set in the class that contains the samples will only exclude at most an ϵ fraction of the total mass. Picking the set $\widetilde{S}$ that maximizes the likelihood of those samples, i.e. the set with minimum mass according to the guessed Gaussian distribution, guarantees that the total variation distance between the learned truncated distribution and the true is at most ϵ, if the guess of the parameters was accurate (Lemma 6.8). The proof of Lemma 6.8 can be found in Appendix E.2.

Lemma 6.8. *Let $\mathcal{S}$ be a family of subsets in $\mathbb{R}^d$ and Let $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S^*) = \mathcal{N}_S^*$ be a Normal distribution truncated on the set $S^* \in \mathcal{S}$. Fix $\epsilon \in (0,1), \delta \in (0,1/4)$ and let*

$$N = O\left(\frac{\mathrm{VC}(\mathcal{S})\log(1/\epsilon)}{\epsilon} + \log\left(\frac{1}{\delta}\right)\right)$$

Moreover, let $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}$ be such that $d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}}),\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})) \leq \epsilon$. Assume that we draw N samples $\boldsymbol{x}_i$ from $\mathcal{N}_{S^*}$, Let $\widetilde{S}$ be the solution of the problem

$$\min_S \mathcal{N}(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}};S) \;\; \textit{subject to } x_i \in S \textit{ for all } i \in [n]$$

Then with probability at least $1-\delta$ we have $d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}},\widetilde{S}),\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)) \leq 3\epsilon/(2\alpha)$.

This is because the total variation distance between two densities f and g can be written as $\int(f(x)-g(x))\mathbb{1}_{f(x)>g(x)}dx$. Note that by choosing the set of the smallest mass consistent with the samples, we guarantee that the guess will have higher density at every point apart from those outside the support $\widetilde{S}$. However, as we argued the outside mass is at most ϵ with respect to the true distribution which gives the bound in the total variation distance.

To remove the assumption that the true parameters are known, we build a cover of all possible mean and covariance matrices that the underlying Gaussian might have and run the tournament from Daskalakis and Kamath (2014) to identify the best one (Lemma E.10). While there are $(d/\epsilon)^{O(d^2)}$ such parameters, the number of samples needed for running the tournament is only logarithmic which shows that an additional $\widetilde{O}(d^2/\epsilon^2)$ are sufficient to find a hypothesis in total variation distance ϵ (Lemma 6.9). The proof of Lemma 6.9 can be found in Appendix E.2.

Lemma 6.9. *Let $S \in \mathcal{S}$ be a subset of $\mathbb{R}^d$ and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ be the corresponding truncated normal distribution. Then $\widetilde{O}\left(\mathrm{VC}(\mathcal{S})/\epsilon + d^2/\epsilon^2\right)$ samples are sufficient to find parameters $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S}$ such that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S), \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})) \leq \epsilon$ with probability at least 99%.*

We finally argue that the ϵ error in total variation of the truncated distributions translates to an $O(\epsilon)$ bound in total variation distance of the untruncated distributions (Lemma 6.10). We show that this is true in general and does not depend on the complexity of the set. To prove this statement, we consider two Gaussians with parameters that are far from each other and construct the worst possible set to make their truncated distributions as close as possible. We show that under the requirement that the set contains at least α mass, the total variation distance of the truncated distributions will be large.

Lemma 6.10 (Total Variation of Truncated Normals)**.** *Let $D_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, S_1)$ and $D_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, S_2)$ be two truncated Normal distributions such that*

$$\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1; S_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2; S_2) \geq \alpha.$$

Then,

$$d_{\mathrm{TV}}(D_1, D_2) \geq C_\alpha \, d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)),$$

where $C_\alpha < \alpha/8$ is a positive constant that only depends on α, $C_\alpha = \Omega(\alpha^3)$.

Proof. Without loss of generality we assume that $D_1 = \mathcal{N}(\mathbf{0}, \boldsymbol{I}, S_1)$ and $D_2 = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, S_2)$, where $\boldsymbol{\Lambda}$ is a diagonal matrix. We want to find the worst sets S_1, S_2 so that $d_{\mathrm{TV}}(D_1, D_2)$ is small. If $D_1(S_1 \setminus S_2) \geq \alpha/2$ then the statement holds. Therefore, we consider the set $S = S_1 \cap S_2$ and relax the constraint that the truncated Gaussian D_2 integrates to 1. Taking into account the fact that the set $S = S_1 \cap S_2$ must have at least some mass $\alpha/2$ with respect to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, the following optimization problem provides a lower bound on the total variation distance of D_1 and D_2.

$$\begin{aligned} \min_{S \in \mathcal{S}, \beta > 0} \quad & \frac{1}{\alpha} \int |\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) - \frac{\alpha}{\beta} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}; \boldsymbol{x})| \, \mathbb{1}_S(\boldsymbol{x}) \mathrm{d}x \\ \text{subj. to} \quad & \int \mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) \, \mathbb{1}_S(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \geq \alpha/2, \end{aligned}$$

For any fixed $\beta > 0$ this is a fractional knapsack problem and therefore we should include in the set the points x in order of increasing ratio of weight that is contribution to the L_1 error $|\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) - \frac{\alpha}{\beta} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}; \boldsymbol{x})|$, over value, that is density $\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x})$ until we reach some threshold T. Therefore, the set is defined to be

$$S = \left\{ \boldsymbol{x} \in \mathbb{R}^d : \frac{|\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) - \frac{\alpha}{\beta} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}; \boldsymbol{x})|}{\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x})} \leq T \right\} = \left\{ \boldsymbol{x} \in \mathbb{R}^d : |1 - \exp(p(\boldsymbol{x}))| \leq T \right\},$$

where $p(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{x})^T \boldsymbol{\Lambda}^{-1} (\boldsymbol{\mu} - \boldsymbol{x}) + \frac{1}{2} \boldsymbol{x}^T \boldsymbol{x} + \log(\alpha / (\sqrt{|\boldsymbol{\Lambda}|} \beta))$. Using Theorem E.4 for the degree 2 polynomial $p(\boldsymbol{x})$ and setting $q = 4$, $\gamma = \alpha^2 (\mathbf{E}_{x \sim \mathcal{N}_0} \, p^2(x))^{1/2} / (256 C^2)$, where C is the absolute constant of Theorem E.4, we get that

$$\mathcal{N}_0(\{\boldsymbol{z} : |p(\boldsymbol{z})| \leq \gamma\}) \leq \frac{\alpha}{4}.$$

To simplify notation set $Q = \{\boldsymbol{z} : |p(\boldsymbol{z})| \leq \gamma\}$. Therefore, for any $\boldsymbol{x}$ in the remaining $\alpha/4$ mass of the set S we know that $|p(\boldsymbol{x})| \geq \gamma$. Next, we lower bound γ in terms of the distance of the parameters of the two Gaussians. We have

$$
\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[p^2(\boldsymbol{x})] \geq \mathrm{Var}_{\boldsymbol{x}\sim\mathcal{N}_0}[p(\boldsymbol{x})] &= \mathrm{Var}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{x})^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{\mu}-\boldsymbol{x})+\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right] \\
&= \mathrm{Var}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\sum_{i=1}^{d}\left(\frac{\mu_i}{\lambda_i}x + x^2\frac{(1-1/\lambda_i)}{2}\right)\right] \\
&= \sum_{i=1}^{d}\mathrm{Var}_{x\sim\mathcal{N}(0,1)}\left[\frac{\mu_i}{\lambda_i}x + x^2\frac{(1-1/\lambda_i)}{2}\right] \\
&= \sum_{i=1}^{d}\frac{1}{2}\left(\frac{1}{\lambda_i}-1\right)^2+\frac{\mu_i^2}{\lambda_i^2} \\
&= \frac{1}{2}\left\|\boldsymbol{\Lambda}^{-1}-\boldsymbol{I}\right\|_F^2+\left\|\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\mu}\right\|_2^2
\end{aligned}
$$

Therefore, using the inequality $\sqrt{2}\sqrt{x+y} \geq \sqrt{x}+\sqrt{y}$ we obtain

$$
\begin{aligned}
\gamma &\geq \frac{\alpha^2}{256\sqrt{2}C^2}\left(\frac{1}{\sqrt{2}}\left\|\boldsymbol{\Lambda}^{-1}-\boldsymbol{I}\right\|_F+\left\|\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\mu}\right\|_2\right) \\
&\geq \frac{\alpha^2}{256C^2}d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1),\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)),
\end{aligned}
$$

where we used Lemma E.1. Assume first that $\gamma \leq 1$. We have that the L_1 distance between the functions $f(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})\mathbb{1}_S(\boldsymbol{x})$ and $g(\boldsymbol{x}) = \frac{\alpha}{\beta}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Lambda};\boldsymbol{x})\mathbb{1}_S(\boldsymbol{x})$ is

$$
\begin{aligned}
\int |f(\boldsymbol{x})-g(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[|1-\exp(p(\boldsymbol{x}))|\mathbb{1}_S(\boldsymbol{x})] \geq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\frac{|p(\boldsymbol{x})|}{2}\mathbb{1}_{S\setminus Q}(\boldsymbol{x})\right] \\
&\geq \gamma \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\mathbb{1}_{S\setminus Q}(\boldsymbol{x})\right] \geq \frac{\alpha\gamma}{4} \geq C_\alpha d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1),\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)),
\end{aligned}
$$

where for the first inequality we used the inequality $|1-\mathrm{e}^x| \geq |x|/2$ for $|x| \leq 1$. Note that $C_a = \Omega(\alpha^3)$. If $\gamma > 1$ we have

$$
\int |f(\boldsymbol{x})-g(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[|1-\exp(p(\boldsymbol{x}))|\mathbb{1}_S(\boldsymbol{x})] \geq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\frac{1}{2}\mathbb{1}_{S\setminus Q}(\boldsymbol{x})\right] \geq \alpha/8,
$$

where we used the inequality $|1-\mathrm{e}^x| \geq 1/2$ for $|x| > 1$. □

6.3 Estimation Algorithm for bounded Gaussian Surface Area

In this section, we present the main steps of our estimation algorithm. In later sections, we provide details of the individual components. The algorithm can be thought of in 3 stages.

First Stage In the first stage, our goal is to learn a weighted characteristic function of the underlying set. Even though we cannot access the underlying set directly, for any given function f we can evaluate the expectation $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}[f(\boldsymbol{x})]$ using truncated samples.

This expectation can be equivalently written as $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}[f(\boldsymbol{x})\psi(\boldsymbol{x})]$ for the function

$$\psi(\boldsymbol{x}) \triangleq \frac{\mathbb{1}_S(\boldsymbol{x})}{\alpha^*}\frac{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}{\mathcal{N}(\mathbf{0},\boldsymbol{I};\boldsymbol{x})} = \frac{\mathbb{1}_S(\boldsymbol{x})}{\alpha^*}\frac{\mathcal{N}^*(\boldsymbol{x})}{\mathcal{N}_0(\boldsymbol{x})}.$$

By evaluating the above expectation for different functions f corresponding to the Hermite polynomials $H_V(\boldsymbol{x})$, we can recover $\psi(\boldsymbol{x})$, through its Hermite expansion:

$$\psi(\boldsymbol{x}) = \sum_{V\in\mathbb{N}^d} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[H_V(\boldsymbol{x})\psi(\boldsymbol{x})]H_V(\boldsymbol{x}) = \sum_{V\in\mathbb{N}^d} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}[H_V(\boldsymbol{x})]H_V(\boldsymbol{x}).$$

Of course, it is infeasible to calculate the Hermite expansion for any $V \in \mathbb{N}^d$. In Section 6.3, we show that by estimating only terms of degree at most k, we can achieve a good approximation to ψ where the error depends on the Gaussian surface area of the underlying set S. To do this, we show that most of the mass of the coefficients $c_V = \mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[H_V(\boldsymbol{x})\psi(\boldsymbol{x})]$ is concentrated on low degree terms, i.e. $\sum_{|V|>k} c_V^2$ is significantly small. Moreover, we show that even though we can only estimate the coefficients c_V through sampling, the sampling error is significantly small.

Overall, after the first stage, we obtain a non-negative function ψ_k that is close to ψ. The approximation error guarantees are given in Theorem 6.18.

Second Stage Given the function ψ_k that was recovered in the first stage, our goal is to decouple the influence of the set $\frac{\mathbb{1}_S(\boldsymbol{x})}{\alpha^*}$ and the influence of the underlying Gaussian distribution which corresponds to the multiplicative term $\frac{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}{\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})}$. This would be easy if we had the exact function ψ in hand. In contrast, for the polynomial function ψ_k the problem is significantly challenging as it is only close to ψ on average but not pointwise.

To perform the decoupling and identify the underlying Gaussian we explicitly multiply the function ψ_k with a corrective term of the form $\frac{\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})}{\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};\boldsymbol{x})}$. We set up an optimization problem seeking to minimize the function $C(\boldsymbol{\mu},\boldsymbol{\Sigma})\, \mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}_S^*}[\frac{\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})}{\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};\boldsymbol{x})}\psi_k(\boldsymbol{x})]$ with an appropriate choice of $C(\boldsymbol{\mu},\boldsymbol{\Sigma})$ so that the unique solution corresponds to $(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)$. Under a reparameterization of $(\boldsymbol{u},\boldsymbol{B}) = (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1})$, we show that the corresponding problem is *strongly* convex. Still, optimizing it directly is non-trivial as it involves taking the expectation with respect to the unknown truncated Gaussian. Instead, we perform stochastic gradient descent (SGD) and show that it quickly converges in few steps to point close to the true minimizer (Algorithm 11).

This allows us to recover parameters $(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ so that the total variation distance between the recovered and the true (untruncated) Gaussian is very small, i.e. $d_{TV}\left(\mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}),\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)\right) \le \epsilon$. Theorem 6.4 describes the guarantees of the second stage. Further details are provided in Section 6.3.

Third Stage Given the weighted indicator function ψ_k and the recovered Gaussian $\mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$, we move on to recover the underlying set S. To do this, we compute the function $\frac{\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})}{\mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}};\boldsymbol{x})}\psi_k(\boldsymbol{x})$ and set a threshold at $1/2$. It is easy to check that if there were no errors, i.e. $\psi_k = \psi$ and $d_{TV}\left(\mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}),\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)\right) = 0$, that this thresholding step would correctly identify the set. In Section 6.3 we bound the error guarantees of this approach. We show that it is possible to obtain an estimate $\hat{S}$ of the underlying set so that the mass of the symmetric difference with the true Gaussian is small, i.e. $\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;S\triangle\hat{S}) < \epsilon$. Overall, our algorithm requires at most

$d^{\text{poly}(1/\alpha, 1/\epsilon)\Gamma^2(S)}$, where $\Gamma(S)$ is the Gaussian surface area of the set S and α is a lower-bound on the mass that is assigned by the true Gaussian on the set S. The running time of our algorithm is linear in the number of samples.

The guarantees of the algorithm We first show our algorithmic results under the assumption that the untruncated Gaussian $\mathcal{N}^*$ is known to be in near-isotropic position, see Definition 6.3. We later transform the more interesting case with an unknown mean and an unknown diagonal covariance matrix to the isotropic case. We next restate Theorem 6.4 for convenience.

Theorem 6.11. *Let $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ be a d-dimensional Gaussian distribution that is in $(O(\log(1/\alpha^*)), 1/16)$-isotropic position and consider a set S such that $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S) \geq \alpha$. There exists an algorithm such that for all $\epsilon > 0$, the algorithm uses $n > d^{\text{poly}(1/\alpha)\frac{\Gamma^2(S)}{\epsilon^8}}$ samples and produces, in* poly(n) *time, estimates that, with probability at least* 99%, *satisfy $d_{TV}(\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \leq \epsilon$.*

We can apply this theorem to estimate the parameters of any Gaussian distribution with an unknown mean and an unknown diagonal covariance matrix by bringing the Gaussian to an $(O(\log(1/\alpha^*)), 1/16)$-isotropic position. Lemma E.3 shows that with high probability, we can obtain initial estimates $\widetilde{\boldsymbol{\mu}}_S$ and $\widetilde{\boldsymbol{\Sigma}}_S$ so that $\|\boldsymbol{\Sigma}^{-1/2}(\widetilde{\boldsymbol{\mu}}_S - \boldsymbol{\mu}^*)\|_2^2 \leq O(\log \frac{1}{\alpha})$ and

$$\widetilde{\boldsymbol{\Sigma}}_S \succeq \Omega(\alpha^2)\boldsymbol{\Sigma}^*, \quad \text{and} \quad \left\| \boldsymbol{\Sigma}^{*-1/2} \widetilde{\boldsymbol{\Sigma}}_S \boldsymbol{\Sigma}^{*-1/2} - \boldsymbol{I} \right\|_F^2 \leq O(\log \frac{1}{\alpha}).$$

Given these estimates, we can transform the space so that $\widetilde{\boldsymbol{\mu}}_S = 0$, and $\widetilde{\boldsymbol{\Sigma}}_S = \boldsymbol{I}$. We note that after this transformation, the mean will be at the right distance from 0, while the eigenvalues λ_i of $\boldsymbol{\Sigma}^*$ will all be within the desired range $\frac{15}{16} \leq \lambda_i \leq \frac{16}{15}$ apart from at most $O(\log(1/\alpha))$. This is because the condition $\left\| \boldsymbol{\Sigma}^{*-1/2} \widetilde{\boldsymbol{\Sigma}}_S \boldsymbol{\Sigma}^{*-1/2} - \boldsymbol{I} \right\|_F^2 \leq O(\log \frac{1}{\alpha})$ implies that $\sum_i (1 - \frac{1}{\lambda_i})^2 \leq O(\log(1/\alpha))$. With this observation, since we know of the eigenvectors of Σ^*, we would be able to search over all possible corrections to the eigenvalues to bring the Gaussian in

$(O(\log(1/\alpha)), \frac{1}{16})$-isotropic position as required by Theorem 6.4. We only need to correct $O(\log(1/\alpha))$ of them.

We can form a space of candidate hypotheses for the underlying distribution, for each choice of $O(\log(1/\alpha))$ out of the d vectors along with the all possible scalings. These hypotheses are at most $d^{O(\log(1/\alpha))}$ times $(\log(1/\alpha))^{O(\log(1/\alpha))}$ for all possible scalings. Thus, there are at most $d^{O(\log(1/\alpha))}$ hypotheses. Running the algorithm for each one of them, we would learn at least one distribution and one set that is accurate according to the guarantees of Theorems 6.4. Running the generic hypothesis testing algorithm of Lemma E.10, we can identify one that is closest in total variation distance to the true distribution $\mathcal{N}t_S$. The sample complexity and runtime would thus only increase by at most $d^{O(\log(1/\alpha))}$. As we showed in Lemma 6.10, knowing the truncated Gaussian in total variation distance suffices to learn in accuracy ϵ the parameters of the untruncated distribution. We thus obtain as corollary, that we can estimate the parameters when the covariance is spherical or diagonal. The same results hold when one wants to recover the underlying set in these cases.

Learning a Weighted Characteristic Function

Our goal in this section is to recover using conditional samples from $\mathcal{N}_S^*$ a weighted characteristic function of the set S. In particular, we will show that it is possible to learn a good approximation to the function

$$\psi(\boldsymbol{x}) = \frac{\mathbb{1}_S(\boldsymbol{x})}{\alpha^*} \frac{\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; \boldsymbol{x})}{\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x})} = \frac{\mathbb{1}_S(\boldsymbol{x})}{\alpha^*} \frac{\mathcal{N}^*(\boldsymbol{x})}{\mathcal{N}_0(\boldsymbol{x})}. \tag{6.4}$$

We will later use the knowledge of this function to extract the unknown parameters and learn the set S.

Hermite Concentration

We start by showing that the function $\psi(\boldsymbol{x})$ admits strong Hermite concentration. This means that we can well-approximate $\psi(\boldsymbol{x})$ if we ignore the higher order terms in the Hermite expansion of $\psi(\boldsymbol{x})$.

Theorem 6.12 (Low Degree Approximation). *Let $S_k\psi$ denote the degree k Hermite expansion of function ψ defined in* (6.4). *We have that*

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[(S_k\psi(\boldsymbol{x})-\psi(\boldsymbol{x}))^2\right]=\sum_{|V|\geq k}\hat{\psi}(V)^2\leq \mathrm{poly}(1/\alpha)\left(\frac{\sqrt{\Gamma(S)}}{k^{1/4}}+\frac{1}{k}\right).$$

where $\Gamma(S)$ is the Gaussian surface area of S, and $a<\alpha^*$ is the absolute constant of (6.3).

We note that the Hermite expansion of ψ is well-defined as $\psi(\boldsymbol{x})\in L_2(\mathbb{R}^d,\mathcal{N}_0)$. This can be seen from the following lemma which will be useful in many calculations throughout the paper.

Lemma 6.13. *Let $\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ be two $(B,\frac{1-\delta}{2k})$-isotropic Gaussians for some parameters $B,\delta>0$ and $k\in\mathbb{N}$. It holds*

$$\exp\left(-\frac{13k^2}{\delta}B\right)\leq\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}\left[\left(\frac{\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1;\boldsymbol{x})}{\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2;\boldsymbol{x})}\right)^k\right]\leq\exp\left(\frac{13k^2}{\delta}B\right).$$

Lemma 6.13 applied for $\mathcal{N}_0$ and $\mathcal{N}^*$ for $k=2$ implies that $\psi(\boldsymbol{x})\in L_2(\mathbb{R}^d,\mathcal{N}_0)$.

To get the desired bound for Theorem 6.12 we use the following lemma, which allows us to bound the Hermite concentration of a function f through its noise stability.

Lemma 6.14. *For any function $f:\mathbb{R}^d\mapsto\mathbb{R}$ and parameter $\rho\in(0,1)$, it holds*

$$\sum_{|V|\geq 1/\rho}\hat{f}(V)^2\leq 2\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[f(\boldsymbol{x})^2-f(\boldsymbol{x})T_{1-\rho}f(\boldsymbol{x})\right]$$

Lemma 6.14 was originally shown in Kalai et al. (2005) for indicator functions of sets, but their proof extends to arbitrary real functions. We provide the proof in the appendix for completeness.

Using Lemma 6.14, we can obtain Theorem 6.12 by bounding the noise sensitivity of the function ψ. The following lemma directly gives the desired result.

Lemma 6.15. *For any $\rho \in (0,1)$,*

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\psi(\boldsymbol{x})^2 - \psi(\boldsymbol{x})T_{1-\rho}\psi(\boldsymbol{x})\right] \le \mathrm{poly}(1/\alpha)\left(\sqrt{\Gamma(S)}\rho^{1/4} + \rho\right).$$

To prove Lemma 6.15, we will require the following lemma whose proof is provided in the appendix.

Lemma 6.16. *Let $r(\boldsymbol{x}) \in L_2(\mathbb{R}^d, \mathcal{N}(\mathbf{0}, \boldsymbol{I}))$ be differentiable at every $\boldsymbol{x} \in \mathbb{R}^d$. Then*

$$\frac{1}{2}\mathop{\mathbf{E}}_{(x,z)\sim D_\rho}[(r(\boldsymbol{x}) - r(\boldsymbol{z}))^2] \le \rho \mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\|\nabla r(\boldsymbol{x})\|_2^2\right]$$

We now move on to the proof of Lemma 6.15.

Proof of Lemma 6.15 For ease of notation we define the following distribution

$$D_\rho = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{I} & (1-\rho)\boldsymbol{I} \\ (1-\rho)\boldsymbol{I} & \boldsymbol{I} \end{pmatrix}\right).$$

We also denote by $r(x) = \mathcal{N}^*(\boldsymbol{x})/\mathcal{N}_0(\boldsymbol{x})$ We can now write

$$\begin{aligned}
&2\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\psi(\boldsymbol{x})^2 - \psi(\boldsymbol{x})T_{1-\rho}\psi(\boldsymbol{x})\right] \\
&\quad= \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}\left[\psi(\boldsymbol{x})^2 - \psi(\boldsymbol{x})\psi(\boldsymbol{z})\right] \\
&\quad= \frac{1}{\alpha^{*2}}\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}[\mathbb{1}_S(\boldsymbol{x})r^2(\boldsymbol{x}) - \mathbb{1}_S(\boldsymbol{x})\mathbb{1}_S(\boldsymbol{z})r^2(\boldsymbol{x})] + \\
&\quad\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}[\mathbb{1}_S(\boldsymbol{x})\mathbb{1}_S(\boldsymbol{z})r^2(\boldsymbol{x}) - \mathbb{1}_S(\boldsymbol{x})\mathbb{1}_S(\boldsymbol{z})r(\boldsymbol{x})r(\boldsymbol{z})]
\end{aligned}$$

We bound each of the two terms separately. For the first term, using Schwarz's inequality we get

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}[\mathbb{1}_S(\boldsymbol{x})r^2(\boldsymbol{x}) - \mathbb{1}_S(\boldsymbol{x})\mathbb{1}_S(\boldsymbol{z})r^2(\boldsymbol{x})] \\
&\leq \left(\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}[\mathbb{1}_S(\boldsymbol{x})\mathbb{1}_{\bar{S}}(\boldsymbol{z})]\right)^{1/2}\left(\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}[r^4(\boldsymbol{x})]\right)^{1/2} \\
&\leq (\mathcal{N}S[S])^{1/2}\mathrm{poly}(1/\alpha) \leq \sqrt{\Gamma(S)}\rho^{1/4}\mathrm{poly}(1/\alpha)
\end{aligned}$$

where the bound on the expectation of $r^4(\boldsymbol{x})$ follows from Lemma 6.13 and the last inequality follows from Lemma E.9.

For the second term, we have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}[\mathbb{1}_S(\boldsymbol{x})\mathbb{1}_S(\boldsymbol{z})(r^2(\boldsymbol{x}) - r(\boldsymbol{x})r(\boldsymbol{z}))] \\
&= \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}\left[\mathbb{1}_S(\boldsymbol{x})\mathbb{1}_S(\boldsymbol{z})\left(\frac{r^2(\boldsymbol{x})}{2} + \frac{r^2(\boldsymbol{z})}{2} - r(\boldsymbol{x})r(\boldsymbol{z})\right)\right] \\
&= \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}\left[\mathbb{1}_S(\boldsymbol{x})\mathbb{1}_S(\boldsymbol{z})\frac{1}{2}(r(\boldsymbol{x}) - r(\boldsymbol{z}))^2\right] \\
&\leq \frac{1}{2}\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}\left[(r(\boldsymbol{x}) - r(\boldsymbol{z}))^2\right] \leq \rho \mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_0}[\|\nabla r(\boldsymbol{x})\|_2^2],
\end{aligned}$$

where the last inequality follows from Lemma 6.16. It thus suffices to bound the expectation of the gradient of r. We have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_0}[\|\nabla r(\boldsymbol{x})\|_2^2] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_0}\left[\left\|-\boldsymbol{\Sigma}^{*-1}(\boldsymbol{x} - \boldsymbol{\mu}^*) + \boldsymbol{x}\right\|_2^2 r^2(\boldsymbol{x})\right] \\
&\leq 2\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_0}[\left\|(\boldsymbol{I} - \boldsymbol{\Sigma}^*)^{-1}\boldsymbol{x}\right\|_2^2 r^2(\boldsymbol{x})] + 2\left\|\Sigma^{*-1}\boldsymbol{\mu}^*\right\|_2^2\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_0}[r^2(\boldsymbol{x})] \\
&\leq 2\sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_0}[\|(\boldsymbol{I} - \boldsymbol{\Sigma}^{*-1})\boldsymbol{x}\|_2^4]\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_0}[r^4(\boldsymbol{x})]} + 2\left\|\Sigma^{*-1}\boldsymbol{\mu}^*\right\|_2^2\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_0}[r^2(\boldsymbol{x})] \leq \mathrm{poly}(1/\alpha),
\end{aligned}$$

where the bound on the expectation of $r^4(\boldsymbol{x})$ and $r^2(\boldsymbol{x})$ follows from Lemma 6.13 and the expectation

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_0}\left[\left\|(\boldsymbol{I}-\boldsymbol{\Sigma}^{*-1})\boldsymbol{x}\right\|_2^4\right] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim N_0}\left[\left(\sum_i (1-\lambda_i)^2 x_i^2\right)^2\right]$$
$$\leq 3\left(\sum_i (1-\lambda_i)^2\right)^2 \leq 3\log^2(1/\alpha) \leq \mathrm{poly}(1/\alpha).$$

Learning the Hermite Expansion

In this section we deal with the sample complexity of estimating the coefficients of the Hermite expansion. We have

$$c_V = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}[H_V(\boldsymbol{x})]$$

Using samples $\boldsymbol{x}_i$ from $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)$, we can estimate this expectation empirically with the unbiased estimate

$$\widetilde{c}_V = \frac{\sum_{i=1}^N H_V(\boldsymbol{x}_i)}{N}.$$

We now show an upper bound for the variance of the above estimate. The proof of this lemma can be found in Appendix E.3.

Lemma 6.17. *Let $\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)$ be the unknown truncated Gaussian. The variance of the following unbiased estimator of the Hermite coefficients $\widetilde{c}_V = \frac{\sum_{i=1}^N H_V(\boldsymbol{x}_i)}{N}$, is upper bounded*

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}[(\widetilde{c}_V - c_V)^2] \leq \mathrm{poly}(1/\alpha)\frac{5^{|V|}}{N}.$$

Theorem 6.18. *Let S be an arbitrary (Borel) subset of $\mathbb{R}^d$. Let α be the constant of* (6.3). *Let $\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)$ be the corresponding truncated Gaussian in $(O\log(1/\alpha), 1/16)$-isotropic* *position (see Definition 6.3), Then, for the estimate*

$$\psi_k(\boldsymbol{x}) = \max\left(0, \sum_{V:0\le|V|\le k} \widetilde{c}_V H_V(\boldsymbol{x})\right), \quad \widetilde{c}_V = \frac{\sum_{i=1}^N H_V(\boldsymbol{x}_i)}{N}$$

it holds for $k \ll d$, $\Gamma(S) > 1$,

$$\mathop{\mathbf{E}}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}\left[\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[(\psi_k(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2\right]\right] \le \mathrm{poly}(1/\alpha)\left(\frac{\sqrt{\Gamma(S)}}{k^{1/4}} + \frac{(5d)^k}{N}\right).$$

Alternatively, for $k = \mathrm{poly}(1/\alpha)\Gamma(S)^2/\epsilon^4$ we obtain that with $N = d^{\mathrm{poly}(1/\alpha)\Gamma(S)^2/\epsilon^4}$ samples, with probability at least 9/10, it holds $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[(\psi_{N,k}(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2] \le \epsilon$.

Proof. Instead of considering the positive part of the Hermite expansion, we will prove the claim for the empirical Hermite expansion of degree k and N samples

$$p_{N,k} = \sum_{V:0\le|V|\le k} \widetilde{c}_V H_V(\boldsymbol{x}).$$

As usual we denote by $S_k\psi(\boldsymbol{x})$ the true (exact) Hermite expansion of degree k of $\psi(\boldsymbol{x})$. Using the inequality $(a-b)^2 \le 2(a-c)^2 + 2(c-b)^2$ we obtain

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[(p_{N,k}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right] \le 2\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[(p_{N,k}(\boldsymbol{x}) - S_k\psi(\boldsymbol{x}))^2\right] + 2\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[(S_k\psi(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2\right]$$

Since Hermite polynomials form an orthonormal system with respect to $\mathcal{N}_0$, we obtain

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[(p_{N,k}(\boldsymbol{x}) - S_k\psi(\boldsymbol{x}))^2\right] &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\left(\sum_{V:0\le|V|\le k}(\widetilde{c}_V - c_V)H_V(\boldsymbol{x})\right)^2\right] \\ &= \sum_{V:0\le|V|\le k}(\widetilde{c}_V - c_V)^2.\end{aligned}$$

Using Lemma 6.17 we obtain

$$\mathop{\mathbf{E}}_{x_1,\ldots,x_N\sim\mathcal{N}^*}\left[\sum_{V:0\le|V|\le k}(\widetilde{c}_V-c_V)^2\right]\le\frac{\mathrm{poly}(1/\alpha)}{N}\sum_{V:0\le|V|\le k}5^{|V|}$$
$$\le\frac{\mathrm{poly}(1/\alpha)}{N}\binom{d+k}{k}5^k,$$

where we used the fact that the number of all multi-indices V of d elements such that $0\le|V|\le k$ is $\binom{d+k}{k}$. Moreover, from Theorem 6.12 we obtain that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}\left[(S_k\psi(\boldsymbol{x})-\psi(\boldsymbol{x}))^2\right]\le\mathrm{poly}(1/\alpha)\left(\frac{\sqrt{\Gamma(S)}}{k^{1/4}}+\frac{1}{k}\right).$$

The theorem follows. □

Optimization of Gaussian Parameters

In this section we show that we can formulate a convex objective function that can be optimized to yield the unknown parameters $\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*$ of the truncated Gaussian. Let S be the unknown (Borel) subset of $\mathbb{R}^d$ such that $\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;S)=\alpha^*$ and let $\mathcal{N}t_S=\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)$ be the corresponding truncated Gaussian.

To find the parameters $\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*$, we define the function

$$M_f(\boldsymbol{u},\boldsymbol{B})\triangleq\mathop{\mathbf{E}}_{x\sim\mathcal{N}t_S}\left[\mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\mathbf{0},\boldsymbol{I};\boldsymbol{x})f(\boldsymbol{x})\right]\tag{6.5}$$

where $h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})=\frac{\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}}{2}-\frac{\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S+\widetilde{\boldsymbol{\mu}}_S\widetilde{\boldsymbol{\mu}}_S^T))}{2}-\boldsymbol{u}^T(\boldsymbol{x}-\widetilde{\boldsymbol{\mu}}_S)+\frac{d}{2}\log 2\pi$.

We will show that the minimizer of $M_f(\boldsymbol{u},\boldsymbol{B})$ for the polynomial function $f=\psi_k$, will satisfy $(\boldsymbol{B}^{-1}\boldsymbol{u},\boldsymbol{B}^{-1})\approx(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)$. Note that $M_f(\boldsymbol{u},\boldsymbol{B})$ can be estimated through samples. Our goal will be to optimize it through stochastic gradient descent.

In order to make sure that SGD algorithm for M_{ψ_k} converges fast in the parameter space we need to project after every iteration to some subset of the space as we will see in more details later in this Section. Assuming that the pair $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ is in $(\sqrt{\log(1/\alpha^*)}, 1/16)$-isotropic position we define the following set

$$\mathcal{D} = \left\{ (\boldsymbol{u}, \boldsymbol{B}) \mid (\boldsymbol{B}^{-1}\boldsymbol{u}, \boldsymbol{B}^{-1}) \quad \text{is in } (c \cdot \log(1/\alpha^*), 1/16)\text{-isotropic position} \right\} \tag{6.6}$$

Where c is the universal constant guaranteed to exist from Section 6.1 so that

$$\max\left\{ \|\boldsymbol{\mu}^* - \tilde{\boldsymbol{\mu}}\|_{\boldsymbol{\Sigma}^*}, \left\|\boldsymbol{\Sigma}^* - \tilde{\boldsymbol{\Sigma}}\right\|_F \right\} \leq c \cdot \log(1/\alpha^*).$$

It is not hard to see that $\mathcal{D}$ is a convex set and that for any $(\boldsymbol{u}, \boldsymbol{B})$ the projection to $\mathcal{D}$ can be done efficiently. For more details we refer to Lemma 8 of Daskalakis et al. (2018). Since after every iteration of our algorithm we project to $\mathcal{D}$ we will assume for the rest of this Section that $(\boldsymbol{u}, \boldsymbol{B}) \in \mathcal{D}$.

An equivalent formulation of $M_f(\boldsymbol{u}, \boldsymbol{B})$ that will be useful for the analysis of the SGD algorithm is

$$\begin{aligned}
&M_f(\boldsymbol{u}, \boldsymbol{B}) && (6.7)\\
&= e^{-\frac{1}{2}\left(\operatorname{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S+\widetilde{\boldsymbol{\mu}}_S\widetilde{\boldsymbol{\mu}}_S^T)))+\boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{u}-\boldsymbol{u}^T\widetilde{\boldsymbol{\mu}}_S\right)} \sqrt{|\boldsymbol{B}|} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}t_S}\left[\frac{\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x})}{\mathcal{N}(\boldsymbol{B}^{-1}\boldsymbol{u}, \boldsymbol{B}^{-1}; \boldsymbol{x})} f(\boldsymbol{x})\right] &&\\
&\triangleq C_{\boldsymbol{u},\boldsymbol{B}} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}t_S}\left[\frac{\mathcal{N}_0(\boldsymbol{x})}{\mathcal{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})} f(\boldsymbol{x})\right] && (6.8)
\end{aligned}$$

Lemma 6.19. *For $(\boldsymbol{u}, \boldsymbol{B}) \in \mathcal{D}$, we have that* $\operatorname{poly}(\alpha) \leq C_{\boldsymbol{u},\boldsymbol{B}} \leq \operatorname{poly}(1/\alpha)$.

Proof. We have that

$$\begin{aligned}
|2\log C_{\boldsymbol{u},\boldsymbol{B}}| &= \left|\operatorname{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S + \widetilde{\boldsymbol{\mu}}_S\widetilde{\boldsymbol{\mu}}_S^T))) + \boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{u} - \boldsymbol{u}^T\widetilde{\boldsymbol{\mu}}_S - \log|\boldsymbol{B}|\right| \\
&= \left|\operatorname{tr}(\boldsymbol{B}-\boldsymbol{I}) + \operatorname{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{I})) + \boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{u} - \log|\boldsymbol{B}|\right| \\
&\leq |\operatorname{tr}(\boldsymbol{B}-\boldsymbol{I}) - \log|\boldsymbol{B}|| + \left|\operatorname{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{I}))\right| + \left|\boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{u}\right|
\end{aligned}$$

We now bound each of the terms separately. Let $\lambda_1, ..., \lambda_d$ be the eigenvalues of $\boldsymbol{B}$.

1. For the first term, we have that

$$|\mathrm{tr}(\boldsymbol{B}-\boldsymbol{I}) - \log|\boldsymbol{B}|| = |\sum_{i=1}^{d}(\lambda_i - 1 - \log\lambda_i)| \le \sum_{i=1}^{d}\frac{(\lambda_i-1)^2}{\lambda_i} \le \frac{\|\boldsymbol{B}-\boldsymbol{I}\|_F^2}{\lambda_{min}}$$

   where we used the fact that $0 \le x - 1 - \log x \le \frac{(x-1)^2}{x}$ for all $x > 0$.

2. For the second term, we have that $\left|\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{I}))\right| \le \|\boldsymbol{B}-\boldsymbol{I}\|_F\|\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{I}\|_F$

3. For the third term, we have that $\left|\boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{u}\right| = \boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{B}^{-1}\boldsymbol{u} \le \lambda_{max}\|\boldsymbol{B}^{-1}\boldsymbol{u}\|_2^2$

Now from the assumption $(\boldsymbol{u}, \boldsymbol{B}) \in \mathcal{D}$ we have that $\|\boldsymbol{B}-\boldsymbol{I}\|_F \le O(\sqrt{\log(1/\alpha^*)})$, $\left\|\boldsymbol{B}^{-1}\boldsymbol{u}\right\|_2 \le O(\sqrt{\log(1/\alpha^*)})$, $\lambda_{min} \ge 15/16$ and $\lambda_{max} \le 17/16$. Also from Lemma E.3 we get that $\left\|\tilde{\boldsymbol{\Sigma}}_S - \boldsymbol{I}\right\|_F \le O(\sqrt{\log(1/\alpha^*)})$ and hence $|2\log C_{\boldsymbol{u},\boldsymbol{B}}| \le O(\log(1/\alpha^*))$. This means that $C_{\boldsymbol{u},\boldsymbol{B}} = \mathrm{poly}(1/\alpha)$ and the lemma follows. □

The Objective Function and its Approximation

To show that the minimizer of the function M_{ψ_k} is a good estimator for the unknown parameters $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$, we consider the function M_f', defined as $M_f(\boldsymbol{u}, \boldsymbol{B}) = \mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}t_S}\left[\mathrm{e}^{h'(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x})f(\boldsymbol{x})\right]$ for $h'(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x}) = \frac{\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}}{2} - \frac{\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\boldsymbol{\Sigma}_S + \boldsymbol{\mu}_S\boldsymbol{\mu}_S^T))}{2} - \boldsymbol{u}^T(\boldsymbol{x} - \boldsymbol{\mu}_S) + \frac{d}{2}\log 2\pi$. This function corresponds to an ideal situation where we know the parameters $\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S$ exactly. Similarly to (6.8), we can write M_f' as $C_{\boldsymbol{u},\boldsymbol{B}}'\,\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}t_S}\left[\frac{\mathcal{N}_0(\boldsymbol{x})}{\mathcal{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}f(\boldsymbol{x})\right]$. We argue that both M_f and M_f' are convex.

Claim 6.20. *For any function $f : \mathbb{R}^d \mapsto \mathbb{R}_{\ge 0}$, $M_f(\boldsymbol{u}, \boldsymbol{B})$ and $M_f'(\boldsymbol{u}, \boldsymbol{B})$ are convex functions of the parameters $(\boldsymbol{u}, \boldsymbol{B})$.*

Proof. We show the statement for M_f. The proof for M_f' is identical. The proof follows by computing the Hessian of M_f and arguing that it is positive semidefinite.

The gradient with respect to $(\boldsymbol{u}, \boldsymbol{B})$ is

$$\nabla M_f(\boldsymbol{u}, \boldsymbol{B}) = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)} \left[\nabla h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x}) \mathrm{e}^{h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x}) f(\boldsymbol{x}) \right]$$
$$= \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)} \left[\begin{pmatrix} \frac{1}{2} \left(\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S \tilde{\boldsymbol{\mu}}_S^T \right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x} \end{pmatrix} \mathrm{e}^{h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x}) f(\boldsymbol{x}) \right] \tag{6.9}$$

Moreover, the Hessian is

$$\mathcal{H}_{M_f}(\boldsymbol{u}, \boldsymbol{B}) = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)} \left[\begin{pmatrix} \frac{1}{2} \left(\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S \tilde{\boldsymbol{\mu}}_S^T \right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x} \end{pmatrix} \begin{pmatrix} \frac{1}{2} \left(\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S \tilde{\boldsymbol{\mu}}_S^T \right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x} \end{pmatrix}^T \mathrm{e}^{h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x}) f(\boldsymbol{x}) \right]$$

which is clearly positive semidefinite since for any $\boldsymbol{z} \in \mathbb{R}^{d \times d + d}$ we have

$$\boldsymbol{z}^T \mathcal{H}_{M_f}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)} \left[\left(\boldsymbol{z}^T \begin{pmatrix} \frac{1}{2} \left(\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S \tilde{\boldsymbol{\mu}}_S^T \right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x} \end{pmatrix} \right)^2 \mathrm{e}^{h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x}) f(\boldsymbol{x}) \right] \geq 0.$$

□

We now argue that the minimizer of the convex function M'_ψ for the weighted characteristic function $\psi(\boldsymbol{x}) = \frac{\mathbb{1}_S(\boldsymbol{x})}{\alpha^*} \frac{\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; \boldsymbol{x})}{\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x})}$ is $(\boldsymbol{u}, \boldsymbol{B}) = (\boldsymbol{\Sigma}^{*-1}, \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*)$.

Claim 6.21. *The minimizer of $M'_\psi(\boldsymbol{u}, \boldsymbol{B})$ is $(\boldsymbol{u}, \boldsymbol{B}) = (\boldsymbol{\Sigma}^{*-1}, \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*)$.*

Proof. The gradient of M'_ψ with respect to $(\boldsymbol{u},\boldsymbol{B})$ is

$$\begin{aligned}
&\nabla M'_\psi(\boldsymbol{u},\boldsymbol{B})\\
&\quad= \mathop{\mathbf{E}}_{x\sim\mathcal{N}t_S}\left[\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{\Sigma}_S-\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\right)^\flat\\ \boldsymbol{\mu}_S-\boldsymbol{x}\end{pmatrix} \mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})\frac{\mathbb{1}_S(\boldsymbol{x})}{\alpha^*}\frac{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}{\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})}\right]\\
&\quad= \mathop{\mathbf{E}}_{x\sim\mathcal{N}t_S}\left[\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{\Sigma}_S-\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\right)^\flat\\ \boldsymbol{\mu}_S-\boldsymbol{x}\end{pmatrix} \mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\frac{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}{\alpha^*}\right]
\end{aligned}$$

For $(\boldsymbol{u},\boldsymbol{B})=(\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^{*-1})$, this is equal to

$$\begin{aligned}
&\nabla M'_\psi(\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^{*-1})\\
&\quad= C_{\boldsymbol{u},\boldsymbol{B}}\cdot\mathop{\mathbf{E}}_{x\sim\mathcal{N}t_S}\left[\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{\Sigma}_S-\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\right)^\flat\\ \boldsymbol{\mu}_S-\boldsymbol{x}\end{pmatrix}\frac{1}{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}\frac{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}{\alpha^*}\right]\\
&\quad= \frac{C_{\boldsymbol{u},\boldsymbol{B}}}{\alpha^*}\cdot\mathop{\mathbf{E}}_{x\sim\mathcal{N}t_S}\left[\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{\Sigma}_S-\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\right)^\flat\\ \boldsymbol{\mu}_S-\boldsymbol{x}\end{pmatrix}\right]
\end{aligned}$$

where $C_{\boldsymbol{u},\boldsymbol{B}}$ that does not depend on x. This is equal to 0 by definition of $\boldsymbol{\mu}_S$ and $\boldsymbol{\Sigma}_S$. □

We want to show that the minimizer of M_{ψ_k} is close to that of M'_ψ. To do this, we bound the difference of the two functions pointwise. The proof of the following lemma is technical and can be found in Appendix E.4.

Lemma 6.22 (Pointwise Approximation of the Objective Function)**.** *Assume that we use Lemma E.2 to estimate $\tilde{\boldsymbol{\mu}}_S,\tilde{\boldsymbol{\Sigma}}_S$ with $\epsilon=\frac{1}{\mathrm{poly}(1/\alpha^*)}\epsilon'$ and Theorem 6.18 with $\epsilon=\frac{1}{p(1/\alpha^*)}\epsilon'^2$ then*

$$\left|M_{\psi_k}(\boldsymbol{u},\boldsymbol{B})-M'_\psi(\boldsymbol{u},\boldsymbol{B})\right|\leq\epsilon'.$$

Now that we have established that M_{ψ_k} is a good approximation of M'_ψ we will prove that we can optimize M_{ψ_k} and get a solution that is very close to the optimal solution of M'_ψ.

Optimization of the Approximate Objective Function

Our goal in this section is to prove that using sample access to $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$ we can find the minimum of the function M_{ψ_k} defined in the previous section. First of all recall that M_{ψ_k} can be written as an expectation over $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$ in the following way

$$M_{\psi_k}(\boldsymbol{u}, \boldsymbol{B}) \triangleq \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N} t_S} \left[\mathrm{e}^{h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) \psi_k(\boldsymbol{x}) \right].$$

In Section 6.3 we prove that we can learn the function ψ_k and hence M_{ψ_k} can be written as

$$M_{\psi_k}(\boldsymbol{u}, \boldsymbol{B}) = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N} t_S} \left[m_{\psi_k}(\boldsymbol{u}, \boldsymbol{B}; x) \right]$$

where $m_{\psi_k}(\boldsymbol{u}, \boldsymbol{B}; x) = \mathrm{e}^{h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) \psi_k(\boldsymbol{x})$, and for any $\boldsymbol{u}, \boldsymbol{B}$ and $\boldsymbol{x}$ we can compute $m_{\psi_k}(\boldsymbol{u}, \boldsymbol{B}; x)$. Since M_{ψ_k} is convex we are going to use stochastic gradient descent to find its minimum. To prove the convergence of SGD and bound the number of steps that SGD needs to converge we will use the the formulation developed in Chapter 14 of Shalev-Shwartz and Ben-David (2014c). To be able to use their results we have to define for any $(\boldsymbol{u}, \boldsymbol{B})$ a random vector $\boldsymbol{v}(\boldsymbol{u}, \boldsymbol{B})$ and prove the following

Unbiased Gradient Estimation

$$\mathbf{E}\left[\boldsymbol{v}(\boldsymbol{u}, \boldsymbol{B})\right] = \nabla M_{\psi_k},$$

Bounded Step Variance

$$\mathbf{E}\left[\left\|\boldsymbol{v}(\boldsymbol{u}, \boldsymbol{B})\right\|_2^2\right] \leq \rho,$$

Strong Convexity for any $\boldsymbol{z} \in \mathcal{D}$ it holds

$$\boldsymbol{z}^T \mathcal{H}_{M_f}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} \geq \lambda.$$

We start with the definition of the random vector $\boldsymbol{v}$. Given a sample $\boldsymbol{x}$ from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$, for any $(\boldsymbol{u}, \boldsymbol{B})$ we define

$$\boldsymbol{v}(\boldsymbol{u}, \boldsymbol{B}) = \nabla_{\boldsymbol{u},\boldsymbol{B}}\, m_{\psi_k}(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x}) \tag{6.10}$$

$$= \begin{pmatrix} \frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S \tilde{\boldsymbol{\mu}}_S^T\right)^{\flat} \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x} \end{pmatrix} \mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x}) \psi_k(\boldsymbol{x}) \tag{6.11}$$

observe that the randomness of $\boldsymbol{v}$ only comes from the random sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$. The fact that $\boldsymbol{v}(\boldsymbol{u}, \boldsymbol{B})$ is an unbiased estimator of $\nabla M_f(\boldsymbol{u}, \boldsymbol{B})$ follows directly from the fact calculation of $\nabla M_f(\boldsymbol{u}, \boldsymbol{B})$ in Section 6.3. For the other two properties that we need we have the following lemmas. The following lemma bounds the variance of the step of the SGD algorithm. It's rather technical proof can be found in Appendix E.4.

Lemma 6.23 (BOUNDED STEP VARIANCE)**.** *Let α be the constant of* (6.3)*. For every $(\boldsymbol{u}, \boldsymbol{B}) \in \mathcal{D}$ it holds*

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N} t_S}\left[\|\boldsymbol{v}(\boldsymbol{u}, \boldsymbol{B})\|_2^2\right] \leq \mathrm{poly}(1/\alpha) \cdot d^{2k},$$

We are now going to prove the strong convexity of the objective function M_{ψ_k}. For this we are going to use a known anti-concentration result (Theorem E.4) for polynomial functions over the Gaussian measure. See Appendix E.1.

The following lemma shows that our objective is strongly convex as long as the guess $\boldsymbol{u}, \boldsymbol{B}$ remains in the set $\mathcal{D}$. Its proof can be found in Appendix E.4.

Lemma 6.24 (STRONG CONVEXITY)**.** *Let α be the absolute constant of* (6.3)*. For every $(\boldsymbol{u}, \boldsymbol{B}) \in \mathcal{D}$, any $\boldsymbol{z} \in \mathbb{R}^d$ such that $\|\boldsymbol{z}\|_2 = 1$ and the first d^2 coordinated of $\boldsymbol{z}$ correspond to a symmetric matrix, then*

$$\boldsymbol{z}^T \mathcal{H}_{M_f}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} \geq \mathrm{poly}(\alpha),$$

Recovering the Unconditional Mean and Covariance

The framework that we use for proving the fast convergence of our SGD algorithm is summarized in the following theorem and the following lemma.

Theorem 6.25 (Theorem 14.11 of Shalev-Shwartz and Ben-David (2014c).)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$. Assume that f is λ-strongly convex, that $\mathbf{E}\left[\boldsymbol{v}^{(i)} \mid \boldsymbol{w}^{(i-1)}\right] \in \partial f(\boldsymbol{w}^{(i-1)})$ and that $\mathbf{E}\left[\left\|\boldsymbol{v}^{(i)}\right\|_2^2\right] \le \rho^2$. Let $\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w}\in\mathcal{D}} f(\boldsymbol{w})$ be an optimal solution. Then,*

$$\mathbf{E}\left[f(\bar{\boldsymbol{w}})\right] - f(\boldsymbol{w}^*) \le \frac{\rho^2}{2\lambda T}\left(1 + \log T\right),$$

where $\bar{\boldsymbol{w}}$ is the output projected stochastic gradient descent with steps $\boldsymbol{v}^{(i)}$ and projection set $\mathcal{D}$ after T iterations.

Lemma 6.26 (Lemma 13.5 of Shalev-Shwartz and Ben-David (2014c).)**.** *If f is λ-strongly convex and $\boldsymbol{w}^*$ is a minimizer of f, then, for any $\boldsymbol{w}$ it holds that*

$$f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \ge \frac{\lambda}{2}\left\|\boldsymbol{w} - \boldsymbol{w}^*\right\|_2^2.$$

Now we have all the ingredients to present the proof of Theorem 6.4.

The proof of Theorem 6.4 The estimation procedure starts by computing the polynomial function ψ_k using $d^{\mathrm{poly}(1/\alpha^*)\frac{\Gamma^2(S)}{\epsilon'^8}}$ samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$ as explained in Theorem 6.18 to get error $\mathrm{poly}(\alpha^*)\epsilon'^2$. Then we compute $\tilde{\boldsymbol{\mu}}_S$ and $\tilde{\boldsymbol{\Sigma}}_S$ as explained in Section 6.1 with $\epsilon = \frac{q(\alpha^*)}{8p(1/\alpha^*)}(\epsilon')^2$ where p comes from Lemma 6.22 and q comes from Lemma 6.24. Our estimators for $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ are the outputs of Algorithm 11.

We analyze the accuracy of our estimation by proving that the minimum of M_{ψ_k} is close in the parameter space to the minimum of M'_{ψ}. Let $\boldsymbol{u}', \boldsymbol{B}'$ be the minimum of the convex function M'_{ψ} and $\boldsymbol{u}_k, \boldsymbol{B}_k$ be the minimum of the convex function M_{ψ_k}. Using Lemma 6.22 we have the following relations

$$\left|M'_{\psi}(\boldsymbol{u}', \boldsymbol{B}') - M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}')\right| \le \epsilon', \qquad \left|M'_{\psi}(\boldsymbol{u}_k, \boldsymbol{B}_k) - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k)\right| \le \epsilon'$$

and also

$$M'_{\psi}(\boldsymbol{u}', \boldsymbol{B}') \leq M'_{\psi}(\boldsymbol{u}_k, \boldsymbol{B}_k), \qquad M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k) \leq M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}').$$

These relations imply that

$$\begin{aligned}
&\left|M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}') - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k)\right| \\
&= M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}') - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k) \\
&\leq M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}') - M'_{\psi}(\boldsymbol{u}', \boldsymbol{B}') + M'_{\psi}(\boldsymbol{u}_k, \boldsymbol{B}_k) - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k) \\
&\leq \left|M'_{\psi}(\boldsymbol{u}', \boldsymbol{B}') - M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}')\right| + \left|M'_{\psi}(\boldsymbol{u}_k, \boldsymbol{B}_k) - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k)\right| \leq 2\epsilon'.
\end{aligned}$$

But from Lemma 6.24 and Lemma 6.26 we get that $\left\| \begin{pmatrix} \boldsymbol{B}'^{\flat} \\ \boldsymbol{u}' \end{pmatrix} - \begin{pmatrix} \boldsymbol{B}_k^{\flat} \\ \boldsymbol{u}_k \end{pmatrix} \right\|_2 \leq \frac{\epsilon'}{2}$. Now we can apply the Claim 6.21 which implies that

$$\left\| \begin{pmatrix} (\boldsymbol{\Sigma}^{*-1})^{\flat} \\ \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^* \end{pmatrix} - \begin{pmatrix} \boldsymbol{B}_k^{\flat} \\ \boldsymbol{u}_k \end{pmatrix} \right\|_2 \leq \frac{\epsilon'}{2}. \tag{6.12}$$

Therefore it suffices to find $(\boldsymbol{u}_k, \boldsymbol{B}_k)$ with accuracy $\epsilon'/2$ to get our theorem.

Let $\boldsymbol{w}^* = \begin{pmatrix} \boldsymbol{B}_k^{\flat} \\ \boldsymbol{u}_k \end{pmatrix}$ To prove that Algorithm 11 converges to $\boldsymbol{w}^*$ we use Theorem 6.25 which together with Markov's inequality, Lemma 6.23 and Lemma 6.24 gives us

$$\mathbf{Pr}\left(M_{\psi_k}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{B}}) - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k) \geq \mathrm{poly}(1/\alpha^*) \cdot \frac{d^{2k}}{T}\left(1 + \log(T)\right) \right) \leq \frac{1}{3}. \tag{6.13}$$

To get our estimation we first repeat the SGD procedure $K = \log(1/\delta)$ times independently, with parameters T, λ each time. We then get the set of estimates $\mathcal{E} = \{\bar{\boldsymbol{w}}_1, \bar{\boldsymbol{w}}_2, \ldots, \bar{\boldsymbol{w}}_K\}$. Because of (6.13) we know that, with high probability $1 - \delta$, for at least the $2/3$ of the points $\bar{\boldsymbol{w}}$ in $\mathcal{E}$ it is true that $M_{\psi_k}(\boldsymbol{w}) - M_{\psi_k}(\boldsymbol{w}^*) \leq \eta$ where $\eta = \mathrm{poly}(1/\alpha^*) \cdot \frac{d^{2k}}{T}\left(1 + \log(T)\right)$. Moreover we will prove later that

$M_{\psi_k}(\boldsymbol{w}) - M_{\psi_k}(\boldsymbol{w}^*) \leq \eta$ and this implies $\|\boldsymbol{w} - \boldsymbol{w}^*\| \leq c \cdot \eta$, where c is a universal constant. Therefore with high probability $1 - \delta$ for at least the $2/3$ of the points $\bar{\boldsymbol{w}}, \bar{\boldsymbol{w}}'$ in $\mathcal{E}$ it is true that $\|\boldsymbol{w} - \boldsymbol{w}'\| \leq 2c \cdot \eta$. Hence if we set $\hat{\boldsymbol{w}}$ to be a point that is at least $2c \cdot \eta$ close to more that the half of the points in $\mathcal{E}$ then with high probability $1 - \delta$ we have that $f(\bar{\boldsymbol{w}}) - f(\boldsymbol{w}^*) \leq \eta$. Hence we can we lose probability at most δ if we condition on the event

$$M_{\psi_k}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{B}}) - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k) \leq \text{poly}(1/\alpha^*) \cdot \frac{d^{2k}}{T} \left(1 + \log(T)\right).$$

Using once again Lemma 6.26 we get that

$$\left\| \begin{pmatrix} \hat{\boldsymbol{B}}^\flat \\ \hat{\boldsymbol{u}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{B}_k^\flat \\ \boldsymbol{u}_k \end{pmatrix} \right\|_2 \leq \frac{\epsilon'}{2}.$$

which together with (6.12) implies

$$\left\| \begin{pmatrix} \hat{\boldsymbol{B}}^\flat \\ \hat{\boldsymbol{u}} \end{pmatrix} - \begin{pmatrix} (\boldsymbol{\Sigma}^{*-1})^\flat \\ \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^* \end{pmatrix} \right\|_2 \leq \frac{\epsilon'}{2}.$$

and the theorem follows as closeness in parameter distance implies closeness in total variation distance for the corresponding untruncated Gaussian distributions.

Algorithm 11 Projected Stochastic Gradient Descent. Given access to samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$.

```
procedure SGD($T, \lambda$)                                  ▷ $T$: number of steps, $\lambda$: parameter.
    $\boldsymbol{w}^{(0)} = \begin{pmatrix} (\boldsymbol{B}^{(0)})^\flat \\ \boldsymbol{u}^{(0)} \end{pmatrix} \leftarrow \begin{pmatrix} (\tilde{\boldsymbol{\Sigma}}_S^{-1})^\flat \\ \tilde{\boldsymbol{\Sigma}}_S^{-1} \tilde{\boldsymbol{\mu}}_S \end{pmatrix}$
    for $i = 1, \ldots, T$ do
        Sample $\boldsymbol{x}^{(i)}$ from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$
        $\eta_i \leftarrow \frac{1}{\lambda \cdot i}$
        $\begin{pmatrix} (\boldsymbol{B}^{(i-1)})^\flat \\ \boldsymbol{u}^{(i-1)} \end{pmatrix} \leftarrow \boldsymbol{w}^{(i-1)}$
        $\boldsymbol{v}^{(i)} \leftarrow \begin{pmatrix} \frac{1}{2} \left( \boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)T} - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S \tilde{\boldsymbol{\mu}}_S^T \right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}^{(i)} \end{pmatrix} \mathrm{e}^{h(\boldsymbol{u}^{(i-1)}, \boldsymbol{B}^{(i-1)}; \boldsymbol{x}^{(i)})} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x}^{(i)}) \psi_k \left( \boldsymbol{x}^{(i)} \right)$
    ▷ From (6.9).
        $\boldsymbol{r}^{(i)} \leftarrow \boldsymbol{w}^{(i-1)} - \eta_i \boldsymbol{v}^{(i)}$
        $\boldsymbol{w}^{(i)} \leftarrow \arg\min_{\boldsymbol{w} \in \mathcal{D}} \left\| \boldsymbol{w} - \boldsymbol{r}^{(i)} \right\|_2^2$          ▷ From Lemma 8 of Daskalakis et al.
    (2018).
    $\begin{pmatrix} \hat{\boldsymbol{B}}^\flat \\ \hat{\boldsymbol{u}} \end{pmatrix} \leftarrow \frac{1}{T} \sum_{i=1}^T \boldsymbol{w}^{(i)}$
    $\hat{\boldsymbol{\Sigma}} \leftarrow \hat{\boldsymbol{B}}^{-1}$
    $\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{B}}^{-1} \hat{\boldsymbol{u}}$
    return $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$
```

Recovering the Set

In this section we prove that, given only positive examples from an unknown truncated Gaussian distribution, that is samples from the conditional distribution on the truncation set, one can in fact learn the truncation set. We only give here the main result, for details see Appendix E.5.

Theorem 6.27 (Recovering the Set)**.** *Let $\mathcal{S}$ be a class of measurable sets with Gaussian surface area at most $\Gamma(\mathcal{S})$. Let $\mathcal{N}^*$ be a Gaussian in $(O(\log(1/\alpha), 1/16))$-isotropic position. Then, given $d^{\mathrm{poly}(1/\alpha)\Gamma(\mathcal{S})^2/\epsilon^{32}}$ samples from the conditional distribution*

$\mathcal{N}_S^*$ we can recover an indicator of the set $\widetilde{S}$ such that with probability at least 99% it holds $\mathbf{Pr}_{x \sim \mathcal{N}^*}[\widetilde{S}(x) \neq \mathbb{1}_S(x)] \leq \epsilon$.

6.4 Lower Bound for Learning the Mean of a Truncated Normal

In this section we prove our information-theoretic lower bound, Theorem 6.6, which we restate below for conveniece.

Theorem 6.28. *There exists a family of sets $\mathcal{S}$ with $\Gamma(\mathcal{S}) = O(d)$ such that any algorithm that draws m samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}, S)$ and computes an estimate $\widetilde{\boldsymbol{\mu}}$ with $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq 1$ must have $m = \Omega(2^{d/2})$.*

Proof. Let $H = [-1, 1]^{d+1}$ be the $d + 1$-dimensional cube. We will also use the left and right subcubes $H_+ = [-1, 0] \times [-1, 1]^d$, $H_- = [0, 1] \times [-1, 1]^d$ respectively. Let $\mathcal{N}_+ = \mathcal{N}(\boldsymbol{e}_1, \boldsymbol{I})$ and $\mathcal{N}_- = \mathcal{N}(-e_1, \boldsymbol{I})$. We denote by r the (scaled) pointwise minimum of the two densities truncated at the cube H, that is

$$r(x) = \frac{\min(\mathcal{N}_+(H; \boldsymbol{x}), \mathcal{N}_-(H; \boldsymbol{x}))}{c} = \frac{\mathbb{1}_H(\boldsymbol{x})}{c} \min(\mathcal{N}_+(\boldsymbol{x}), \mathcal{N}_-(\boldsymbol{x})),$$

where $c = 1 - d_{\mathrm{TV}}(\mathcal{N}_+, \mathcal{N}_-)$.

To simplify notation we assume that we work in $\mathbb{R}^{d+1}$ instead of $\mathbb{R}^d$. Let $V = (v_1, \ldots, v_d) \in \{+1, -1\}^d$. For every V we define the set $G_V = H \cap \{\boldsymbol{y} \in \mathbb{R}^d : y_i v_i \geq 0\}$. We also define the subcubes $H_V = [0, 1] \times G_V$. We consider the following subset of H parameterized by the 2^d parameters $t_V \in [0, 1]$ and $\delta \in [-1, 1]$.

$$S_+ = [-1 + \delta, 0] \times [-1, 1]^d \cup \bigcup_{V \in \{-1, +1\}^d} [0, t_V] \times G_V$$

We will argue that there exists a distribution D^+ on the values t_V such that on expectation $d_{\mathrm{TV}}(\mathcal{N}_+^{S_+}, \mathcal{N}_-^{S_-})$ is $O(2^{-d})$. We show how to construct the distribution D_+ since the construction for D_- is the same. In fact we will show that both distributions are very close to $r(x)$. Notice that for some $(t, \boldsymbol{y}) \in \mathbb{R}^{d+1}$ we have We draw each t_V independently from the distribution with cdf

$$F(t) = \mathbb{1}_{[0,1)}(t)(1 - \mathrm{e}^{-2t}) + \mathbb{1}_{[1,+\infty)}(t)$$

Notice that for $t \in (0,1)$ and any $\boldsymbol{y} \in \mathbb{R}^d$ we have that $1 - F(t) = \mathcal{N}_-(t, \boldsymbol{y}) / \mathcal{N}_+(t, \boldsymbol{y})$.

After we draw all t_V from F we choose δ so that $\mathcal{N}_+(S_+; x) = c$. We will show that on expectation over the t_V we have $\delta = 0$, which means that no correction is needed. In fact we show something stronger, namely that for all $x \in H_+$ we have that $\mathbf{E}_{S_+ \sim D_+}[N_+(S_+; \boldsymbol{x})] = r(\boldsymbol{x})$. Assume that $x \in H_V$. Indeed,

$$\begin{aligned}\mathop{\mathbf{E}}_{S_+ \sim D_+}[\mathcal{N}_+(S_+; \boldsymbol{x})] &= \frac{\mathcal{N}_+(\boldsymbol{x})}{c} \mathop{\mathbf{E}}_{S_+ \sim D_+}[\mathbb{1}_{S_+}(\boldsymbol{x})] = \frac{\mathcal{N}_+(\boldsymbol{x})}{c} \mathop{\mathbf{E}}_{S_+ \sim D_+}[\mathbb{1}_{\{x_1 \le t_V\}}] \\ &= \frac{\mathcal{N}_+(\boldsymbol{x})}{c}(1 - F(t_V)) = \frac{\mathcal{N}_-(\boldsymbol{x})}{c} = r(\boldsymbol{x})\end{aligned}$$

Moreover, observe that for all $\boldsymbol{x} \in H_- \cap S_+$ we have that $N_+(S_+; \boldsymbol{x}) = r(\boldsymbol{x})$ always (with probability 1). We now argue that in order to have constant probability to distinguish $N_+(S_+)$ from $r(x)$ one needs to draw $\Omega(2^d)$ samples. Since the expected density of $N_+(S_+)$ matches $r(x)$ for all $x \in H_+$, to be able to distinguish the two distributions one needs to observe at least two samples in the same cube H_V. Since we have 2^d disjoint cubes H_V the probability of a sample landing in H_V is at most $1/2^d$. Therefore, using the birthday problem, to have constant probability to observe a collision one needs to draw $\Omega(\sqrt{2^d}) = \Omega(2^{d/2})$ samples. Since for all $x \in H_- \cap S_+$, $N_+(S_+)$ exactly matches $r(x)$, to distinguish between the two distributions one needs to observe a sample $\boldsymbol{x}$ with $-1 + \delta < x_1 < -1$. Due to symmetry, N_+ assigns to all cubes H_V equal probability, call that p. Moreover, we have that $c = 2^{d+1} p$. Now let p_V be the random variable corresponding to the probability that N_+ assigns to $[0, t_V] \times G_V$. We have that $\mathbf{E}_{t_V \sim F}[p_V] = p$ for all V. Since the independent random variables p_V are bounded in $[0, 1/2^d]$, Hoeffding's inequality implies that $|\sum_{V \in \{-1,1\}^d}(p_V - p)| < 1/2^{d/2}$ with probability at least $1 - 2/\mathrm{e}^2$. This means that with probability at least $3/4$ one will need to draw

$\Omega(2^{d/2})$ samples in order to observe one with $x_1 < -1 + \delta$.

Since any set S in our family $\mathcal{S}$ has almost everywhere (that is except the set of its vertices which a finite set and thus of measure zero) smooth boundary we may use the following equivalent (see e.g. Nazarov (2003a)) definition of its surface area

$$\Gamma(S) = \int_{\partial S} \mathcal{N}_0(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}),$$

where $\mathrm{d}\sigma(x)$ is the standard surface measure on $\mathbb{R}^d$. Without loss of generality we assume that S corresponds to the set S_+ defined above (the proof is the same if we consider a set S_-). We have

$$\partial S \subseteq \bigcup_{V \in \{+1,-1\}^d} (\{t_V\} \times G_V) \cup \partial([-1,+1]^{d+1}) \cup \bigcup_{i=1}^{d+1} \{\boldsymbol{x} : x_i = 0\}.$$

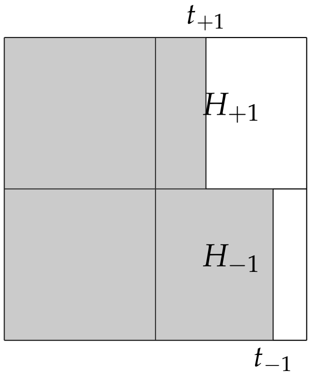
By the definition of Gaussian surface area it is clear that $\Gamma(A \cup B) \leq \Gamma(A) + \Gamma(B)$. From Table 1.1 we know that $\Gamma([-1,+1]^{d+1}) = O(\sqrt{\log d})$. Moreover, we know that a single halfspace has surface area at most $\sqrt{2/\pi}$ (see e.g. Klivans et al. (2008)). Therefore $\Gamma\left(\bigcup_{i=1}^{d+1} \{x : x_i = 0\}\right) \leq \sum_{i=1}^{d+1} \sqrt{2/\pi} = O(d)$. Finally, we notice that for any point x on the hyperplane $\{\boldsymbol{x} : x_1 = 0\}$ and any $\boldsymbol{y}$ on $\{\boldsymbol{x} : x_1 = c\}$ (for any $c \geq 0$), we have $\mathcal{N}_0(\boldsymbol{x}) \geq \mathcal{N}_0(\boldsymbol{y})$. Therefore, the surface area of each set $t_V \times G_V$ is maximized for $t_V = 0$. In this case $\bigcup_{V \in \{+1,-1\}^d} (\{t_V\} \times G_V) \subseteq \{\boldsymbol{x} : x_1 = 0\}$, which implies that the set $\bigcup_{V \in \{+1,-1\}^d} (\{t_V\} \times G_V)$ contributes at most $\sqrt{2/\pi}$ to the total surface area. Putting everything together, we have that $\Gamma(S) = O(d)$.

□

6.5 Identifiability with bounded Gaussian Surface Area

In this section we investigate the sample complexity of the problem of estimating the parameters of a truncated Gaussian using a different approach that does not

Figure 6.4: The set S_+ when $d = 1$.



depend on the VC dimension of the family $\mathcal{S}$ of the truncation sets to be finite. For example, we settle the sample complexity of learning the parameters of a Gaussian distribution truncated at an unknown convex set (recall that the class of convex sets has infinite VC dimension). Our method relies on finding a tuple $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$ of parameters so that the moments of the corresponding truncated Gaussian $\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$ are all close to the moments of the unknown truncated Gaussian distribution, for which we have unbiased estimates using samples. The main question that we need to answer to determine the sample complexity of this problem is how many moments are needed to be matched in order to be sure that our guessed parameters are close to the parameters of the unknown truncated Gaussian. We state now the main result. Its proof is based on Lemma 6.31 and can be found in Appendix E.6. We show Theorem 6.5, which we restate here for convenience.

Theorem 6.29 (Moment Matching)**.** *Let $\mathcal{S}$ be a family of subsets of $\mathbb{R}^d$ of bounded Gaussian surface area $\Gamma(\mathcal{S})$. Moreover, assume that if T is an affine map and $T(\mathcal{S}) = \{T(S) : S \in \mathcal{S}\}$ is the family of the images of the sets of $\mathcal{S}$, then it holds $\Gamma(T(\mathcal{S})) = O(\Gamma(\mathcal{S}))$. For some $S \in \mathcal{S}$, let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ be an unknown truncated Gaussian. $d^{O(\Gamma(\mathcal{S})/\epsilon^4)}$ samples are sufficient to find parameters $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S}$ such that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S), \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})) \leq \epsilon$.*

The key lemma of this section is Lemma 6.31. It shows that if two truncated normals are in total variation distance ϵ then there exists a moment where they differ. The main idea is to prove that there exists a polynomial that approximates well the indicator of the set $\{f_1 > f_2\}$. Notice that the total variation distance between two densities can be written as $\int \mathbb{1}_{\{f_1 > f_2\}}(\boldsymbol{x}) f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) \mathrm{d}x$. In our proof we use the chi squared divergence, which for two distributions with densities f_1, f_2 is defined as

$$D_{\chi^2}(f_1 \| f_2) = \int \frac{(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))^2}{f_2(\boldsymbol{x})} \mathrm{d}x$$

To prove it we need the following nice fact about chi squared divergence between Gaussian distributions. In general chi squared divergence may be infinite for some pairs of Gaussians. In the following lemma we prove that for any pair of Gaussians, there exists another Gaussian N such that $D_{\chi^2}(N_1 \| N)$ $D_{\chi^2}(N_2 \| N)$ are finite even if $D_{\chi^2}(N_1 \| N_2) = \infty$.

Lemma 6.30. *Let $N_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, and $N_2 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_2)$ be two Normal distributions that satisfy the conditions of Lemma E.3. Then there exists a Normal distribution N such that*

$$D_{\chi^2}(N_1 \| N), D_{\chi^2}(N_2 \| N) \leq \exp\Bigg(2 \left\| \boldsymbol{\Sigma}_1^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \right\|_2 + \frac{1}{2} \max(1, \|\boldsymbol{\Sigma}_1\|_2) \left\| \boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{-1/2} - \boldsymbol{I} \right\|_F^2 \Bigg)$$

Now we state the main lemma of this section. We give here a sketch of its proof. It's full version can be found in Appendix E.6.

Lemma 6.31. *Let $\mathcal{S}$ be a family of subsets of $\mathbb{R}^d$ of bounded Gaussian surface area $\Gamma(\mathcal{S})$. Moreover, assume that if T is an affine map and $T(\mathcal{S}) = \{T(S) : S \in \mathcal{S}\}$ is the family of the images of the sets of $\mathcal{S}$, then it holds $\Gamma(T(\mathcal{S})) = O(\Gamma(\mathcal{S}))$. Let $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, S_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, S_2)$ be two truncated Gaussians with densities f_1, f_2 respectively. Let $k = O(\Gamma(\mathcal{S})/\epsilon^4)$. If $d_{\mathrm{TV}}(f_1, f_2) \geq \epsilon$, then there exists a $V \in \mathbb{N}^d$ with $|V| \leq k$ such that*

$$\left| \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, S_1)} [\boldsymbol{x}^V] - \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, S_2)} [\boldsymbol{x}^V] \right| \geq \epsilon / d^{O(k)}.$$

Proof sketch. Let $W = S_1 \cap S_2 \cap \{f_1 > f_2\} \cup S_1 \setminus S_2$, that is the set of points where the first density is larger than the second. We now write the L_1 distance between f_1, f_2 as

$$\int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} = \int \mathbb{1}_W(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$$

Denote $p(\boldsymbol{x})$ the polynomial that will do the approximation of the L_1 distance. From Lemma 6.30 we know that there exists a Normal distribution within small chi-squared divergence of both $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Call the density function of this distribution $g(\boldsymbol{x})$. We have

$$\begin{aligned} &\Big| \int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} - \int p(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})) \Big| \qquad (6.14) \\ &\quad \le \int |\mathbb{1}_W(\boldsymbol{x}) - p(\boldsymbol{x})| \, |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \\ &\quad \le \int |\mathbb{1}_W(\boldsymbol{x}) - p(\boldsymbol{x})| \sqrt{g(\boldsymbol{x})} \, \frac{|f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|}{\sqrt{g(\boldsymbol{x})}} \mathrm{d}\boldsymbol{x} \\ &\quad \le \sqrt{\int (\mathbb{1}_W(\boldsymbol{x}) - p(\boldsymbol{x}))^2 g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}} \sqrt{\int \frac{(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))^2}{g(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}}, \qquad (6.15) \end{aligned}$$

where we use Schwarzs' inequality. From Lemma 6.30 we know that

$$\int \frac{f_1(\boldsymbol{x})^2}{g(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \le \int \frac{\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1; \boldsymbol{x})^2}{g(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} = \exp(\mathrm{poly}(1/\alpha)).$$

Similarly, $\int \frac{f_2(\boldsymbol{x})^2}{g(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} = \exp(\mathrm{poly}(1/\alpha))$. Therefore we have,

$$\begin{aligned} \Big| \int |f_1(\boldsymbol{x}) - & \quad f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} - \int p(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})) \Big| \\ \le \exp(\mathrm{poly}(1/\alpha)) & \sqrt{\int (\mathbb{1}_W(\boldsymbol{x}) - p(\boldsymbol{x}))^2 g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}} \end{aligned}$$

Recall that $g(\boldsymbol{x})$ is the density function of a Gaussian distribution, and let $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ be the parameters of this Gaussian. Notice that it remains to show that there exists a good approximating polynomial $p(\boldsymbol{x})$ to the indicator function $\mathbb{1}_W$. We can now transform the space so that $g(\boldsymbol{x})$ becomes the standard normal. Notice that this is an affine transformation that also transforms the set W; Since the Gaussian surface area is "invariant" under linear transformations

Since $\mathbb{1}_W \in L^2(\mathbb{R}^d, \mathcal{N}_0)$ we can approximate it using Hermite polynomials. For some $k \in \mathcal{N}$ we set $p(\boldsymbol{x}) = S_k \mathbb{1}_W(x)$, that is

$$p_k(\boldsymbol{x}) = \sum_{V:|V| \leq k} \widehat{\mathbb{1}_W}_V H_V(\boldsymbol{x}).$$

Combining Lemma 6.14 and Lemma E.9 we obtain

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_0} [(\mathbb{1}_W(\boldsymbol{x}) - p_k(x))^2] = O\left(\frac{\Gamma(\mathcal{S})}{k^{1/2}}\right).$$

Therefore, $\left| \int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} - \int p_k(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})) \right| = \exp(\mathrm{poly}(1/\alpha)) \frac{\Gamma(\mathcal{S})^{1/2}}{k^{1/4}}$. Ignoring the dependence on the absolute constant α, to achieve error $O(\epsilon)$ we need degree $k = O(\Gamma(\mathcal{S})^2/\epsilon^4)$.

To complete the proof, it remains to obtain a bound for the coefficients of the polynomial $q(\boldsymbol{x}) = p_k(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}))$. Using known facts about the coefficients of Hermite polynomials we obtain that $\|q(\boldsymbol{x})\|_\infty \leq \binom{d+k}{k}^2 (4d)^{k/2} (O(1/\alpha^2))^k$. To conclude the proof we notice that we can pick the degree k so that

$$\left| \int q(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})) \right| = \left| \sum_{V:|V| \leq k} \boldsymbol{x}^V (f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})) \right| \geq \epsilon/2.$$

Since the maximum coefficient of $q(\boldsymbol{x})$ is bounded by $d^{O(k)}$ we obtain the result. □

6.6 VC-dimension vs Gaussian Surface Area

We use two different complexity measures of the truncation set to get sample complexity bounds, the VC-dimension and the Gaussian Surface Area (GSA) of the class of the sets. As we already mentioned in the introduction, there are classes, for example convex sets, that have bounded Gaussian surface area but infinite VC-dimension. However, this is not the main difference between the two complexity measures in our setting. Having a class with bounded VC-dimension means that the empirical risk minimization needs finite samples. To get an efficient algorithm we still need to *implement the ERM for this specific class*. Therefore, it is not clear whether it is possible to get an algorithm that works for all sets of bounded VC-dimension. On the other hand, bounded GSA means that we can approximate the weighted indicator of the set using its low order Hermite coeffients. This approximation works for all sets of bounded GSA and does not depend on the specific class of sets. Therefore, using GSA we manage to get a unified approach that learns the parameters of the underlying Gaussian distribution using only the assumption that the truncation set has bounded GSA. In other words, our approach uses the information of the class that the truncation set belongs only to decide how large the degree of the approximating polynomial should be. Having said that, it is an interesting open problem to design algorithms that learn the parameters of the Gaussian and use the information that the truncation set belongs to some class (e.g. intersection of k-halfspaces) to beat the runtime of our generic approach that only depends on the GSA of the class.

6.7 Further Related Work

Our work is related to the field of robust statistics as it can robustly learn a Gaussian even in the presence of an adversary erasing samples outside a certain set. Recently, there has been a lot of theoretical work doing robust estimation of the parameters of multi-variate Gaussian distributions in the presence of arbitrary corruptions to a small ε fraction of the samples, allowing for both deletions of samples and additions of samples that can also be chosen adaptively Diakonikolas et al. (2016b); Charikar et al. (2017); Lai et al. (2016b); Diakonikolas et al. (2017a, 2018c). When the corruption of the data is so powerful it is easy to see that the estimation error of the parameter depends on ϵ and cannot shrink to 0 as the number of samples grows to infinity. In our model the corruption is more restrictive but in return our results show how to estimate the parameters of a multi-variate Gaussian distribution *to arbitrary* accuracy even when the fraction of corruption is any constant less than 1.

Our work also has connections with the literature of learning from positive examples. At the heart of virtually all of the results in this literature is the use of the exact knowledge of the original non-truncated distribution to be able to generate fake negative examples, e.g. Denis (1998); Letouzey et al. (2000). When the original distribution is uniform, better algorithms are known. Diakonikolas et al. De et al. (2014) gave efficient learning algorithms for DNFs and linear threshold functions, Frieze et al. Frieze et al. (1996) and Anderson et al. Anderson et al. (2013) gave efficient learning algorithms for learning d-dimensional simplices. Another line of work proves lower bounds on the sample complexity of recovering an unknown set from positive examples. Goyal et al. Goyal and Rademacher (2009) showed that learning a convex set in d-dimensions to accuracy ϵ from positive samples, uniformly distributed inside the set, requires at least $2^{\Omega(\sqrt{d/\epsilon})}$ samples, while the work of Eldan (2011) showed that $2^{\Omega(\sqrt{d})}$ samples are necessary even to estimate the mass of the set. To the best of our knowledge, no matching upper bounds are known for those results. Our estimation result implies that $d^{\mathrm{poly}(\frac{1}{\epsilon})\sqrt{d}}$ are sufficient to learn the set and its mass when given positive samples from a Gaussian truncated on the convex set.

7 NON-PARAMETRIC TRUNCATED STATISTICS: LEARNING SMOOTH-DENSITIES

7.1 Formal Statement of Results

Definitions and Preliminaries

Notation. Let $K \subseteq \mathbb{R}^d$ and $B \in \mathbb{R}_+$, we define $L_\infty(K, B)$ to be the set of all functions $f : K \to \mathbb{R}$ such that $\max_{\boldsymbol{x} \in K} |f(\boldsymbol{x})| \leq B$. We may use $L_\infty(B)$ instead of $L_\infty([0,1]^d, B)$. We also define $\operatorname{diam}_p(K) = \sup_{\boldsymbol{x},\boldsymbol{y} \in K} \|\boldsymbol{x} - \boldsymbol{y}\|_p$ where $\|\cdot\|_p$ is the usual ℓ_p-norm of vectors. Let $\mathbb{R}^{d \times \cdots (k \text{ times}) \cdots \times d}$ be the set of k-order tensors of dimension d, which for simplicity we will denote by $\mathbb{R}^{d^k}$. For $\boldsymbol{\alpha} \in \mathbb{N}^d$, we define the factorial of the multi-index $\boldsymbol{\alpha}$ to be $\boldsymbol{\alpha}! = (\alpha_1!) \cdots (\alpha_d!)$. Additionally for any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d$ we define $\boldsymbol{z}^{\boldsymbol{\alpha}} = z_1^{\alpha_1} \cdots z_d^{\alpha_d}$ and in particular $(\boldsymbol{x} - \boldsymbol{y})^{\boldsymbol{\alpha}} = (x_1 - y_1)^{\alpha_1} \cdots (x_d - y_d)^{\alpha_d}$.

Remark 7.1. *Throughout the paper, for simplicity of exposition, we will consider the support K of the densities that we aim to learn to be the hypercube $[0,1]^d$. Our results hold for arbitrary convex sets with the following property $[a,b]^d \subseteq K \subseteq [c,d]^d$. Then all our results will be modified by multiplying with a function of $R \triangleq \frac{d-c}{b-a}$. We will add a note in our theorems to specify the function of R in every case, but we will keep our main statements for $K = [0,1]^d$ to simplify our statements. The set K should not be confused with the survival set S from which we see the samples, which can be an arbitrary measurable subset of K.*

Multivariate Polynomials

When we use a polynomial to define a probability distribution, as we will see in Section 7.1, the constant term of the polynomial has to be determined from the rest of the coefficients so that the resulting function integrates to 1. For this reason we can ignore the constant term. A polynomial q of d variables, degree k, and zero constant term is a function $q : \mathbb{R}^d \to \mathbb{R}$ of the form

$$q(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d, 0 < |\boldsymbol{\alpha}| \le k} v_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}} \tag{7.1}$$

where $v_{\boldsymbol{\alpha}} \in \mathbb{R}$. The monomials of degree $\le k$ can be indexed by a multi-index $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \le k$ and any polynomial belongs to the vector space defined by the monomials as per (7.1).

To associate the space of polynomials with a Euclidean space we can use any ordering of monomials, for example the lexicographic ordering. Using this ordering we can define the *monomial profile of degree k*, $\boldsymbol{m}_k : \mathbb{R}^d \to \mathbb{R}^{t_k - 1}$, as $(\boldsymbol{m}_k(\boldsymbol{x}))_{\boldsymbol{\alpha}} = \boldsymbol{x}^{\boldsymbol{\alpha}}$ where $t_k = \binom{d+k}{k}$ is equal to the number of monomials with d variables and degree at most k and where we abuse notation to index a coordinate in $\mathbb{R}^{t_k - 1}$ via a multi-index $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \le k$ and $\boldsymbol{\alpha} \neq \boldsymbol{0}$; this can be formally done using the lexicographic ordering. Therefore the vector space of polynomials is homomorphic to the vector space $\mathbb{R}^{t_k - 1}$ via the following correspondence

$$\boldsymbol{v} \in \mathbb{R}^{t_k - 1} \leftrightarrow \boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x}) \triangleq q_{\boldsymbol{v}}(\boldsymbol{x}). \tag{7.2}$$

We denote by $\mathcal{Q}_{d,k}$ the space of polynomials of degree at most k with d variables and zero constant term, where we might drop d from the notation if it is clear from context.

High-order Derivatives and Taylor's Theorem

In this section we will define the basic concepts about high order derivatives of a multivariate real-valued function $f : K \to \mathbb{R}$, where $K \subseteq \mathbb{R}^d$.

Fix $k \in \mathbb{N}$ and let $\boldsymbol{u} \in [d]^k$. We define the order k derivative of f with index $\boldsymbol{u} = (u_1, \ldots, u_k)$ as $\mathbf{D}^k_{\boldsymbol{u}} f(\boldsymbol{x}) = \frac{\partial^k f}{\partial x_{u_1} \cdots \partial x_{u_k}}(\boldsymbol{x})$, observe that $\mathbf{D}^k_{\boldsymbol{u}} f$ is a function from S to $\mathbb{R}$. The order k derivative of f at $\boldsymbol{x} \in S$ is then the tensor $\mathbf{D}^k f(\boldsymbol{x}) \in \mathbb{R}^{d^k}$ where the entry of $\mathbf{D}^k f(\boldsymbol{x})$ that corresponds to the index $\boldsymbol{u} \in [d]^k$ is $(\mathbf{D}^k f(\boldsymbol{x}))_{\boldsymbol{u}} = \mathbf{D}^k_{\boldsymbol{u}} f(\boldsymbol{x})$. Because of the symmetry of the partial derivatives the k-th order derivatives can be indexed with a multi-index $\boldsymbol{\alpha} \triangleq (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$, with $|\boldsymbol{\alpha}| = \sum_{i=1}^{d} \alpha_i = k$, as follows $\mathbf{D}_{\boldsymbol{\alpha}} f(\boldsymbol{x}) = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d}(\boldsymbol{x})$. Observe that the k-order derivative of f is a function $\mathbf{D}^k f : K \to \mathbb{R}^{d^k}$.

Norm of High-order Derivative. There are several ways to define the norm of the tensor and hence the norm of a k-order derivative of a multi-variate function. Here we will define only the norm that we will use in the rest of the paper as follows

$$\left\|\mathbf{D}^k f\right\|_\infty \triangleq \sup_{\boldsymbol{x} \in K} \max_{\boldsymbol{u} \in [d]^k} \left|\mathbf{D}^k_{\boldsymbol{u}} f(\boldsymbol{x})\right| = \sup_{\boldsymbol{x} \in K} \max_{\boldsymbol{u} \in [d]^k} \left|\frac{\partial^k f}{\partial x_{u_1} \cdots \partial x_{u_k}}(\boldsymbol{x})\right|. \tag{7.3}$$

Observe that this definition depends on the set K, but for ease of notation we eliminate K from the notation of the norm and we make sure that this set will be clear from the context. For most part of the paper K is the box $[0,1]^d$.

In order to present the main application of Taylor's Theorem that we are using in the rest of the paper we need the definition of a k-order Taylor approximation of a multi-variate function.

Definition 7.2. *(Taylor Approximation) Let $f : K \to \mathbb{R}$ be a $(k+1)$-times differentiable function on the convex set $K \subseteq \mathbb{R}^d$. Then we define $\bar{f}_k(\cdot; \boldsymbol{x})$ to be the k-order Taylor approximation of f around $\boldsymbol{x}$ as follows*

$$\bar{f}_k(\boldsymbol{y}; \boldsymbol{x}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d, |\boldsymbol{\alpha}| \le k} \frac{\mathbf{D}_{\boldsymbol{\alpha}} f(\boldsymbol{x})}{\boldsymbol{\alpha}!} (\boldsymbol{y} - \boldsymbol{x})^{\boldsymbol{\alpha}}.$$

We are now ready to state the main application of Taylor's Theorem. For a proof of this theorem together with a full statement of the multi-dimensional Taylor's Theorem we refer to the Appendix F.1.

Theorem 7.3 (Corollary of Taylor's Theorem)**.** *Let $K \subseteq \mathbb{R}^d$ and $f : K \to \mathbb{R}$ be a $(k+1)$-times differentiable function such that* $\mathrm{diam}_\infty(K) \le R$ *and* $\left\|\mathbf{D}^{k+1} f\right\|_\infty \le W$*,* *then for any $\boldsymbol{x}, \boldsymbol{y} \in K$ it holds that*

$$\left|f(\boldsymbol{y}) - \bar{f}_k(\boldsymbol{y}; \boldsymbol{x})\right| \leq \left(\frac{15d}{k}\right)^{k+1} \cdot R^{k+1} \cdot W.$$

Bounded and High-order Smooth Function

In this section we define the set of functions that our statistical Taylor theorem applies. It is also the domain of function with respect to which we are solving the non-parametric truncated density estimation problem. This set of functions is very similar to the functions consider for interpolation of probability densities from exponential families Barron and Sheu (1991). We note that in this paper our goal is to solve a much more difficult problem since our goal is to do extrapolation instead of interpolation. We call the set of function that we consider *bounded and high-order smooth functions*.

Definition 7.4 (*Bounded and High-order Smooth Functions*)**.** *Let $K = [0, 1]^d$, we define the set $\mathcal{L}(B, M)$ of functions $f : K \to \mathbb{R}$ for which the following conditions are satisfied.*

- (Bounded Value) *It holds that* $\max_{\boldsymbol{x} \in K} |f(\boldsymbol{x})| \leq B$.
- (High-Order Smoothness) *For any natural number k with $k \geq k_0$, f is k-times continuously differentiable and it holds that $\left\|\mathbf{D}^k f\right\|_\infty \leq M^k$.*

We note that the definition of the class $\mathcal{L}$ depends on k_0 as well but for ease of notation we don't keep track of this dependence and we treat k_0 as a constant throughout the paper.

The above definition can be extended to convex sets K but for ease of notation we fix $K = [0, 1]^d$ as discussed in Remark 7.1. Next we provide examples of bounded and high-order smooth functions.

Example 7.5. *Let q be a polynomial of degree k, with $q \in L_\infty(B)$ then obviously $q \in \mathcal{L}(B, 0)$. Also for $d = 1$, the trigonometric functions* $\sin, \cos$ *lie inside $\mathcal{L}(1, 1)$. The exponential function $x \mapsto \exp(c \cdot x)$ lies inside $\mathcal{L}(e^{\max(c,0)}, c)$ when $K = [0, 1]$. Also if $f \in \mathcal{L}(B, M)$ and $g \in \mathcal{L}(B', M')$ then $f + g \in \mathcal{L}(B + B', M + M')$. If $f \in \mathcal{L}(B, M, 0)$*

and $g \in \mathcal{L}(B', M', 0)$ then $f \cdot g \in \mathcal{L}(B \cdot B', 2 \cdot M \cdot M', 0)$. On the other hand the log *function is not in $\mathcal{L}(B, M)$ for any B, M since $\left\|\mathbf{D}^k \log\right\|_\infty \geq (k-1)!$ and hence it cannot be bounded by any exponential function.*

Probability Distributions

We are now ready to define the main object that we study in our paper, which is probability distributions with a given log-density function.

Definition 7.6. *Let $S \subseteq \mathbb{R}^d$ and let $f : S \to \mathbb{R}$ such that $\int_S \exp(f(\boldsymbol{x}))d\boldsymbol{x} < \infty$. We define the distribution $D(f, S)$ with log-density f supported on S to be the distribution with density*

$$D(f, S; \boldsymbol{x}) = \frac{\mathbb{1}_S(\boldsymbol{x})\ e^{f(\boldsymbol{x})}}{\int_S e^{f(\boldsymbol{x})}\, \mathrm{d}x} = \mathbb{1}_S(\boldsymbol{x})\ \exp(f(\boldsymbol{x}) - \psi(f, S)),$$

where $\psi(f, S) = \log \int_S e^{f(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$. If f is equal to a k degree polynomial $q_{\boldsymbol{v}} \in \mathcal{Q}_k$ with coefficients $\boldsymbol{v} \in \mathbb{R}^{t_k - 1}$ then instead of $D(q_{\boldsymbol{v}}, S)$ we abuse notation and we write $D(\boldsymbol{v}, S)$. Finally, let $T \subseteq S$, we define $D(f, S; T) = \int_T D(f, S; \boldsymbol{x}) \mathrm{d}\boldsymbol{x}$.

Our main focus in this paper is on probability distributions $D(f, [0, 1]^d)$ where $f \in \mathcal{L}(B, M)$ for some known parameters B, M. More specifically our main goal is to approximation the density of $D(f, [0, 1]^d)$ given samples from $D(f, S)$, where S is a measurable subset of $[0, 1]^d$.

As we already mentioned in Section 1.5, in this work we provide provable extrapolation of non-parametric density functions from samples, i.e., given samples from the conditional density on some subset S of the support we want to recover the shape of the density function *outside* of S. We consider densities proportional to $e^{f(x)}$, where f is a sufficiently smooth function. Our observation consists of samples from a density proportional to $\mathbb{1}_S(x)e^{f(x)}$, where S is a known (via a membership oracle) subset of the support. For this problem to even be well-posed we need further assumptions on the density function. Even if we are given the exact conditional density $\mathbb{1}_S(x)e^{f(x)}$, it is easy to see that, if $f \notin C_\infty$, i.e., if f is not infinitely times differentiable everywhere in the whole support, there is no hope to extrapolate its curve outside of S; for a simple example, if we observe a density proportional to $e^{|x|}$ truncated in $(-\infty, 0]$ we cannot extrapolate this density to $(0, +\infty)$, because we cannot distinguish whether we are observing truncated samples from e^{-x} or $e^{|x|}$. On the other hand, if the log-density f is analytic and sufficiently smooth, then the value of f at every x can be determined only from local information, namely its derivatives at a single point. This well known property of analytic functions is quantified by Taylor's remainder theorem. In this work we build upon this intuition and show that even given *samples* from a sufficiently smooth density and even if these samples are *conditioned in a small subset of the support* we can still determine the function in the entire support and most importantly this can be done in a statistically and computationally efficient way.

Our first result shows that the polynomial which maximizes the likelihood with respect to the *conditional* distribution $D(f, S)$ (let us call this polynomial the "MLE polynomial") approximates the density $e^{f(x)}$ *everywhere* on $[0, 1]$, i.e. the MLE polynomial has small extrapolation error. Observe, that this result cannot follow just from the fact that for example the Taylor polynomial extrapolates, because the MLE polynomial and the Taylor polynomial are in principle very different. While it is conceptually clear that the MLE polynomial of sufficiently large degree will have small interpolation error and hence will approximate well the density inside S, our result is the first to show that the same polynomial has small extrapolation error and hence approximates the density on the entire interval $[0, 1]$.

Theorem 7.7 (Information Projection Extrapolation Error). *Let $I = [0, 1] \subseteq \mathbb{R}$ and $f : I \mapsto \mathbb{R}$ be a function that is $(k+1)$-times continuously differentiable on I, with $\left|f^{(k+1)}(x)\right| \leq M^{k+1}$ for all $x \in I$. Let $S \subseteq I$ be a measurable set such that* $\mathrm{vol}(S) > 0$, *and p be a polynomial of degree at most k defined as*

$$p = \operatorname*{argmin}_{q \in \mathcal{Q}_k} \mathrm{D}_{\mathrm{KL}}(D(f, S) \| D(q, S)) .$$

Then, it holds that $\mathrm{D}_{\mathrm{KL}}(D(f,I)\|D(p,I)) \leq e^{W_k} W_k^2$, $W_k = \frac{M^{k+1}}{(k+1)!}$.

Extending the previous result to multivariate densities is significantly more challenging. The reason is that multivariate polynomial interpolation is much more intricate and is a subject of active research, see for example the survey Gasca and Sauer (2000). Instead of trying to characterize the properties of the exact MLE polynomial we give an alternative method for obtaining multivariate extrapolation guarantees that does not rely on multivariate polynomial interpolation. Our approach uses the additional assumption that the set S from which we observe samples has non-trivial volume, that is $\mathrm{vol}(S) \geq a$ for some $\alpha > 0$. Under this natural assumption we obtain the following theorem.

Theorem 7.8 (Extrapolation Error of MLE)**.** *Let* $K = [0,1]^d \subseteq \mathbb{R}^d$*,* $f \in \mathcal{L}(B,M)$ *be function supported on* K*, and* $S \subseteq K$ *be a measurable subset of* K *such that* $\mathrm{vol}(S) \geq \alpha$*. Moreover, define*

$$k = \widetilde{\Omega}\left(\frac{d^3 M}{\alpha^2} + \log\left(\frac{2^B}{\epsilon}\right)\right), \quad D = \{\boldsymbol{v} : \max_{\boldsymbol{x}\in K} \left|\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})\right| \leq 3B\},$$

and $r_k^* = \min_{\boldsymbol{u}\in D} \mathrm{D}_{\mathrm{KL}}(D(f,S)\|D(\boldsymbol{u},S))$*. Then, for every* $\boldsymbol{u} \in D$ *such that*

$$\mathrm{D}_{\mathrm{KL}}(D(f,S)\|D(\boldsymbol{u},S)) \leq r_k^* + \exp\left(-\widetilde{\Omega}\left(\frac{d^3 M}{\alpha^2} + B\right)\right) \cdot \left(\frac{1}{\epsilon}\right)^{-\Omega(\log(d/\alpha))},$$

it holds that $d_{\mathrm{TV}}(D(f,K), D(\boldsymbol{u},K)) \leq \epsilon$*.*

So far we have argued about the extrapolation error of the population MLE polynomial, i.e., we assume that we have access to the population distribution $D(f,S)$ and that we can maximize the population MLE with no error. Therefore, our next step is to show how we can incorporate the statistical error from the access to only finitely many samples from $D(f,S)$ and to provide an efficient algorithm that computes the MLE polynomial with small enough approximation loss.

Theorem 7.9 (Multi-Dimensional Statistical Taylor Theorem)**.** *Let* $f : [0,1]^d \to \mathbb{R}$ *with* $f \in \mathcal{L}(B,M)$ *and* $S \subseteq [0,1]^d$ *such that* $\mathrm{vol}(S) \geq \alpha$*. Let* $k = \widetilde{\Omega}(d^3 M/\alpha^2 + B) +$

$2\log(1/\epsilon)$*, then there exists an algorithm that uses* $N = 2^{\widetilde{O}(d^4 M/\alpha^2 + Bd)} \cdot (1/\epsilon)^{O(d+\log(1/\alpha))}$ *samples from* $D(f, S)$*, runs in* $\mathrm{poly}(N)$ *time and outputs a vector of coefficients* $\boldsymbol{v}$ *such that* $d_{\mathrm{TV}}(D(f,K), D(q_{\boldsymbol{v}}, K)) \leq \epsilon$*.*

7.2 Single Dimensional Densities

In this section we show our Statistical Taylor Theorem for single-dimensional densities. We keep this analysis separate from our main multi-dimensional theorem for several reasons. First, there exists a great body of work on single-dimensional non-parametric estimation problems in the vanilla setting and more specifically in truncated estimation problems this is the only setting that has been considered so far. Therefore, it is easier to compare the estimators and results that we get with the existing results. In fact this is the strategy that is followed in other multi-dimensional non-parametric estimation problems, e.g., see Wu (2010). Another reason is that in the single dimensional setting we are able to obtain a slightly stronger information theoretic result using more elementary tools, although the analysis of our efficient algorithmic procedure is the same as in multiple dimensions. Finally, the single dimensional setting serves as a nice example where the difference between interpolation and extrapolation is more clear.

In this section our goal is to estimate the density of the distribution $D(f, [0,1])$ using only samples from $D(f, S)$, where the log-density f is a bounded and high-order smooth function, i.e. $f \in \mathcal{L}(B, M)$, and S is a measurable subset of $[0,1]$. As a first step we need to understand what is a sufficient degree for a polynomial to well-approximate (Section 7.2) this is the part that is different compared to the multi-dimensional case that we present in Section 7.3. Them in Section 7.2 we state the application of our general multi-dimensional statistical and computational result to the single dimensional case where the assumptions and guarantees have a simpler form.

Identifying the Sufficient Degree

In this section we assume population access to $D(f, S)$ and our goal is to identify a polynomial q such that $D(q, [0, 1])$ approximates $D(f, [0, 1])$. In particular, we want to answer the question: if q minimizes the KL-divergence between $D(q, S)$ and $D(f, S)$, what can we say about the total variation distance between $D(q, [0, 1])$ and $D(f, [0, 1])$? Moreover, how does this depend on the degree of q? As the degree of q grows, it certainly allows $D(q, S)$ to come closer to $D(f, S)$. The natural thing to expect hence is that the same is true of $D(q, [0, 1])$, coming closer to $D(f, [0, 1])$. Unfortunately, this is not necessarily the case, because it could be that, as the degree of the polynomial increases, the approximant $D(q, S)$ overfits to $D(f, S)$, so extrapolating to $[0, 1]$ fails to give a good approximation to $D(f, [0, 1])$. This is the main technical difficulty of this section.

In the next theorem, we show is that if the function is high-order smooth, then there is some threshold beyond which we get better approximations using higher degrees, i.e. overfitting does not happen for any degree above some threshold. We illustrate this behavior in Example 7.11. We now restate Theorem 7.7 for convenience.

Theorem 7.10 (Information Projection Extrapolation Error)**.** *Let $I = [0, 1] \subseteq \mathbb{R}$ and $f : I \mapsto \mathbb{R}$ be a function that is $(k+1)$-times continuously differentiable on I, with $\left|f^{(k+1)}(x)\right| \leq M^{k+1}$ for all $x \in I$. Let $S \subseteq I$ be a measurable set such that* $\mathrm{vol}(S) > 0$, *and p be a polynomial of degree at most k defined as*

$$p = \operatorname*{argmin}_{q \in \mathcal{Q}_k} \mathrm{D_{KL}}(D(f, S) \| D(q, S)) .$$

Then, it holds that $\mathrm{D_{KL}}(D(f, I) \| D(p, I)) \leq e^{W_k} W_k^2$, $W_k = \frac{M^{k+1}}{(k+1)!}$.

The proof of Theorem 7.7 can be found in Appendix F.2. A weaker version of this theorem and be proved by applying the multi-dimensional Lemma 7.16 for $d = 1$. Nevertheless, Theorem 7.7 is slightly stronger its proof is more elementary. We also note that we can prove a more general version of Theorem 7.7 where I

is any interval $[a, b]$. The difference in the guarantees is that the term W_k will be multiplied by R^{k+1} where $R \triangleq b - a$.

To convey the motivation for our theorem and illustrate its guarantees, we use the following example.

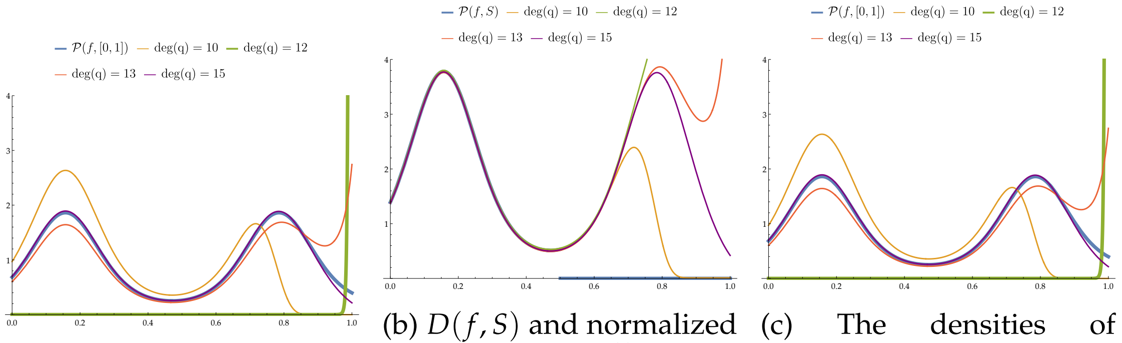
Example 7.11. *Let $f(x) = \sin(10 \cdot x)$ and $S = [0, 1/2]$. Our goal is to estimate the probability distribution $D(f, [0, 1])$ using only samples from $D(f, S)$. The guarantees of Theorem 7.7 are illustrated in Figure 7.1 where we can see the density of the distributions $D(f, [0, 1])$, $D(f, S)$, $D(q, S)$ and $D(q, [0, 1])$ for various values of the degree of q, where q is always chosen to minimize the KL-divergence between $D(f, S)$ and $D(q, S)$, i.e., using no further information about $D(f, [0, 1])$. As we see, $D(q, S)$ approximates $D(f, S)$ very well for each one of the presented degrees, with a marginal improvement in the quality of approximation as the degree of q is increased.*

An important observation is that for small values of the degree of q the approximation error between $D(f, [0, 1])$ and $D(q, [0, 1])$ is not monotone in the degree of q. In particular, when the degree of q is 10 *then $D(q, [0, 1])$ is reasonably close to $D(f, [0, 1])$ while when the degree of q becomes* 12 *then $D(q, [0, 1])$ is way off. This suggests that the overfitting occurs for degree equal to* 12*. This is the point of Theorem 7.7. It guarantees that, for degree greater than a threshold, overfitting cannot happen and $D(q, [0, 1])$ will always be a good approximation to $D(f, [0, 1])$.*

Handling the Optimization Error

Our next goal is to provide a version of Theorem 7.7 where approximation error is also introduced, due to finite samples.

Theorem 7.12 (Approximate Information Projection Extrapolation Error)**.** *Let $I = [0, 1]$, $f \in \mathcal{L}(B, M)$ be a function supported on I, $S \subseteq I$ be a measurable set such that $\mathrm{vol}(S) \geq \alpha$, D_k be the convex set $D_k = \{p \in \mathcal{Q}_k : p \in L_\infty(I, 3B)\}$., where $k = \Omega(M + \log(1/\epsilon))$, and let also $r_k^* = \min_{p \in D_k} \mathrm{D}_{\mathrm{KL}}(D(f, S) \| D(p, S))$. If some q*



(a) The densities of $D(f,[0,1])$ and $D(f,S)$.

(b) $D(f,S)$ and normalized on S functions $e^{q(\cdot)}$ for different $\deg(q)$.

(c) The densities of $D(f,[0,1])$ and $D(q,[0,1])$ for different $\deg(q)$.

Figure 7.1: In figure (a) we can see the probability density functions of the distributions $D(f,[0,1])$ and $D(f,S)$. In figure (b) we have the density of $D(f,S)$ together with the functions $\exp(q(x))$ for various degrees of q normalized so that the integral on S is 1. As we can see all the degrees approximate very well the conditional density but they have completely different behavior outside S. In figure (c) we can see the densities $D(f,[0,1])$ and $D(q,[0,1])$ for various degrees of q. The difference between (b) and (c) is that in (c) the functions are normalized so that their integral over $[0,1]$ is equal to 1 whereas in (b) the integral over S is equal to 1.

with $q \in D_k$ satisfies

$$\mathrm{D_{KL}}(D(f,S) \| D(q,S)) \le r_k^* + 2^{-O(k\log(1/\alpha)+B)}, \tag{7.4}$$

then it holds that $d_{\mathrm{TV}}(D(f,I),D(q,I)) \le \epsilon$.

The proof of Theorem 7.12 can be found in Appendix F.2. Now what is left to do is to argue that we can efficiently compute a polynomial q that satisfies (7.4). Unfortunately, the proof of this step is not simplified in the single dimensional case and we need to invoke our general multi-dimensional theorem with the assumptions and guarantees simplified due to the single dimensionality of the distribution. For more details about the specific algorithm that we use we refer to Section 7.3. A theorem that summarizes all these steps is the following.

Theorem 7.13 (1-D Statistical Taylor Theorem). *Let $I = [0,1]$, $f \in \mathcal{L}(B,M)$ be a function supported on I, and $S \subseteq I$ be a (measurable) subset of I such that* $\mathrm{vol}(S) \ge$

α. There exists an algorithm that draws $N = 2^{\widetilde{O}((M+\log(1/\epsilon))\log(1/\alpha)+B)}$ samples from $D(f,S)$, runs in time polynomial in the number of samples, and outputs a vector of coefficients $\boldsymbol{v}$ such that $d_{\mathrm{TV}}(D(f,K),D(q_{\boldsymbol{v}},K)) \le \epsilon$.

Proof. We define the convex set $D_k = \left\{\boldsymbol{v} \in \mathbb{R}^m \ : \ \max_{\boldsymbol{x}\in[0,1]^d} \left|\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})\right| \le 3B\right\}$, where $m = \binom{d+k}{k} - 1$. From Theorem 7.12 we know that it suffices to fix some degree $k = O(M + \log(1/\epsilon))$ and find a candidate $\boldsymbol{v}$ such that the conditional Kullback-Leibler divergence $\mathrm{D_{KL}}(D(f,S)\|D(q_{\boldsymbol{v}},S)) \le \min_{\boldsymbol{u}\in D_k} \mathrm{D_{KL}}(D(f,S)\|D(q_{\boldsymbol{u}},S)) + 2^{-O(k\log(1/\alpha)+B)}$. Therefore, from Theorem 7.20 we obtain that with $N = 2^{\widetilde{O}(k\log(1/\alpha)+B)}$ samples and in time poly(N), we can compute such a candidate. □

7.3 Multi-Dimensional Densities

In this section we show the general form of our Statistical Taylor Theorem that applies to multi-dimensional densities. Although the techniques used in this section and involved and possibly of independent interest, our strategy to prove this theorem is similar to the strategy that we followed in Section 7.2: (1) we identify the sufficient degree that we need to use, (2) we handle approximation errors that we get as a result of finite sample, and (3) we design an efficient algorithm to compute the solutions that are information-theoretically shown to exist. Putting all these together we prove the following which is the main theorem of our paper. We now restate Theorem 7.9 for convenience.

Theorem 7.14 (Multi-Dimensional Statistical Taylor Theorem)**.** *Let $f : [0,1]^d \to \mathbb{R}$ with $f \in \mathcal{L}(B,M)$ and $S \subseteq [0,1]^d$ such that $\mathrm{vol}(S) \ge \alpha$. Let $k = \widetilde{\Omega}(d^3M/\alpha^2 + B) + 2\log(1/\epsilon)$, then there exists an algorithm that uses $N = 2^{\widetilde{O}(d^4M/\alpha^2+Bd)} \cdot (1/\epsilon)^{O(d+\log(1/\alpha))}$ samples from $D(f,S)$, runs in* poly(N) *time and outputs a vector of coefficients $\boldsymbol{v}$ such that $d_{\mathrm{TV}}(D(f,K),D(q_{\boldsymbol{v}},K)) \le \epsilon$.*

Towards proving the above theorem the main bottleneck is that in the multi-dimensional polynomial interpolation theory there are no sufficient tools to prove the extrapolation properties of an estimator that can be computed efficiently. Surprisingly, our formulation of the extrapolation problem in the language of density estimation enables us to use strong anti-concentration results to prove extrapolation results for polynomial approximations.

We begin this section with a presentation of our main lemma in this direction which we call "Distortion of Conditioning Lemma" and we believe is of independent interest and could be useful in other multi-dimensional extrapolation problems.

Identifying the Sufficient Degree – The Distortion of Conditioning Lemma

The goal of this section is to identify the sufficient degree so that the MLE polynomial approximates well the true density in the whole domain $K = [0,1]^d$, i.e., it has small extrapolation error. We restate Theorem 7.8 for convenience.

Theorem 7.15 (Extrapolation Error of MLE). *Let $K = [0,1]^d \subseteq \mathbb{R}^d$, $f \in \mathcal{L}(B, M)$ be function supported on K, and $S \subseteq K$ be a measurable subset of K such that $\mathrm{vol}(S) \geq \alpha$. Moreover, define*

$$k = \widetilde{\Omega}\left(\frac{d^3 M}{\alpha^2} + \log\left(\frac{2^B}{\epsilon}\right)\right),\ D = \{\boldsymbol{v} : \max_{\boldsymbol{x} \in K} \left|\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})\right| \leq 3B\},$$

and $r_k^* = \min_{\boldsymbol{u} \in D} \mathrm{D_{KL}}(D(f,S) \| D(\boldsymbol{u},S))$. Then, for every $\boldsymbol{u} \in D$ such that

$$\mathrm{D_{KL}}(D(f,S) \| D(\boldsymbol{u},S)) \leq r_k^* + \exp\left(-\widetilde{\Omega}\left(\frac{d^3 M}{\alpha^2} + B\right)\right) \cdot \left(\frac{1}{\epsilon}\right)^{-\Omega(\log(d/\alpha))},$$

it holds that $d_{\mathrm{TV}}(D(f,K), D(\boldsymbol{u},K)) \leq \epsilon$.

The first step in proving Theorem 7.8 is to understand the approximation error as a function of the degree that we use when we have access to the population distribution $D(f,S)$. This is established in the following lemma whose proof can be found in Appendix F.3.

Lemma 7.16 (Approximation of Log-density). *Let $K \subseteq \mathbb{R}^d$ be a convex set centered at the origin $\mathbf{0}$ of diameter $\mathrm{diam}_\infty(K) \leq R$ and let $f \in \mathcal{L}(B, M)$ be a function supported on K. There exists polynomial $p(\boldsymbol{x}) = \boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x}) \in \mathcal{Q}_{d,k}$ such that for every $S \subseteq K$ it holds*

$$\mathrm{D}_{\mathrm{KL}}(D(f,S) \| D(\boldsymbol{v},S)) \leq 2\left(\frac{15MRd}{k}\right)^{k+1}, \quad \textit{and} \quad \|\mathbb{1}_K p\|_\infty \leq 2B + \left(\frac{15MRd}{k}\right)^{k+1}.$$

From Lemma 7.16 we obtain that by choosing $\mathbf{0} \in K$, there exists $\boldsymbol{v}$ such that $\left\|\mathbb{1}_K \boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})\right\|_\infty \leq 2B + (15Md/k)^{k+1}$. Moreover, from the same lemma we have that $\min_{\boldsymbol{w} \in D} \mathrm{D}_{\mathrm{KL}}(D(f,S) \| D(\boldsymbol{w},S)) \leq \mathrm{D}_{\mathrm{KL}}(D(f,S) \| D(\boldsymbol{v},S)) \leq 2\,(15Md/k)^{k+1}$. To simplify notation set $r_k = 2\,(15Md/k)^{k+1}$. Now, let $q(\boldsymbol{x}) = \boldsymbol{u}^T \boldsymbol{m}_k(\boldsymbol{x})$ be any approximate minimizer in D of the KL-divergence between $D(q,S)$ and $D(f,S)$ that satisfies

$$\mathrm{D}_{\mathrm{KL}}(D(f,S) \| D(\boldsymbol{u},S)) \leq \min_{\boldsymbol{w} \in D} \mathrm{D}_{\mathrm{KL}}(D(f,S) \| D(\boldsymbol{w},S)) + \bar{\epsilon} \leq r_k + \bar{\epsilon}.$$

This bound implies the following via Pinsker's inequality and the subadditivity of the square root

$$d_{\mathrm{TV}}(D(f,S), D(\boldsymbol{u},S)) \leq \sqrt{r_k} + \sqrt{\bar{\epsilon}}. \tag{7.5}$$

Our next step is to relate the conditional total variation $d_{\mathrm{TV}}(D(f,S), D(\boldsymbol{u},S))$ with the total variation in the whole domain $d_{\mathrm{TV}}(D(f,K), D(\boldsymbol{u},K))$. For this we develop a novel extrapolation technique based on anti-concentration of polynomial functions. In particular we use the following Theorem from Carbery and Wright (2001).

Theorem 7.17 (Theorem 2 of Carbery and Wright (2001)). *Let $K = [0,1]^d$ and let $p : \mathbb{R}^d \mapsto \mathbb{R}$ be a polynomial of degree at most k. If $q \in [1, +\infty]$, then there exists absolute constant C such that for any $\gamma > 0$ it holds*

$$\left(\int_K |p(\boldsymbol{x})|^{q/k} \mathrm{d}\boldsymbol{x}\right)^{1/q} \int_K \mathbb{1}\{|p(\boldsymbol{x})| \leq \gamma\} \mathrm{d}\boldsymbol{x} \leq C\gamma^{1/k} \, \min(q,d)\,.$$

This anti-concentration result is very useful for extrapolation because it can be used to bound the behavior of a polynomial function even outside the region from which we get the samples. This is the main idea of the following lemma which is one of the main technical contributions of the paper and we believe that it is of independent interest.

Lemma 7.18 (Distortion of Conditioning)**.** *Let $K = [0,1]^d$ and let p, q be polynomials of degree at most k such that $p, q \in L_\infty(K, B)$. For every $S \subseteq K$ with $\mathrm{vol}(S) > 0$ it holds that*

$$e^{-2B}\mathrm{vol}(S) \ \le \frac{d_{\mathrm{TV}}(D(p,K), D(q,K))}{d_{\mathrm{TV}}(D(p,S), D(q,S))} \le \ 8e^{5B}\frac{(2C\min(d,2k))^k}{\mathrm{vol}(S)^{k+1}}\,,$$

where C is the absolute constant of Theorem 7.17.

Remark 7.19. *Both Theorem 7.17 and Lemma 7.18 hold for the more general case where K is an arbitrary convex subset of $\mathbb{R}^d$. We choose to state this weaker expression for ease of notation and to be coherent with the rest of the paper.*

Unfortunately it is still not clear how to apply Lemma 7.18 to equation (7.18) because Lemma 7.18 assumes that both the distributions that we are comparing have as a log-density a bounded degree polynomial. Nevertheless, we can use a sequence of triangle inequalities togethet with and the multi-variate Taylor's Theorem (see Theorem 7.3) to combine Lemma 7.18 and equation (7.18) from which we can prove Theorem 7.8 by choosing the appropriate value for $\bar{\epsilon}$ as we explain in detail in Appendix F.3. The proof of the Distortion of Conditioning Lemma in presented in Appendix F.3.

Computing the MLE

In this section we describe an efficient algorithm that solves the Maximum Likelihood problem that we need in order to apply Theorem 7.8.

The efficient algorithm that we design in this section solves the following problem: given sample access to the conditional distribution $D(f, S)$ and fix a degree k, our algorithm finds a polynomial p of degree k that approximately minimizes the distance $\mathrm{D_{KL}}(D(f,S)\|D(\boldsymbol{u},S))$. More precisely we prove the following.

Theorem 7.20. *Let $f : [0,1]^d \to \mathbb{R}$ and $S \subseteq [0,1]^d$ with $\mathrm{vol}(S) \geq \alpha$. Fix a degree $k \in \mathbb{N}$ and a parameter $C > 0$ and define $D = \{\boldsymbol{v} \;:\; \max_{\boldsymbol{x}\in[0,1]^d} |\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})| \leq C\}$. There exists an algorithm that draws $N = 2^{O(dk)}(C^2/\epsilon)^2$ samples from $D(f,S)$, runs in time $2^{O(dk+C)}/(\alpha\epsilon^2)$, and outputs $\hat{\boldsymbol{v}} \in D$ such that $\mathrm{D_{KL}}(D(f,S)\|D(\hat{\boldsymbol{v}},S)) \leq \min_{\boldsymbol{u}\in D} \mathrm{D_{KL}}(D(f,S)\|D(\boldsymbol{u},S)) + \epsilon$, with probability at least 99%.*

The algorithm that we use for proving Theorem 7.20 is Projected Stochastic Gradient Descent with projection set D. In order to prove the guarantees of Theorem 7.20 we have to prove: (1) an upper bound on the number of steps that the PSGD algorithm needs, (2) find an efficient procedure to project to the set D. For the second we can use the celebrated algorithm by Renegar for the existential theory of reals Renegar (1992a,b), as we explain in detail in the appendix. To bound the number of steps that the PSGD performs we use the following lemma.

Lemma 7.21 (Theorem 14.8 of Shalev-Shwartz and Ben-David (2014c))**.** *Let $R, \rho > 0$. Let f be a convex function, $D \subseteq \mathbb{R}^d$ be a convex set of bounded diameter, $\mathrm{diam}_2(D) \leq R$, and let $\boldsymbol{w}^* \in \mathrm{argmin}_{\boldsymbol{w}\in D} f(\boldsymbol{w})$. Consider the following Projected Gradient Descent (PSGD) update rule $\boldsymbol{w}_{t+1} = \mathrm{proj}_D(\boldsymbol{w}_t - \eta \boldsymbol{v}_t)$, where $\boldsymbol{v}_t$ is an unbiased estimate of $\nabla f(\boldsymbol{w})$. Assume that PSGD is run for T iterations with $\eta = \sqrt{R^2\rho^2/T}$. Assume also that for all t, $\|\boldsymbol{v}_t\|_2 \leq \rho$ with probability 1. Then, for any $\epsilon > 0$, in order to achieve $\mathbf{E}[f(\bar{\boldsymbol{w}})] - f(\boldsymbol{w}^*) \leq \epsilon$ it suffices that $T \geq R^2\rho^2/\epsilon^2$.*

From the above lemma we can see that it remains to find an upper bound on the diameter of the set D and an upper bound on the norm of the stochastic gradient $\|\boldsymbol{v}_t\|$. The latter follows from some algebraic calculations whereas the first one from tight bounds on the coefficients of a polynomial with bounded values Ben-David et al. (2018). A detailed proof of Theorem 7.20 is presented in the Appendix F.3.

Putting Everything Together – The Proof of Theorem 7.9

From Theorem 7.8 we have that if we fix the degree $k = O(d^3 M/\alpha^2 + B) + 2\log(1/\epsilon)$ then it suffices to optimize the function $L(\boldsymbol{v})$ of Equation F.7 constrained in the convex set

$$D = \left\{ \boldsymbol{v} \in \mathbb{R}^m \;:\; \left\| \mathbb{1}_K \boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x}) \right\|_\infty \leq 3B \right\}.$$

From Theorem 7.8 we have that a vector $\boldsymbol{v}$ with optimality gap $2^{-\widetilde{\Omega}(d^3 M/\alpha^2 + B)}(1/\epsilon)^{-\Omega(\log(d/\alpha)}$ achieves the extrapolation guarantee $d_{\mathrm{TV}}(D(f,K), D(\boldsymbol{v}_T, K)) \leq \epsilon$. From Theorem 7.20 we have that there exists an algorithm that achieves this optimality gap with sample complexity

$$N = B^2 2^{O(dk)} 2^{\widetilde{O}(d^3 M/\alpha^2 + B)} (1/\epsilon)^{O(\log(d/\alpha))} = 2^{\widetilde{O}(d^4 M/\alpha^2 + Bd)} (1/\epsilon)^{d + \log(1/\alpha)}\,.$$

□

8 LEARNING FROM COARSE DATA

8.1 Formal Statement of Results

Notation and Preliminaries For a graph G, we denote by $\boldsymbol{L}_G$ its Laplacian matrix. We denote $\mathcal{B}(\boldsymbol{x}, \rho)$ the Euclidean ball of radius ρ centered at $\boldsymbol{x}$; we simply refer to $\mathcal{B}$ if the radius and the center are clear from the context and we denote the associated sphere $\partial\mathcal{B}$, i.e., its boundary. The probability simplex is denoted by Δ^n and discrete distributions D supported on $[n]$ will usually be represented by their associated probability vectors $\boldsymbol{p} \in \Delta^n$. For any distribution D, we overload the notation and we use the same notation for the corresponding density and denote $D(S) = \sum_{x \in S} D(x)$ for any $S \subseteq [n]$. We denote the support of the probability distribution D by $\mathrm{supp}(D)$. The d-dimensional Gaussian distribution will be denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. When the covariance matrix is known, we simplify to $\mathcal{N}(\boldsymbol{\mu})$. For a set $S \subseteq \mathbb{R}^d$, we let $\mathcal{N}_S$ denote the conditional Gaussian distribution on the set S, i.e., $\mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) = \mathbf{1}\{\boldsymbol{x} \in S\}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) / \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S)$. We denote Φ (resp. ϕ) the cdf (resp. pdf) of the standard Normal distribution. The total variation distance of $\boldsymbol{p}, \boldsymbol{q} \in \Delta^n$ is $d_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q}) = \max_{S \subseteq [n]} \boldsymbol{p}(S) - \boldsymbol{q}(S) = \|\boldsymbol{p} - \boldsymbol{q}\|_1 / 2$. Let D be a joint distribution over labeled examples $\mathcal{X} \times \mathcal{Z}$, with $\mathcal{X}$ be the input space and $\mathcal{Z}$ the label space. A statistical query (SQ) oracle $\mathrm{STAT}(D, \tau)$ with tolerance parameter $\tau \in [0, 1]$ takes as input a statistical query defined by a real-valued function $q : \mathcal{X} \times \mathcal{Z} \to [-1, 1]$ and outputs an estimate of $\mathbf{E}_{(x,z) \sim D}[q(x, z)]$ that is accurate to within an additive $\pm\tau$.

We start by describing the generative model of coarsely labeled data in the supervised setting. We model coarse labels as subsets of the domain of all possible fine labels.

Definition 8.1 (Generative Process of Coarse Data with Context)**.** *Let $\mathcal{X}$ be an arbitrary domain, and let $\mathcal{Z} = \{1, \ldots, k\}$ be the discrete domain of all possible fine labels. We generate coarsely labeled examples as follows:*

1. *Draw a finely labeled example (x, z) from a distribution D on $\mathcal{X} \times \mathcal{Z}$.*

2. *Draw a coarsening partition $\mathcal{S}$ (of $\mathcal{Z}$) from a distribution π.*

3. *Find the unique set $S \in \mathcal{S}$ that contains the fine label z.*

4. *Observe the coarsely labeled example (x, S).*

We denote D_π the distribution of the coarsely labeled example (x, S).

In the supervised setting, our main focus is to answer the following question.

Question 8.2. *Can we train a model, using coarsely labeled examples $(x, S) \sim D_\pi$, that classifies finely labeled examples $(x, z) \sim D$ with accuracy comparable to that of a classifier that was trained on examples with fine grained labels?*

Definition 8.1 does not impose any restrictions on the distribution over partitions π. It is clear that if partitions are very rough, e.g., we split $\mathcal{Z}$ into two large disjoint subsets, we lose information about the fine labels and we cannot hope to train a classifier that performs well over finely labeled examples. In order for Question 8.2 to be information theoretically possible, we need to assume that the partition distribution π preserves fine-label information. The following definition quantifies this by stating that reasonable partition distributions π are those that preserve the total variation distance between different distributions supported on the domain of the fine labels $\mathcal{Z}$. We remark that the following definition does not require $\mathcal{D}$ to be supported on pairs (x, z) but is a general statement for the unsupervised version of the problem, see also Definition 8.10.

Definition 8.3 (Information Preserving Partition Distribution)**.** *Let $\mathcal{Z}$ be any domain and let $\alpha \in (0, 1]$. We say that π is an α-information preserving partition distribution if for every two distributions D^1, D^2 supported on $\mathcal{Z}$, it holds that $d_{\mathrm{TV}}(D^1_\pi, D^2_\pi) \geq \alpha \cdot d_{\mathrm{TV}}(D^1, D^2)$, where $d_{\mathrm{TV}}(D^1, D^2)$ is the total variation distance of D^1 and D^2.*

Our first result is a positive answer to Question 8.2 in essentially full generality: we show that concept classes that are efficiently learnable in the Statistical Query

(SQ) model, Kearns (1998), are also learnable from coarsely labeled examples. Our result is similar in spirit with the result of Kearns (1998), where it is proved that SQ learnability implies learnability under random classification noise.

Theorem 8.4 (SQ from Coarsely Labeled Examples)**.** *Consider a distribution D_π over coarsely labeled examples in $\mathbb{R}^d \times [k]$, (see Definition 8.1) with α-information preserving partition distribution π. Let $q : \mathbb{R}^d \times [k] \to [-1, 1]$ be a query function, that can be evaluated on any input in time T, and $\tau, \delta \in (0, 1)$. There exists an algorithm (Algorithm 12), that draws $N = \widetilde{O}(k^4/(\tau^3 \alpha^2) \log(1/\delta))$ coarsely labeled examples from D_π and, in* $\mathrm{poly}(N, T)$ *time, computes an estimate $\hat{r}$ such that, with probability at least $1 - \delta$, it holds $\left| \mathbf{E}_{(x,z)\sim D}[q(x, z)] - \hat{r} \right| \leq \tau$.*

Learning Parametric Distributions from Coarse Samples. In many important applications, instead of a discrete distribution over fine labels, a continuous parametric model is used. A popular example is when the domain $\mathcal{Z}$ of Definition 8.1 is the entire Euclidean space $\mathbb{R}^d$, and the distribution of finely labeled examples is a Gaussian distribution whose parameters possibly depend on the context x. Such censored regression settings are known as Tobit models Tobin (1958); Maddala (1986); Gourieroux (2000). Lately, significant progress has been made from a computational point of view in such censored/truncated settings in the distribution specific setting, e.g., when the underlying distribution is Gaussian Daskalakis et al. (2018); Kontonis et al. (2019), mixtures of Gaussians Nagarajan and Panageas (2019), linear regression Daskalakis et al. (2019); Ilyas et al. (2020); Daskalakis et al. (2020). In this distribution specific setting, we consider the most fundamental problem of learning the mean of a Gaussian distribution given coarse data.

Definition 8.5 (Coarse Gaussian Data)**.** *Consider the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^\star)$, with mean $\boldsymbol{\mu}^\star \in \mathbb{R}^d$ and identity covariance matrix. We generate a sample as follows:*

1. *Draw $\mathbf{z}$ from $\mathcal{N}(\boldsymbol{\mu}^\star)$.*

2. *Draw a partition $\mathbf{\Sigma}$ (of $\mathbb{R}^d$) from π.*

3. *Observe the set $S \in \mathbf{\Sigma}$ that contains $\mathbf{z}$.*

We denote the distribution of S as $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$.

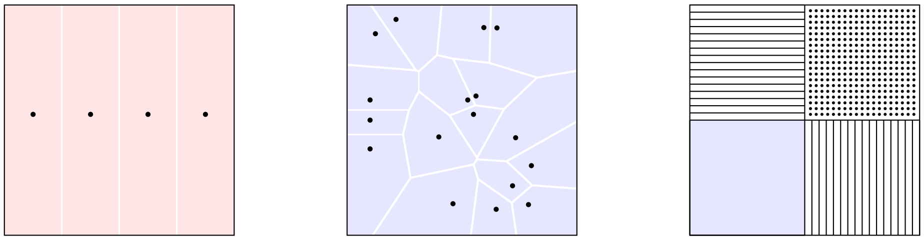
Remark 8.6. *We remark that we only require membership oracle access to the subsets of the partition $\mathcal{S}$. A set $S \subseteq \mathbb{R}^d$ corresponds to a membership oracle $\mathcal{O}_S : \mathbb{R}^d \to \{0, 1\}$ that given $\boldsymbol{x} \in \mathbb{R}^d$ outputs whether the point lies inside the set S or not.*

We first study the above problem, from a computational viewpoint. For the corresponding problems in censored and truncated statistics no geometric assumptions are required for the sets: in Daskalakis et al. (2018) it was shown that an efficient algorithm exists for arbitrarily complex truncation sets. In contrast in our more general model of coarse data we show that having sets with geometric structure is necessary. In particular we require that every set of the partition is convex, see Figure 8.2(b,c). We show that when the convexity assumption is dropped, learning from coarse data is a computationally hard problem even under a mixture of very simple sets.

Theorem 8.7 (Hardness of Matching the Observed Distribution with General Partitions)**.** *Let π be a general partition distribution. Unless $\mathrm{RP} = \mathrm{NP}$, no algorithm with sample access to $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$, can compute, in $\mathrm{poly}(d)$ time, a $\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^d$ such that $d_{\mathrm{TV}}(\mathcal{N}_\pi(\widetilde{\boldsymbol{\mu}}), \mathcal{N}_\pi(\boldsymbol{\mu}^\star)) < 1/d^c$ for some absolute constant $c > 1$.*

We prove our hardness result using a reduction from the well known MAX-CUT problem, which is known to be NP-hard, even to approximate Håstad (2001). In our reduction, we use partitions that consist of simple sets: fat hyperplanes, ellipsoids and their complements: the computational hardness of this problem is rather inherent and not due to overly complicated sets.

On the positive side, we identify a geometric property that enables us to design a computationally efficient algorithm for this problem: namely we require all the sets of the partitions to be *convex*, e.g., Figure 8.2(b,c). We remark that having finite or countable subsets, is not a requirement of our model. For example, we can handle convex partitions of the form (c) that correspond to the output distribution of a ReLU neural network, see Wu et al. (2019). We continue with our theorem for learning Gaussians from coarse data.



(a) Non-Identifiable Case (b) Convex Partition Case (c) ReLU Case

Figure 8.1: Convex Partitions of $\mathbb{R}^2$.

Figure 8.2: (a) is a very rough partition, that makes learning the mean impossible: Gaussians $\mathcal{N}((0,z))$ centered along the same vertical line $(0,z)$ assign exactly the same probability to all cells of the partitions and therefore, $d_{\mathrm{T}V}(\mathcal{N}_\pi((0,z_1)),\mathcal{N}_\pi((0,z_2)))=0$: it is impossible to learn the second coordinate of the mean. (b) is a convex partition of $\mathbb{R}^2$, that makes recovering the Gaussian possible. (c) is the convex partition corresponding to the output distribution of one layer ReLU networks. When both coordinates are positive, we observe a fine sample (black points correspond to singleton sets). When exactly one coordinate (say $\boldsymbol{x}_1$) is positive, we observe the line $L_z=\{\boldsymbol{x}:\boldsymbol{x}_2<0,\boldsymbol{x}_1=z>0\}$ that corresponds to the ReLU output $(\boldsymbol{x}_1,0)$. If both coordinates are negative, we observe the set $\{\boldsymbol{x}:\boldsymbol{x}_1<0,\boldsymbol{x}_2<0\}$, that corresponds to the point $(0,0)$.

Theorem 8.8 (Gaussian Mean Estimation with Convex Partitions)**.** *Let $\epsilon,\delta\in(0,1)$. Consider the generative process of coarse d-dimensional Gaussian data $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$, as in Definition 8.5. Assume that the partition distribution π is α-information preserving and is supported on convex partitions of $\mathbb{R}^d$. The following hold.*

1. *The empirical log-likelihood objective*

$$\mathcal{L}_N(\boldsymbol{\mu})=\frac{1}{N}\sum_{i=1}^{N}\log\mathcal{N}(\boldsymbol{\mu};S_i)$$

   is concave with respect to $\boldsymbol{\mu}$ where the sets S_i for $i\in[N]$ are i.i.d. samples from $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$.

2. *There exists an algorithm, that draws $N=\widetilde{O}(d/(\epsilon^2\alpha^2)\log(1/\delta))$ samples from $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$ and computes an estimate $\widetilde{\boldsymbol{\mu}}$ that satisfies $d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}}),\mathcal{N}(\boldsymbol{\mu}^\star))\leq\epsilon$, with*

probability at least $1 - \delta$.

Our algorithm for mean estimation of a Gaussian distribution relies on the log-likelihood being concave when the partitions are convex. We remark that, similar to our approach, one can use the concavity of likelihood to get efficient algorithms for regression settings, e.g., Tobit models, where the mean of the Gaussian is given by a linear function of the context $\boldsymbol{A}\boldsymbol{x}$ for some unknown matrix $\boldsymbol{A}$.

8.2 Supervised Learning from Coarse Data

In this section, we consider the problem of *supervised* learning from coarse data. In this setting, there exists some underlying distribution over finely labeled examples, D. However, we have sample access only to the distribution associated with coarsely labeled examples D_π, see Definition 8.1. As discussed in Section 8.1, under this setting, even problems that are naturally convex when we have access to examples with fine labels, become non-convex when we introduce coarse labels (e.g., multiclass logistic regression). The main result of this section is Theorem 8.4, which allows us to compute statistical queries over finely labeled examples.

Overview of the Proof of Theorem 8.4

In order to simulate a statistical query we take a two step approach. Our first building block considers the unsupervised version of the problem, see Definition 8.10, i.e., we marginalize the context x and try to learn the distribution of the fine labels z given coarse samples S. This can be viewed as learning a general discrete distribution supported on $\mathcal{Z} = \{1, \dots, k\}$ given coarse samples, i.e., subsets of $\mathcal{Z}$. We show that, when the partition distribution π is α-information preserving, this can be done efficiently, see Proposition 8.11. Our algorithm (Algorithm 12) exploits the fact that even though in general having coarse data results in non-concave likelihood objectives, when we consider parametric models (see, for example, the case of logistic regression in Appendix G.2), this is not true when we maximize over all discrete distributions. In Proposition 8.11, we show that $\widetilde{O}(k/(\epsilon\alpha)^2)$ samples are sufficient for this step. For the details of this step, see Section 8.2.

Using the above algorithm, one could try to separately learn the marginal distribution over x, D_x and the distribution of the fine labels z *conditional on some fixed* x; let us denote this distribution as D_z^x. Then one could generate finely labeled examples (x, z) and use them to estimate the query $\mathbf{E}_{(x,z)\sim D}[q(x, z)]$. The reason that this naive approach fails is that it requires many coarse examples (x, S) with exactly the same value of x. Unless the domain $\mathcal{X}$ is very small, the probability that we observe samples with the same value of x is going to be tiny. In order to overcome this obstacle, at a high level, our approach is to split the domain $\mathcal{X}$ into larger sets and then, learn the conditional distribution of the labels, not on a fixed point x, but on these larger sets of non-trivial mass.

Intuitively, in order to have an effective partition of the domain $\mathcal{X}$, we want to group together points x whose values $q(x, z)$ are close. Since z belongs in a discrete domain $\mathcal{Z} = [k]$, we can decompose the query $q(x, z)$ as $q(x, z) = \sum_{i=1}^{k} q(x, i)\mathbf{1}\{z = i\}$. We estimate the value of $\mathbf{E}_{(x,z)\sim D}[q(x, i)\mathbf{1}\{z = i\}]$ separately. To find a suitable reweighting of the domain $\mathcal{X}$, we perform rejection sampling, accepting a pair $(x, S) \sim D$ with probability $q(x, i)$ [1]: points x that have small value $q(x, i)$ contribute less in the expectation and are less likely to be sampled. After performing this rejection sampling process based on x, we have pairs (x, S), conditional that x was accepted. Now, using our previous maximum likelihood learner of Proposition 8.11 we learn the marginal distribution over fine labels and use it to answer the query. We provide the details of this rejection sampling step in the full proof of Theorem 8.4, see Section 8.2.

For a description of the corresponding algorithm that simulates statistical queries, see Algorithm 12. To keep the presentation simple we state the algorithm for the case where the query function $q(x, z)$ is positive. It is straightforward to generalize it for general queries, see Section 8.2.

Remark 8.9 (Empirical Likelihood Approach)**.** *One could try to use the empirical*

[1] It is easy to handle the case where this function takes negative values, see the proof of Theorem 8.4.

Algorithm 12 Statistical Queries from Coarse Labels.

```
Input: Query $q : \mathcal{X} \times \mathcal{Z} \mapsto (0,1]$, tolerance $\tau \in [0,1]$, confidence $\delta \in [0,1]$.
Oracle: Access to coarsely labeled samples $(x, S) \sim D_\pi$, $\pi$ is $\alpha$-information preserving.
Output: Estimate $\widehat{r}$ such that $\left| \mathbf{E}_{(x,z)\sim D}[q(x,z)] - \widehat{r} \right| \leq \tau$ with probability at least $1 - \delta$.

procedure StatQuery($q, \tau, \delta$)
    Compute $\widehat{r}_i \leftarrow \mathrm{SQ}(q, i, O(\tau/k), \delta/k)$ for any $i \in \mathcal{Z}$.
    Output $\widehat{r} \leftarrow \sum_{i=1}^{k} \widehat{r}_i$.

procedure SQ($q, i, \rho, \delta$)
    Draw $N_1 = \widetilde{\Theta}(\frac{\log(1/\delta)}{\rho^2})$ samples $(x_j, S_j)$ from $D_\pi$.
    Compute $\widehat{\mu}_i \leftarrow \frac{1}{N_1} \sum_{j=1}^{N_1} q(x_j, i)$.
    if $\widehat{\mu}_i \leq \rho$ do
        Output $\widehat{r}_i \leftarrow 0$.
    end
    Draw $N_2 = \widetilde{\Theta}(\frac{k \log(1/\delta)}{\rho^3 \alpha^2})$ samples $(x_j, S_j)$ from $D_\pi$. ▷ $\widetilde{\Theta}(\frac{k^4 \log(1/\delta)}{\tau^3 \alpha^2})$ examples overall.
    $T_{accept} \leftarrow \emptyset$.    ▷ Training set of accepted samples.
    Add $S_j$ in $T_{accept}$ with probability $q(x_j, i)$, $\forall j \in [N_2]$.    ▷ Rejection Sampling Process.
    Compute $\widetilde{D}$ using Proposition 8.11 with input $(T_{accept}, \rho, \delta)$.
    Output $\widehat{r}_i \leftarrow \widehat{\mu}_i \cdot \widetilde{D}(i)$.
```

likelihood directly over the coarsely labeled data (as defined in Owen (2001)). However, in general, these empirical likelihood objectives are non-convex when the data are coarse and therefore it is computationally hard to optimize them directly. Our approach for simulating statistical queries consists of two ingredients: reweighting the feature space via rejection sampling in order to group together points and learning discrete distributions from coarse data. To learn the discrete distributions (Section 8.2), we use a (direct) empirical likelihood approach similar to that of Owen (1988); Owen et al. (1990); Owen (2001). However, our main contribution is the use of rejection sampling to reduce the initial non-convex problem to the special case of learning a discrete distribution (with small support) from coarse data *which, as we prove, is a tractable (convex) problem. For more connections with censored statistics techniques, we refer the reader to Thomas and Grunkemeier (1975); Owen (1988); Gill et al. (1997); Owen (2001).*

Learning Marginals Over Fine Labels

In this subsection, we deal with *unsupervised* learning from coarse data in discrete domains. Although this is an ingredient of our main result for simulating statistical queries in a supervised setting where labeled data (x, S) are given, the result of this section does not depend on the points x and concerns the unsupervised version of the problem. To keep the notation simple, we will use D to denote a distribution over finite labels $\mathcal{Z}$.

Definition 8.10 (Generative Process of Coarse Data)**.** *Let $\mathcal{Z}$ be a discrete domain and D be a distribution supported on $\mathcal{Z}$. Moreover, let π be a distribution supported on partitions of $\mathcal{Z}$. We consider the following generative process:*

1. *Draw z from D.*
2. *Draw a partition $\mathcal{S}$ from the distribution over all partitions π.*
3. *Observe the set $S \in \mathcal{S}$ that contains z.*

We denote the distribution of S as D_π.

The assumption that we require is that the partition distribution π is α-information preserving, see Definition 8.3. At this point we give some examples of information preserving partition distributions. We first observe that $\alpha = 0$ if and only if the problem is not identifiable. For instance, if π is supported only on the partition $\mathbf{\Sigma} = \{\{1, 2\}, \{3, \dots, k\}\}$, the problem is not identifiable, since, for example, the fine label 1 is indistinguishable from the fine label 2. The value $\alpha = 1$ is attained when the partition totally preserves the distribution distance. Intuitively, the value $1 - \alpha$ corresponds to the distortion that the coarse labeling introduces to a finely labeled dataset.

In many cases most fine labels may be missing. Consider two data providers that use different methods to round their samples. The rounding's uncertainty can be viewed as a coarse labeling of the data. Assume that we add discrete (balanced Bernoulli) noise ξ to some true value $x \in [0..k]$. Consider two partitions $\{\mathbf{\Sigma}_1, \mathbf{\Sigma}_2\}$ with $\mathbf{\Sigma}_1 = \{\{0,1\},\{2,3\},\ldots,\{k-1,k\},\{k+1\}\}$ and $\mathbf{\Sigma}_2 = \{\{0\},\{1,2\},\ldots\{k-1,k\}\}$. Observe that, when $x+\xi$ is odd, we can think of the rounded sample, as a draw from $\mathbf{\Sigma}_1$ and when $x+\xi$ is even, as a draw from $\mathbf{\Sigma}_2$. This example shows that we can capture the problem of deconvolution of two distributions D_1, D_2, where one of them is known and we observe samples $x_1 + x_2, x_i \sim D_i$.

The following proposition establishes the sample complexity of unsupervised learning of discrete distributions with coarse data. Our goal is to compute an estimate of the discrete distribution $D^\star$ with probability vector $\boldsymbol{p}^\star \in \Delta^k$ from N coarse samples $S_1, \ldots, S_N$ drawn from the distribution $D^\star_\pi$. Our algorithm maximizes the empirical likelihood. Analyzing the empirical log-likelihood objective $\mathcal{L}_N(\boldsymbol{p}) = \frac{1}{N}\sum_{n=1}^N \log\big(\sum_{i\in S_n} \boldsymbol{p}_i\big)$, where $\boldsymbol{p} \in \Delta^k$ is a guess probability vector, we observe that the problem is concave and, therefore, can be efficiently optimized (e.g., by gradient descent).

Proposition 8.11. *Let $\mathcal{Z}$ be a discrete domain of cardinality k and let D be a distribution supported on $\mathcal{Z}$. Moreover, let π be an α-information preserving partition distribution for some $\alpha \in (0,1]$. Then, with $N = \widetilde{O}(k/(\epsilon^2\alpha^2)\log(1/\delta))$ samples from D_π and in time polynomial in the number of samples N, we can compute a distribution $\widetilde{D}$ supported on $\mathcal{Z}$ such that $d_{\mathrm{TV}}(\widetilde{D}, D) \le \epsilon$.*

Proof. Let $D^\star$ be the target discrete distribution, supported on a discrete domain of size k, and let $\boldsymbol{p}^\star \in \Delta^k$ be the corresponding probability vector. For some distribution D supported on a discrete domain of size k, we define the following population log-likelihood objective.

$$\mathcal{L}(D) = \mathop{\mathbf{E}}_{S\sim D^\star_\pi}[\log D(S)] = \mathop{\mathbf{E}}_{S\sim D^\star_\pi}\Big[\log\big(\sum_{i\in S} D(i)\big)\Big]\,. \tag{8.1}$$

Since D is a discrete distribution for simplicity we may identify with its probability vector $\boldsymbol{p}$, where $\boldsymbol{p}_i = D(i)$. Therefore, for any $\boldsymbol{p}$ in the probability simplex Δ^k, we define

$$\mathcal{L}(\boldsymbol{p}) = \mathop{\mathbf{E}}_{S \sim D_\pi^\star} \left[\log \sum_{i \in S} \boldsymbol{p}_i \right] . \tag{8.2}$$

The corresponding empirical log-likelihood objective after drawing N independent samples $S_1, \ldots, S_N$ from $D_\pi^\star$ is given by

$$\mathcal{L}_N(\boldsymbol{p}) = \frac{1}{N} \sum_{n=1}^{N} \log \left(\sum_{i \in S_n} \boldsymbol{p}_i \right) . \tag{8.3}$$

We first observe that the log-likelihood (both the population and the empirical) is a concave function and therefore can be efficiently optimized (e.g., by gradient descent). Thus, our main focus in this proof is to bound its sample complexity. We first observe that when the guess $\boldsymbol{p} \in \Delta^k$ has some very biased coordinates, i.e., for some subset S the corresponding $\boldsymbol{p}_i$'s are close to 0, the probability of a set S, $\sum_{i \in S} p_i$ will be close to zero and therefore $\log \left(\sum_{i \in S} \boldsymbol{p}_i \right)$ will be large. Thus, we have to restrict our search to a subset of the probability simplex, i.e., have $\boldsymbol{p}_i \geq \epsilon/k$. We set $\widetilde{\Delta}^k = \{ \boldsymbol{p} \in \Delta^k, \boldsymbol{p}_i \geq \epsilon/k$ for all $i = 1, \ldots, k$ $\}$. We now prove that, given roughly $k/(\epsilon^2 \alpha^2)$ samples, we can guarantee that probability vectors that are far from the optimal vector $\boldsymbol{p}^\star$ will also be significantly sub-optimal in the sense that they are far from being maximizers of the empirical log-likelihood.

Claim 8.12. *Let $N \geq \widetilde{\Omega}(k/(\epsilon^2\alpha^2) \log(1/\delta))$. With probability at least $1 - \delta$, we have that, for every $\boldsymbol{p} \in \widetilde{\Delta}^k$ such that $\|\boldsymbol{p} - \boldsymbol{p}^\star\|_1 \geq \epsilon$, it holds*

$$\max_{\boldsymbol{q} \in \widetilde{\Delta}^k} \mathcal{L}_N(\boldsymbol{q}) - \mathcal{L}_N(\boldsymbol{p}) \geq \Omega \left((\epsilon\alpha)^2 \right) .$$

Proof. We first construct a cover of the probability simplex $\widetilde{\Delta}^k$ by discretizing each coordinate $\boldsymbol{p}_i$ to integer multiples of $O((\epsilon^{3/2}\alpha/k)^2)$. The resulting cover $\mathcal{C}$ contains $O((k/(\epsilon^{3/2}\alpha))^{2k})$ elements. We first observe that we can replace any element $\boldsymbol{p} \in \widetilde{\Delta}^k$ with an element $\boldsymbol{p}'$ inside our cover $\mathcal{C}$ without affecting the value of the objective $\mathcal{L}_N(\boldsymbol{p})$ by a lot. In particular, using the fact that $x \mapsto \log(x)$ is

$1/r$-Lipschitz in the interval $[r, +\infty)$, we have that for any set $S \subseteq \{1, \dots, k\}$ it holds

$$\Big| \log \Big(\sum_{i \in S} \boldsymbol{p}_i \Big) - \log \Big(\sum_{i \in S} \boldsymbol{q}_i \Big) \Big| \leq \frac{1}{\sum_{i \in S} \boldsymbol{p}_i} \Big| \sum_{i \in S} (\boldsymbol{p}_i - \boldsymbol{q}_i) \Big| \leq \frac{k}{\epsilon} \|\boldsymbol{p} - \boldsymbol{q}\|_1 ,$$

where we used the fact that, since $\boldsymbol{p} \in \widetilde{\Delta}^k$, it holds $\boldsymbol{p}_i \geq \epsilon/k$. Therefore, when we round each coordinate of a vector $\boldsymbol{p}$ to the closest integer multiple of $O((\epsilon^{3/2}\alpha/k)^2)$ we get a vector $\boldsymbol{p}' \in \mathcal{C}$ such that for any set S it holds $|\log(\sum_{i \in S} \boldsymbol{p}_i) - \log(\sum_{i \in S} \boldsymbol{q}_i)| \leq \epsilon^2\alpha^2/6$ which implies that the empirical log-likelihood satisfies $|\mathcal{L}_N(\boldsymbol{p}) - \mathcal{L}_N(\boldsymbol{p}')| \leq \epsilon^2\alpha^2/6$. We will now show that, with high probability, any element $\boldsymbol{p}$ of the cover $\mathcal{C}$ such that $\|\boldsymbol{p} - \boldsymbol{p}^\star\|_1 \geq \epsilon$, satisfies $\mathcal{L}_N(\boldsymbol{p}^\star) - \mathcal{L}_N(\boldsymbol{p}) \geq \epsilon^2\alpha^2/2$. We will use the following concentration result on likelihood ratios.

Lemma 8.13 (Proposition 7.27 of Massart (2007)). *Let D_1, D_2 be two distributions (on any domain) with positive density functions f, g respectively. For any $x \in \mathbb{R}$, it holds*

$$\mathop{\mathbf{Pr}}_{x_1, \dots, x_N \sim D_1} \left[\frac{1}{N} \sum_{n=1}^{N} \log \frac{f(x_n)}{g(x_n)} \leq (d_{\mathrm{TV}}(D_1, D_2))^2 - 2x/N \right] \leq e^{-x} .$$

Using the above lemma with $x = O(\log(|\mathcal{C}|/\delta)) = O(k \log(k/(\epsilon\delta)))$ and

$$N = \Theta(k \log(k/(\epsilon\delta))/(\alpha^2\epsilon^2)) ,$$

we obtain that, with probability at least $1 - \delta/|\mathcal{C}|$, it holds $\mathcal{L}_N(\boldsymbol{p}^\star) - \mathcal{L}_N(\boldsymbol{p}) \geq d_{\mathrm{TV}}(D_\pi, D_\pi^\star)^2 - \alpha^2\epsilon^2/2$. From the union bound, we obtain that the same is true for all vectors $\boldsymbol{p} \in \mathcal{C}$ with probability at least $1 - \delta$. We are now ready to finish the proof of the claim. Let $\boldsymbol{p} \in \widetilde{\Delta}^k$ be any probability vector such that $\|\boldsymbol{p} - \boldsymbol{p}^\star\|_1 \geq \epsilon$. Let $\bar{\boldsymbol{p}} \in \widetilde{\Delta}^k$ be the maximizer of the empirical likelihood constrained on $\widetilde{\Delta}^k$, i.e., $\bar{\boldsymbol{p}} = \arg\max_{\boldsymbol{q} \in \widetilde{\Delta}^k} \mathcal{L}_N(\boldsymbol{q})$ and let $\widetilde{\boldsymbol{p}}^\star$ be the closest vector of the cover $\mathcal{C}$ to $\boldsymbol{p}^\star$. We have

$$\mathcal{L}_N(\bar{\boldsymbol{p}}) - \mathcal{L}_N(\boldsymbol{p}) \geq \mathcal{L}_N(\widetilde{\boldsymbol{p}}^\star) - \mathcal{L}_N(\boldsymbol{p}) \geq \mathcal{L}_N(\boldsymbol{p}^\star) - \epsilon^2\alpha^2/6 - \mathcal{L}_N(\boldsymbol{p}) .$$

The first inequality holds since both $\bar{\boldsymbol{p}}$ and $\widetilde{\boldsymbol{p}}^\star$ lie in $\widetilde{\Delta}^k$. The second inequality holds since we can replace the point of the cover $\widetilde{\boldsymbol{p}}^\star \in \mathcal{C}$, with each closest point in the simplex $\boldsymbol{p}^\star$ without affecting the likelihood value by a lot. Finally, since $\boldsymbol{p}$ lies in $\widetilde{\Delta}^k$, we can replace it with a point $\boldsymbol{p}'$ in the cover with $\|\boldsymbol{p}' - \boldsymbol{p}^\star\|_1 \geq \epsilon$, and get that

$$\mathcal{L}_N(\bar{\boldsymbol{p}}) - \mathcal{L}_N(\boldsymbol{p}) \geq \mathcal{L}_N(\boldsymbol{p}^\star) - \epsilon^2\alpha^2/6 - \mathcal{L}_N(\boldsymbol{p}') - \epsilon^2\alpha^2/6\,,$$

and, since $\mathcal{L}_N(\boldsymbol{p}^\star) - \mathcal{L}_N(\boldsymbol{p}') \geq \epsilon^2\alpha^2/2$, we have that $\mathcal{L}_N(\bar{\boldsymbol{p}}) - \mathcal{L}_N(\boldsymbol{p}) = \Omega(\epsilon^2\alpha^2)\,.$ □

This concludes the proof of Proposition 8.11. □

The Proof of Theorem 8.4

In this subsection, we prove Theorem 8.4. Our goal is to simulate a statistical query oracle which takes as input a query function q with domain $\mathcal{X} \times \mathcal{Z}$ and outputs an estimate of its expectation with respect to finely labeled examples $\mathbf{E}_{(x,z)\sim D}[q(x,z)]$, using coarsely labeled examples. Recall that since we have sample access only to coarsely labeled examples $(x, S) \sim D_\pi$, we cannot directly estimate this expectation. The key idea is to perform rejection sampling on each coarse sample (x, S) with acceptance probability $q(x, j)$ for any fine label $j \in \mathcal{Z}$. Because of the rejection sampling process, this marginal distribution is not the marginal of D on the fine labels $\mathcal{Z}$, but the marginal of D on the fine labels, conditional on the accepted samples. However, the task of estimating from this marginal distribution can be still reduced to the unsupervised problem (see Proposition 8.11) of the previous section. Consider an arbitrary query function $q : \mathcal{X} \times \mathcal{Z} \to [-1, 1]$ and, without loss of generality, let $\mathcal{Z} = [k]$. Recall that D is the joint probability distribution on the finely labeled examples (x, z). We have that

$$\mathop{\mathbf{E}}_{(x,z)\sim D}[q(x,z)] = \sum_{j=1}^{k} \mathop{\mathbf{E}}_{(x,z)\sim D}\Big[q(x,j)\mathbf{1}\{z=j\}\Big] = \sum_{j=1}^{k} \mathop{\mathbf{E}}_{(x,z)\sim D}\Big[q_j(x)\mathbf{1}\{z=j\}\Big]\,. \tag{8.4}$$

Since we would like to estimate the expectation of the query $q(x,z)$ with tolerance τ, it suffices to estimate the expectation of each query $q_j(x)\mathbf{1}\{z=j\}$ with tolerance τ/k for any $j \in [k]$. Hence, it suffices to estimate expectations of the form $\mathbf{E}_{(x,z)\sim D}[f(x)\mathbf{1}\{z=j\}]$ for arbitrary functions $f : \mathcal{X} \to [0,1]$[2] and $j \in [k]$.

Let D_x denote the marginal distribution of the examples $x \in \mathcal{X}$. The algorithm performs rejection sampling. Each coarsely labeled example $(x,S) \sim D_\pi$ is accepted with probability $f(x)$, that does not depend on the coarse label S. Hence, the rejection sampling process induces a distribution D^f over finely labeled examples $(x,z) \in \mathcal{X} \times \mathcal{Z}$ with density

$$D^f(x,z) = \frac{f(x)}{\mathbf{E}_{x\sim D_x}[f(x)]} D(x,z)\,.$$

We remark that, we do not have sample access to D^f because we do not have sample access to the distribution D of the fine examples; we introduced the above notation for the purposes of the proof. Similarly, to D_x, we define D_x^f to be the marginal distribution of x conditional on its acceptance, i.e.,

$$D_x^f(x) = \frac{f(x)}{\mathbf{E}_{x\sim D_x}[f(x)]} D_x(x)\,. \tag{8.5}$$

Let D_z denote the marginal distribution of the fine labels $[k]$ and let $D_z(\cdot|x)$ be the marginal distribution conditional on the example x. We have that

$$\mathop{\mathbf{E}}_{(x,z)\sim D}\Big[f(x)\mathbf{1}\{z=j\}\Big] = \int_{\mathcal{X}} f(x)D(x,j)dx = \int_{\mathcal{X}} f(x)D_x(x)D_z(j|x)dx\,.$$

The above expectation can be equivalently written, by multiplying and dividing by D_x^f,

$$\mathop{\mathbf{E}}_{(x,z)\sim D}\Big[f(x)\mathbf{1}\{z=j\}\Big] = \int_{\mathcal{X}} \Big(\frac{f(x)D_x(x)}{D_x^f(x)}\Big)\Big(D_x^f(x)D_z(j|x)\Big)dx\,.$$

[2] Any function $f : \mathcal{X} \to [-1,1]$ can be decomposed into $f = f^+ - f^-$ with $f^+, f^- \geq 0$ and, by linearity of expectation, it suffices to work with functions f with image in $[0,1]$.

The first term in the integral is equal to $\mathbf{E}_{x\sim D_x}[f(x)]$, by substituting Equation (8.5) and, hence, is constant. The second term corresponds to the probability of observing the fine label j, given an example x, that has been accepted from the rejection sampling process. Similarly, to the marginal D_z, we define D_z^f to be the marginal distribution of the fine labels z conditional on acceptance. Hence, we can write

$$\mathop{\mathbf{E}}_{(x,z)\sim D}\left[f(x)\mathbf{1}\{z=j\}\right] = \mathop{\mathbf{E}}_{x\sim D_x}[f(x)] \cdot \mathop{\mathbf{Pr}}_{z\sim D_z^f}[z=j]\,. \tag{8.6}$$

The decomposition of the expectation of Equation (8.6) is a key step: we now only need to learn the marginal distribution of fine labels conditional on acceptance D_z^f.

Recall that our goal is to estimate the left-hand side expectation of Equation (8.6) with tolerance τ/k. We claim that it suffices to estimate each term of the right hand side product of Equation (8.6) with tolerance $\tau/(2k)$. This is implied from the following: consider an estimate $\widetilde{\mu}$ of the value $\mathbf{E}_{x\sim D_x}[f(x)]$ and an estimate $\widetilde{p}$ of the value $\mathbf{Pr}_{z\sim D_z^f}[z=j]$. Then, using Equation (8.6), we have that

$$\left|\widetilde{\mu}\cdot\widetilde{p} - \mathop{\mathbf{E}}_{(x,z)\sim D}[f(x)\mathbf{1}\{z=j\}]\right| = \left|\widetilde{\mu}\cdot\widetilde{p} - \mathop{\mathbf{E}}_{x\sim D_x}[f(x)]\cdot \mathop{\mathbf{Pr}}_{z\sim D_z^f}[z=j]\right|,$$

and, hence, by adding and subtracting the term $\widetilde{\mu}\ \mathbf{Pr}_{z\sim D_z^f}[z=j]$, using the triangle inequality and, since both $\mathbf{E}_{x\sim D_x}[f(x)]$ and $\mathbf{Pr}_{z\sim D_z^f}[z=j]$ are at most 1, we get that

$$\left|\widetilde{\mu}\cdot\widetilde{p} - \mathop{\mathbf{E}}_{(x,z)\sim D}[f(x)\mathbf{1}\{z=j\}]\right| \le \left|\widetilde{\mu} - \mathop{\mathbf{E}}_{x\sim D_x}[f(x)]\right| + \left|\widetilde{p} - \mathop{\mathbf{Pr}}_{z\sim D_z^f}[z=j]\right|.$$

We will show that $O(k^4/(\tau^3\alpha^2)\log(1/\delta))$ samples are sufficient to bound each term of the right hand side by $\tau/(2k)$, with high probability. In order to estimate the expectation $\mathbf{E}_{(x,z)\sim D}[q(x,z)]$, the algorithm applies (in parallel) the above process k times with $f = q_j$ for any $j \in [k]$ (using Equation (8.4)) using a single training set of size $N = O(k^4/(\tau^3\alpha^2)\log(1/\delta))$ drawn from the distribution D_π of coarsely labeled examples. Moreover, the running time is polynomial in the number of samples N. To conclude the proof, it suffices to show the following claims.

Claim 8.14. *There exists an algorithm that, uses* $N = \widetilde{O}(k^4/(\tau^3\alpha^2)\log(1/\delta))$ *samples from* D_π *and computes an estimate* $\widetilde{p}$*, that satisfies* $\left|\widetilde{p} - \mathbf{Pr}_{z\sim D_z^f}[z=j]\right| \leq \tau/(2k)$ *, with probability at least* $1-\delta$*.*

Proof. Recall that the distribution D_z^f is the marginal distribution of the fine labels $z \in \mathcal{Z} = [k]$, conditional that the example $x \sim D_x^f$, i.e., that the example $x \in \mathcal{X}$ has been accepted by the rejection sampling process. Hence, the distribution D_z^f is supported on $\mathcal{Z}$. We can then directly apply Proposition 8.11, using as training set the set of *accepted* coarsely labeled samples (x, S) and can compute an estimate $\widetilde{D}$, that is ϵ-close in total variation distance to D_z^f. By setting $\epsilon = \tau/(2k)$, the algorithm uses $\widetilde{O}(k^3/(\tau^2\alpha^2)\log(1/\delta))$ samples from the set of accepted samples and outputs the estimate $\widetilde{p} = \widetilde{D}(j)$. For the example $x \in \mathcal{X}$, the acceptance probability $f(x)$ can be considered $\Omega(\tau/k)$. Otherwise, we can set the desired expectation equal to 0. Hence, the algorithm needs to draw in total $\widetilde{O}(k^4/(\tau^3\alpha^2)\log(1/\delta))$ samples from D_π in order to compute an estimate $\widetilde{p}$ that satisfies

$$\left|\widetilde{p} - \mathop{\mathbf{Pr}}_{z\sim D_z^f}[z=j]\right| \leq \tau/(2k)\,,$$

with probability at least $1-\delta$. □

Claim 8.15. *There exists an algorithm that, uses* $N = O((k^2/\tau^2)\log(1/\delta))$ *samples from* D_π *and computes an estimate* $\widetilde{\mu}$*, that satisfies* $\left|\widetilde{\mu} - \mathbf{E}_{x\sim D_x}[f(x)]\right| \leq \tau/(2k)$ *, with probability at least* $1-\delta$*.*

Proof. The algorithms draws N coarsely labeled examples from D_π and computes the estimate $\widetilde{\mu} = \frac{1}{N}\sum_{i=1}^N f(x_i)$. From the Hoeffding bound, since the estimate is a sum of independent bounded random variables, we get

$$\mathbf{Pr}\left[\left|\widetilde{\mu} - \mathop{\mathbf{E}}_{x\sim D_x}[f(x)]\right| \geq \tau/(2k)\right] \leq 2\exp(-N\tau^2/(2k^2))\,.$$

Using $N = O((k^2/\tau^2)\log(1/\delta))$ samples, the algorithm estimates the desired expectation with error $\tau/(2k)$, with probability at least $1-\delta$. Note that, if $\widetilde{\mu} <$

$\tau/(2k)$, the algorithm can output 0, since the estimated value will lie in the desired tolerance interval. □

8.3 Learning Gaussians from Coarse Data

In this section, we focus on an unsupervised learning problem with coarse data. Recall that we have already solved such a problem in the discrete setting as an ingredient of our supervised learning result, see Section 8.2. In this section, we study the fundamental problem of learning a Gaussian distribution given coarse data. In Section 8.3, we show that, under general partitions, this problem is NP-hard. In Section 8.3, we show that we can efficiently estimate the Gaussian mean under convex partitions of the space.

Computational Hardness under General Partitions

In this section, we consider general partitions of the d-dimensional Euclidean space, that may contain non-convex subsets. For instance, a compact convex body and its complement define a non-convex partition of $\mathbb{R}^d$. In order to get this computational hardness result, we reduce from MAX-CUT and make use of its hardness of approximation (see Håstad (2001)). Recall that MAX-CUT can be viewed as a maximization problem, where the objective function corresponds to a particular quadratic function (associated with the Laplacian matrix of the given graph instance) and the constraints restrict the solution to lie in the Boolean hypercube (the constraints can be seen geometrically as the intersection of bands, see Figure 8.3).

We first define MAX-CUT and a variant of MAX-CUT where the optimal cut score is given as part of the input. Let $G = (V, E)$ be a graph[3] with d vertices. A *cut* is a partition of V into two subsets S and $S' = V \setminus S$ and the value of the cut (S, S') is $c(S, S') = \sum_{u,v \in E} \mathbf{1}\{u \in S, v \in S'\}$. The goal of the problem is find the maximum value cut in G, i.e., to partition the vertices into two sets so that the number of

[3] We are going to work with graphs with unit weights.

edges crossing the cut is maximized. We can define MAX-CUT as the following maximization problem for the graph $G = (V, E)$ with $|V| = d$:

$$\max \sum_{(i,j)\in E} (\boldsymbol{x}_i - \boldsymbol{x}_j)^2\,, \;\text{ subj. to }\; \boldsymbol{x}_i \in \{-1, +1\}\; \forall i \in [d]\,.$$

The objective function is the quadratic form $\boldsymbol{x}^T \boldsymbol{L}_G \boldsymbol{x}$, where $\boldsymbol{L}_G$ is the Laplacian matrix of the graph G. We may also assume that the value of the optimal cut is known and is equal to opt.[4] Before proceeding with the overview of the proof, we state a key result of Håstad (2001) about the inapproximability of MAX-CUT .

Lemma 8.16 (Inapproximability of Maximum Cut Problem Håstad (2001))**.** *It is* NP*-hard to approximate* MAX-CUT *to any factor higher than* 16/17*.*

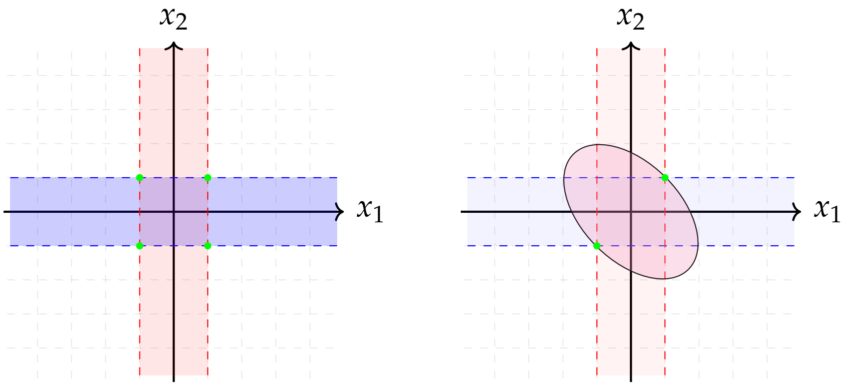


Figure 8.3: The geometry of the MAX-CUT instance. The left figure corresponds to the fat hyperplanes, i.e., the constraints of MAX-CUT and the right figure (the ellipsoid) corresponds to the objective function of MAX-CUT . The green points lie in the Boolean hypercube.

[4] Observe that this problem is still hard, since the maximum value of a cut is bounded by d^2 and, hence, if this problem could be solved efficiently, one would be able to solve MAX-CUT by trying all possible values of opt.

Sketch of the Proof of Theorem 8.7

The first step of the proof is to construct the distribution over partitions of $\mathbb{R}^d$. The Max-Cut problem can be viewed as a collection of $d+1$ non-convex partitions of the d-dimensional Euclidean space. Consider an instance of Max-Cut with $|V| = d$ and optimal cut value opt. Consider the collection of $d+1$ partitions $\mathcal{B} = \{\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_d, \mathcal{T}\}$. We define the partitions as follows: for any $i = 1, \ldots, d$, we let $S_i = \{\boldsymbol{x} : -1 \leq \boldsymbol{x}_i \leq 1\}$ be the sets that correspond to fat hyperplanes of Figure 8.3(a) and the partitions $\mathbf{\Sigma}_i = \{S_i, S_i^c\}$, i.e., pairs of fat hyperplanes and their complements (see Figure 8.4(a,b)). These d partitions will simulate the Max-Cut constraints, i.e., that the solution vector lies in the hypercube $\{-1,1\}^d$. It remains to construct $\mathcal{T}$, which intuitively corresponds to the quadratic objective of Max-Cut .

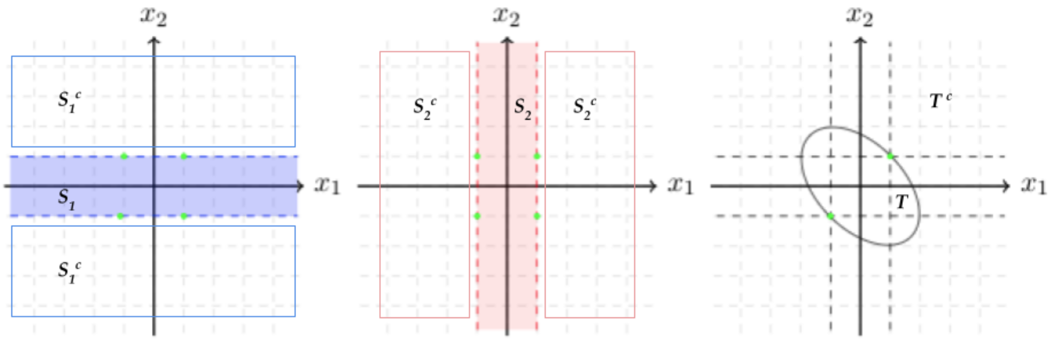


Figure 8.4: The mixture of partitions that corresponds to the Max-Cut problem. In figures (a) and (b), we partition the Euclidean space using fat hyperplanes (the blue set S_1 and the red set S_2 respectively) and their complements $S_1^c = \mathbb{R}^d \setminus S_1$ and $S_2^c = \mathbb{R}^d \setminus S_2$. The third figure (c) partitions $\mathbb{R}^d$ using the ellipsoid $T = \{\boldsymbol{x} : \boldsymbol{x}^T \mathbf{\Sigma}^{-1} \boldsymbol{x} \leq q\}$ and its complement $T^c = \mathbb{R}^d \setminus T$ (for some $d \times d$ covariance matrix $\mathbf{\Sigma}$ and positive real q).

Fix the covariance matrix $\mathbf{\Sigma} = \boldsymbol{L}_G^{-1}\text{opt}$ [5] , i.e., $\mathbf{\Sigma}$ is the inverse of the Laplacian normalized by opt. We let $T = \{\boldsymbol{x} : \boldsymbol{x}^T \mathbf{\Sigma}^{-1} \boldsymbol{x} \leq q\}$ for some positive value q to be

[5] In fact, $\boldsymbol{L}_G$ has zero eigenvalue with eigenvector $(1, \ldots, 1)$: we have to project the Laplacian to the subspace orthogonal to $(1, \ldots, 1)$ to avoid this. We ignore this technicality here for simplicity.

defined later (see Figure 8.3(b) and Figure 8.4(c)). Then, we let $\mathcal{T} = \{T, T^c\}$. We construct a mixture π of these partitions by picking each one uniformly at random, i.e., with probability $1/(d+1)$.

Let us assume that there exists an algorithm that, given access to samples from $\mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})$, with *known covariance* $\boldsymbol{\Sigma}$, computes, in time $\mathrm{poly}(d)$, a mean vector $\boldsymbol{\mu}$ so that the output distributions are matched, i.e., $d_{\mathrm{TV}}(\mathcal{N}_\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}))$ is upper bounded by $1/d^c$ for some absolute constant $c > 1$. Equivalently this means that the mass that $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ assigns to each set S_i and T is within $\mathrm{poly}(1/d)$ of the corresponding mass that $\mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})$ assigns to the same set. There are two main challenges in order to prove the reduction:

1. How can we generate coarse samples from $\mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})$ since $\boldsymbol{\mu}^\star$ is the solution of the MAX-CUT problem and therefore is unknown?

2. Given opt, is it possible to pick the threshold q of the ellipsoid $T = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \le q\}$ so that any vector $\boldsymbol{\mu}$ (rounded to belong in $\{-1, 1\}^d$), that achieves $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; T) \approx \mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; T)$ and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S_i) \approx \mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; S_i)$, also achieves an approximation ratio better than $16/17$ for the MAX-CUT objective ?

The key observation to answer the first question is that, by the rotation invariance of the Gaussian distribution, the probability $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; T) = \mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})}\left[\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \le q\right]$ is a constant p that only depends on the value opt of the MAX-CUT problem. Therefore, having this value p, we can flip a coin with this probability and give the coarse sample T if we get heads and T^c otherwise. Similarly, the value of $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; S_i)$ is an absolute constant that does not depend on $\boldsymbol{\mu}^\star \in \{-1, 1\}^d$ and therefore we can again simulate coarse samples by flipping a coin with probability equal to $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; S_i)$.

To resolve the second question, we first show that any vector $\boldsymbol{\mu}$ that approximately matches the probabilities of the d fat halfspaces, lies very close to a corner of the hypercube, see Lemma 8.20. Therefore, by rounding this guess $\boldsymbol{\mu}$, we obtain exactly a corner of the hypercube without affecting the probability assigned to the ellipsoid constraint by a lot. We then show that any vector of the hypercube that almost matches the probability of the ellipsoid achieves large cut value. In particular, we prove that there exists a value for the threshold q of the ellipsoid $\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \leq q$ that makes the probability $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; T)$ *very sensitive to changes of* $\boldsymbol{\mu}$. Therefore, the only way for the algorithm to match the observed probability is to find a $\boldsymbol{\mu}$ that achieves large cut value. We show the following lemma.

Lemma 8.17 (Sensitivity of Gaussian Probability of Ellipsoids)**.** *Let* $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})$, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *be d-dimensional Gaussian distributions. Let* $\boldsymbol{v}^\star = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\mu}^\star$, $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\mu}$ *and assume that* $\|\boldsymbol{v}\|_2 \leq \|\boldsymbol{v}^\star\|_2 = 1$*. Denote* $q = d + \|\boldsymbol{v}^\star\|_2^2 + \sqrt{2d + 4\|\boldsymbol{v}^\star\|_2^2}$*. Then, assuming d is larger than some sufficiently large absolute constant, it holds that*

$$\left| \mathop{\mathbf{Pr}}_{x \sim \mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})} \left[\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \leq q \right] - \mathop{\mathbf{Pr}}_{x \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \leq q \right] \right| \geq \frac{\|\boldsymbol{v}^\star\|_2^2 - \|\boldsymbol{v}\|_2^2}{6\sqrt{2d+4}} - o(1/\sqrt{d}) \,.$$

Notice that with $\boldsymbol{\Sigma} = \boldsymbol{L}_G^{-1} \mathrm{opt}$, in the above lemma, we have $\|\boldsymbol{v}^\star\|_2^2 = 1$, since $\boldsymbol{\mu}^\star$ achieves cut value opt. By assumption, we know that the learning algorithm can find a guess $\boldsymbol{\mu}$ that makes the left hand side of the inequality of Lemma 8.17 smaller than $\mathrm{poly}(1/d)$. Thus, we obtain that, for d large enough, it must be that $\|\boldsymbol{v}\|_2^2 = \boldsymbol{\mu}^T \boldsymbol{L}_G \boldsymbol{\mu} / \mathrm{opt} \geq 16/17$. Therefore, $\boldsymbol{\mu}$ achieves value greater than $(16/17)\mathrm{opt}$.

Remark 8.18. *The transformation π used in the above hardness result is not information preserving. In Theorem 8.7, we prove that it is computationally hard to find a vector* $\boldsymbol{\mu} \in \mathbb{R}^d$ *that matches in total variation the* observed distribution over coarse labels*. In contrast, as we will see in the upcoming Section 8.3, when the sets of the partitions are convex, we show that there is an efficient algorithm that can solve the same problem and compute some* $\boldsymbol{\mu} \in \mathbb{R}^d$ *such that* $\mathrm{TV}(\mathcal{N}_\pi(\boldsymbol{\mu}^\star), \mathcal{N}_\pi(\boldsymbol{\mu}))$ *is small regardless of whether the transformation π is information preserving. When the transformation is information preserving, we can further show that the vector* $\boldsymbol{\mu}$ *that we compute will be close to* $\boldsymbol{\mu}^\star$*.*

Sensitivity of Gaussian Probabilities

We now prove Lemma 8.17, namely that the probability of an ellipsoid with respect to the Gaussian distribution is sensitive to small changes of its mean.

Proof of Lemma 8.17. We first observe that

$$\begin{aligned}\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\le q\right] &= \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\boldsymbol{x}^T\boldsymbol{x}+2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\le q-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]\\ &= \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\boldsymbol{x}^T\boldsymbol{x}+2\boldsymbol{v}^T\boldsymbol{x}\le q-\|\boldsymbol{v}\|_2^2\right],\end{aligned}$$

where $\boldsymbol{v}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}$. Similarly, we have $\mathbf{Pr}_{x\sim\mathcal{N}(\boldsymbol{\mu}^\star,\boldsymbol{\Sigma})}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\le q\right]=\mathbf{Pr}_{x\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\boldsymbol{x}^T\boldsymbol{x}+2(\boldsymbol{v}^\star)^T\boldsymbol{x}\le q-\|\boldsymbol{v}^\star\|_2^2\right]$, where $\boldsymbol{v}^\star=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}^\star$. From the rotation invariance of the Gaussian distribution, we may assume, without loss of generality, that $\boldsymbol{v}=\|\boldsymbol{v}\|\boldsymbol{e}_1$ and $\boldsymbol{v}^\star=\|\boldsymbol{v}^\star\|\boldsymbol{e}_1$. Notice that $(\|\boldsymbol{v}\|_2+\boldsymbol{x}_1)^2+\sum_{i=2}^d\boldsymbol{x}_i^2$ is a sum of independent random variables. To estimate these probabilities we are going to use the central limit theorem.

Lemma 8.19 (CLT, Theorem 1, Chapter XVI in Feller (1957))**.** *Let $X_1,\dots,X_n$ be independent random variables with $\mathbf{E}[|X_i|^3]<+\infty$ for all i. Let $m_1=\mathbf{E}[\sum_{i=1}^n X_i]$ and $m_j=\sum_{i=1}^n\mathbf{E}[(X_i-\mathbf{E}[X_i])^j]$. Then,*

$$\mathbf{Pr}\left[\frac{(\sum_{i=1}^n X_i)-m_1}{\sqrt{m_2}}\le x\right]-\Phi(x)=m_3\frac{(1-x^2)\phi(x)}{6m_2^{3/2}}+o\left(n/m_2^{3/2}\right),$$

where $\Phi(\cdot)$, resp., $\phi(\cdot)$ is the CDF resp., PDF of the standard normal distribution and the convergence is uniform for all $x\in\mathbb{R}$.

Using the above central limit theorem we obtain

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[(\|\boldsymbol{v}^\star\|_2+\boldsymbol{x}_1)^2+\sum_{i=2}^d\boldsymbol{x}_i^2\le q\right]=\Phi(\bar{q}_1)+O\left(\frac{1}{\sqrt{d}}\right)(1-\bar{q}_1{}^2)\phi(\bar{q}_1)+o\left(1/\sqrt{d}\right),$$

where $\bar{q}_1 = \frac{q-(d+\|v^\star\|_2^2)}{\sqrt{2d+4\|v^\star\|_2^2}}$. Since $q = d + \|v^\star\|^2 + \sqrt{2d + 4\|v^\star\|_2^2}$ we obtain $\bar{q}_1 = 1$ and therefore

$$\Pr_{x\sim\mathcal{N}(\mathbf{0},I)} \left[x^T x + 2(v^\star)^T x \le q - \|v^\star\|_2^2\right] = \Phi(1) + o\left(1/\sqrt{d}\right) .$$

Similarly, from the central limit theorem, we obtain

$$\Pr_{x\sim\mathcal{N}(\mathbf{0},I)} \left[(\|v\|_2 + x_1)^2 + \sum_{i=2}^{d} x_i^2 \le q\right] = \Phi(\bar{q}_2) + O\left(\frac{1}{\sqrt{d}}\right)(1-\bar{q}_2{}^2)\phi(\bar{q}_2) + o\left(1/\sqrt{d}\right) ,$$

where $\bar{q}_2 = \frac{q-(d+\|v\|_2^2)}{\sqrt{2d+4\|v\|_2^2}} = 1 + O(1/\sqrt{d})$. Therefore, we have

$$\Pr_{x\sim\mathcal{N}(\mathbf{0},I)} \left[x^T x + 2v^T x \le q - \|v\|_2^2\right] = \Phi(\bar{q}_2) + o\left(1/\sqrt{d}\right) .$$

Moreover, we have that $\bar{q}_2 \ge 1 + (\|v^\star\|_2^2 - \|v\|_2^2)/(\sqrt{2d + 4\|v\|_2^2})$. Using the fact that d is sufficiently large and standard approximation results on the Gaussian CDF, we obtain

$$\Phi\left(1 + \frac{\|v^\star\|_2^2 - \|v\|_2^2}{\sqrt{2d + 4\|v\|_2^2}}\right) - \Phi(1) \ge (\|v^\star\|_2^2 - \|v\|_2^2)/\left(6\sqrt{2d + 4\|v\|_2^2}\right) ,$$

and, since $\|v\|_2 \le 1$, we conclude that the left-hand side satisfies

$$\Phi\left(1 + \frac{\|v^\star\|_2^2 - \|v\|_2^2}{\sqrt{2d + 4\|v\|_2^2}}\right) - \Phi(1) \ge (\|v^\star\|_2^2 - \|v\|_2^2)/\left(6\sqrt{2d + 4}\right) .$$

The result follows. □

We will also require the following sensitivity lemma about the Gaussian probability of bands, i.e., sets of the form $\{x : |x_i| \le 1\}$. We show that the probabilities of such regions are also sensitive under perturbations of the mean of the Gaussian. This means that any vector $\boldsymbol{\mu}$ that has $\mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[-1 \leq \boldsymbol{x}_i \leq 1\right]$ close to $\mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^\star,\boldsymbol{\Sigma})}\left[-1 \leq \boldsymbol{x}_i \leq 1\right]$ must be very close to a corner of the hypercube.

Lemma 8.20 (Sensitivity of Gaussian Probability of Bands)**.** *Let $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be two d-dimensional Gaussian distributions with $\boldsymbol{e}_i^T \boldsymbol{\Sigma} \boldsymbol{e}_i \leq Q$, and $|\boldsymbol{\mu}_i^\star| = 1$ for all $i \in [d]$. Then, for any $i \in [d]$, it holds that*

$$\left| \underset{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^\star,\boldsymbol{\Sigma})}{\mathbf{Pr}}\left[-1 \leq \boldsymbol{x}_i \leq 1\right] - \underset{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\mathbf{Pr}}\left[-1 \leq \boldsymbol{x}_i \leq 1\right] \right| \geq c \cdot \frac{\min(1, (1 - |\boldsymbol{\mu}_i|)^2)}{Q^4},$$

for some absolute constant $c \in (0, 1]$.

Proof. Let us fix $i \in [d]$, define $\mu^\star$ (resp. μ) for $\boldsymbol{\mu}_i^\star$ (resp. $\boldsymbol{\mu}_i$), and $\sigma^2 = \boldsymbol{\Sigma}_{ii}$. Without loss of generality since both Gaussians have the same variance σ by symmetry we may assume that $\mu^\star = 1$ and $\mu \in [0, +\infty)$. We first deal with the case $\mu > 1$. We have

$$\begin{aligned} &\underset{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^\star,\boldsymbol{\Sigma})}{\mathbf{Pr}}\left[-1 \leq \boldsymbol{x}_i \leq 1\right] - \underset{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\mathbf{Pr}}\left[-1 \leq \boldsymbol{x}_i \leq 1\right] \\ &\qquad\qquad = \underset{t\sim\mathcal{N}(1,\sigma^2)}{\mathbf{E}}\left[\mathbf{1}\{|t| \leq 1\}\left(1 - \frac{\mathcal{N}(\mu,\sigma^2;t)}{\mathcal{N}(1,\sigma^2;t)}\right)\right]. \end{aligned}$$

We have that since $\mu > 1$ the ratio $\frac{\mathcal{N}(\mu,\sigma^2;t)}{\mathcal{N}(1,\sigma^2;t)} = e^{\frac{(\mu-1)(-\mu+2t-1)}{2\sigma^2}}$ is maximized for $t = 1$ and has maximum value $e^{-\frac{(\mu-1)^2}{2\sigma^2}}$. By taking the derivative with respect to σ we observe that the probability that $N(1, \sigma)$ assigns to $[-1, 1]$ is decreasing with respect to σ and therefore it is minimized for $\sigma = 1$. We have that $\mathbf{Pr}_{t\sim\mathcal{N}(1,\sigma)}[-1 < t < 1] = \Omega(1/\sigma)$ and therefore $\mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^\star,\boldsymbol{\Sigma})}\left[-1 \leq \boldsymbol{x}_i \leq 1\right] - \mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[-1 \leq \boldsymbol{x}_i \leq 1\right] \geq C \cdot \left(1 - e^{-\frac{(\mu-1)^2}{2\sigma^2}}\right)$. We can obtain the significantly weaker lower bound of $c \min(1, (1 - |\mu|)^2)$ for some absolute constant $c \in (0, 1]$ by using the inequality $1 - e^{-x} \geq 1/2 \min(1, x)$ that holds for all $x \in [0, +\infty)$.

We now deal with the case $\mu \in [0, 1)$. In that case the expression of their ratio of the densities of $\mathcal{N}(1, \sigma)$ and $\mathcal{N}(\mu, \sigma)$ derived above shows us that they cross at $t = (1+\mu)/2$. Therefore, they completely cancel out in the interval $[\mu, 1]$. We have $\mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[-1 \leq \boldsymbol{x}_i \leq 1] - \mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^\star,\boldsymbol{\Sigma})}[-1 \leq \boldsymbol{x}_i \leq 1] = \mathbf{Pr}_{t\sim\mathcal{N}(\mu,\sigma)}[-1 \leq t \leq \mu] - \mathbf{Pr}_{t\sim\mathcal{N}(1,\sigma)}[-1 \leq t \leq \mu] = \Omega((1-\mu)/(1+\sigma^4))$, where to obtain the last inequality we use standard approximations of Gaussian integrals. Combining the above two cases we obtain the claimed lower bound.

□

The Proof of Theorem 8.7

We are now ready to provide the complete proof of Theorem 8.7. Consider an instance of MAX-CUT with $|V| = d$ and optimal value $\mathrm{opt} = O(d^2)$. Let $\boldsymbol{L}_G$ be the Laplacian matrix of the (connected) graph G. Since the minimum eigenvalue of $\boldsymbol{L}_G$ is 0, we project the matrix onto the subspace V that is orthogonal to $\mathbf{1} = (1, \ldots, 1)$. We introduce a $(d-1) \times d$ partial isometry $\boldsymbol{R}$, that satisfies $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$ and $\boldsymbol{R}\mathbf{1} = \mathbf{0}$, i.e., $\boldsymbol{R}$ projects vectors to the subspace V. We consider $\boldsymbol{L}_G' = \boldsymbol{R}\boldsymbol{L}_G\boldsymbol{R}^T$. It suffices to find a solution $\boldsymbol{x} \in V$ and then project back to $\mathbb{R}^d$: $\boldsymbol{y} = \boldsymbol{R}^T\boldsymbol{x}$. We note that the matrix $\boldsymbol{L}_G'$ is positive definite (the smallest eigenvalue of $\boldsymbol{L}_G'$ is equal to the second smallest eigenvalue of $\boldsymbol{L}_G$) and preserves the optimal score value, in the sense that

$$\mathrm{opt} = \max_{\boldsymbol{y}\in\mathbb{R}^d} \boldsymbol{y}^T\boldsymbol{L}_G\boldsymbol{y} = \max_{\boldsymbol{x}\in\mathbb{R}^d}(\boldsymbol{R}^T\boldsymbol{x})^T\boldsymbol{L}_G(\boldsymbol{R}^T\boldsymbol{x}) = \max_{\boldsymbol{x}\in V}\boldsymbol{x}^T\boldsymbol{L}_G'\boldsymbol{x}\,.$$

Assume that there exists an efficient black-box algorithm $\mathcal{A}$, that, given sample access to a generative process of coarse Gaussian data $\mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})$ with known covariance [6] matrix $\boldsymbol{\Sigma}$, computes an estimate $\widetilde{\boldsymbol{\mu}}$ in $\mathrm{poly}(d)$ time, that satisfies

$$d_{\mathrm{TV}}(\mathcal{N}_\pi(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}), \mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})) < 1/d^c\,.$$

We choose the known covariance matrix to be equal to $\boldsymbol{\Sigma} = (\boldsymbol{L}_G')^{-1}\mathrm{opt}$, where opt is the given optimal MAX-CUT value and let $\boldsymbol{\mu}^\star \in \{-1, 1\}^{d-1}$ be the unknown mean

[6] We remark that our hardness result is stated for identity covariance matrix (and not for an arbitrary known covariance matrix). In order to handle this case, we provide a detailed discussion after the end of the proof of Theorem 8.7.

vector. Recall that, not only the black-box algorithm $\mathcal{A}$, but also the generative process that we design is agnostic to the true mean. However, as we will see the knowledge of the optimal value opt and the fact that the true mean lies in the hypercube $\{-1,1\}^{d-1}$ suffice to generate samples from the true coarse generative process $\mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})$.

In what follows, we will construct such a coarse generative process using the objective function and the constraints of the MAX-CUT problem. Specifically, we will design a collection $\mathcal{B} = \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_{d-1}, \mathcal{T}\}$ of d partitions of the d-dimensional Euclidean space and let the partition distribution π be the uniform probability measure over $\mathcal{B}$.

We define the partitions as follows: for any $i = 1, \dots, d-1$, let $S_i = \{\boldsymbol{x} : -1 \leq \boldsymbol{x}_i \leq 1\}$ and $\boldsymbol{\Sigma}_i = \{S_i, S_i^c\}$. These $d-1$ partitions simulate the integrality constraints of MAX-CUT , i.e., the solution vector should lie in the hypercube $\{-1,1\}^{d-1}$. It remains to construct $\mathcal{T}$, which corresponds to the quadratic objective of MAX-CUT . We let $T = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \leq q\}$, for $q > 0$ to be decided. Then, we let $\mathcal{T} = \{T, T^c\}$. Recall that the known covariance matrix $\boldsymbol{\Sigma} = (\boldsymbol{L}_G')^{-1}\text{opt}$ lies in $\mathbb{R}^{(d-1)\times(d-1)}$ and, so, we will use $d-1$ bands (i.e., fat hyperplanes).

The main question to resolve is how to generate efficiently samples from the designed general partition, i.e., the distribution $\mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})$, *without* knowing the value of $\boldsymbol{\mu}^\star$. The key observation is that, by the rotation invariance of the Gaussian distribution, the probability $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; T) = \mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^\star,\boldsymbol{\Sigma})}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \leq q\right]$ is a constant p that only depends on the value opt of the maximum cut (see the proof of Lemma 8.17). Therefore, having this value p, we can flip a coin with this probability and give the coarse sample T if we get heads and T^c otherwise. At the same time, the value of $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; S_i)$ is an absolute constant that does not depend on $\boldsymbol{\mu}^\star \in \{-1,1\}^{d-1}$ and, therefore, we can again simulate coarse samples by flipping a coin with probability equal to $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; S_i)$. More precisely, since S_i is a symmetric interval around 0, we have that

$$\underset{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^\star,\boldsymbol{\Sigma})}{\mathbf{Pr}}\left[-1 \leq \boldsymbol{x}_i \leq 1\right] = \underset{t\sim\mathcal{N}(1,\boldsymbol{\Sigma}_{ii})}{\mathbf{Pr}}\left[-1 \leq t \leq 1\right].$$

Notice that the above constant only depends on the *known* constant $\boldsymbol{\Sigma}_{ii}$ and can be computed to very high accuracy using well known approximations of the Gaussian integral or rejection sampling. Moreover, all the probabilities $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; S_i), \mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; T)$ are at least polynomially small in $1/d$. In particular, $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; S_i)$, is always larger than $\Omega(1/\sigma) \geq \mathrm{poly}(1/d)$ and smaller than $1/2$ and $\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; T) = \Phi(1) + o(1/\sqrt{d})$ [7], see the proof of Lemma 8.17. Having these values we can generate samples from $\mathcal{N}_\pi$ as follows:

1. Pick one of the d sets $S_1, \dots, S_{d-1}, T$ uniformly at random.
2. Flip a coin with success probability equal to the probability of the corresponding sets and return either the set or its complement.

Giving sample access to the designed oracle with $\mathcal{B} = \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_{d-1}, \mathcal{T}\}$, the black-box algorithm $\mathcal{A}$ computes efficiently and returns an estimate $\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^{d-1}$, that satisfies

$$d_{\mathrm{TV}}(\mathcal{N}_\pi(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}), \mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})) < o(1/d^c)\,.$$

We proceed with two claims: (i) the algorithm's output $\widetilde{\boldsymbol{\mu}}$ should lie in a ball of radius $\mathrm{poly}(1/d)$, centered at one of the vertices of the hypercube $\{-1,1\}^{d-1}$ and (ii) it will hold that the rounded vector $\widehat{\boldsymbol{\mu}} = (\mathrm{sgn}(\widetilde{\boldsymbol{\mu}}_i))_{1 \leq i \leq d-1} \in \{-1,1\}^{d-1}$ will attain a cut score, that approximates the MAX-CUT within a factor larger than $16/17$. By the algorithm's guarantee, since π is the uniform distribution, we get that

$$|\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}; T) - \mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; T)| + \sum_{i=1}^{d-1} |\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}; S_i) - \mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; S_i)| = o(1/d^{c-1})\,.$$

Hence, we get that each of the above d summands is at most $o(1/d^{c-1})$.

Claim 8.21. *It holds that $\|\widetilde{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}\|_\infty < \epsilon$, where $\widetilde{\boldsymbol{\mu}}$ is the black-box algorithm's estimate and $\widehat{\boldsymbol{\mu}}$ its rounding to $\{-1,1\}^{d-1}$.*

[7] $\Phi(\cdot)$ is the CDF of the standard Normal distribution.

Proof. For any coordinate $i \in [d-1]$, we will apply Lemma 8.20 in order to bound the distance between the estimated guess and the true, based on the Gaussian mass gap in each one of the $d-1$ bands.

Note that $|\boldsymbol{\mu}_i^\star| = 1$ for all $i \in [d-1]$. Also, note that the $(d-1) \times (d-1)$ matrix $\boldsymbol{L}'_G$ is positive definite and the minimum eigenvalue $\lambda(\boldsymbol{L}'_G)$ is equal to the second smallest eigenvalue of the $d \times d$ Laplacian matrix $\boldsymbol{L}_G$. It holds that $\lambda(\boldsymbol{L}'_G) > 0$. Hence, the maximum entry of the covariance matrix $\boldsymbol{\Sigma} = (\boldsymbol{L}'_G)^{-1}\mathrm{opt}$ is upper bounded by $1/(\mathrm{opt} \cdot \lambda(\boldsymbol{L}'_G)) < Q = \mathrm{poly}(d)$ for some value Q. Using Lemma 8.20 and the algorithm's guarantee, we have that

$$(|\widetilde{\boldsymbol{\mu}}_i| - 1)^2/Q^4 \leq |\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}; S_i) - \mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; S_i)| = o\left(1/d^{c-1}\right) .$$

For sufficiently large c, we get that each coordinate of the estimated vector $\widetilde{\boldsymbol{\mu}}$ lies in an interval, centered at either 1 or -1 of length $o(1/d^{c-1})$. This implies that $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{w}\|_\infty < \epsilon$ for some $\epsilon = o(1/d^{c-1})$ and some vertex $\boldsymbol{w}$ of the hypercube $\{-1, 1\}^{d-1}$. Hence, we have that $\widetilde{\boldsymbol{\mu}}$ should lie in a ball, with respect to the L_∞ norm, centered at one of the vertices of the $(d-1)$-hypercube with radius of order ϵ and note that this vertex corresponds to the rounded vector $\widehat{\boldsymbol{\mu}}$ of the estimated vector. □

We continue by claiming that the rounded vector $\widehat{\boldsymbol{\mu}}$ attains a MAX-CUT value, that approximates the optimal value opt withing a factor strictly larger than 16/17.

Claim 8.22. *The* MAX-CUT *value of the rounded vector* $\widehat{\boldsymbol{\mu}} \in \{-1, 1\}^{d-1}$ *satisfies*

$$\widehat{\boldsymbol{\mu}}^T \boldsymbol{L}'_G \widehat{\boldsymbol{\mu}} > (16/17) \cdot \mathrm{opt} .$$

Proof. We will make use of Lemma 8.17, in order to get the desired result via the Gaussian mass gap between the two means on the designed ellipsoid. In order to apply this Lemma, note that, for the true mean $\boldsymbol{\mu}^\star$, we have that $\|\boldsymbol{v}^\star\|_2^2 = \|(\boldsymbol{\Sigma}^\star)^{-1/2}\boldsymbol{\mu}^\star\|_2^2 = ((\boldsymbol{\mu}^\star)^T \boldsymbol{L}'_G \boldsymbol{\mu}^\star)/\mathrm{opt} = 1$, since the true mean attains the optimal MAX-CUT score. Similarly, for the rounded estimated mean $\widehat{\boldsymbol{\mu}}$, the associated vector $\widehat{\boldsymbol{v}}$ satisfies $\|\widehat{\boldsymbol{v}}\|_2 \leq 1$, since its cut value is at most opt. So, we can apply Lemma 8.17

with $\boldsymbol{v}^\star = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}^\star$ and $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\mu}}$ and get that

$$\frac{1 - \left(\widehat{\boldsymbol{\mu}}^T \boldsymbol{L}'_G \widehat{\boldsymbol{\mu}}\right)/\mathrm{opt}}{6\sqrt{2d+4}} - o\left(1/\sqrt{d}\right) < o\left(1/d^{c-1}\right) ,$$

which implies that, for some small constant c', the value of the estimated mean satisfies $\widehat{\boldsymbol{\mu}}^T \boldsymbol{L}'_G \widehat{\boldsymbol{\mu}} > (1 - c' - 1/d^{c-1})\mathrm{opt}$. This implies that the algorithm $\mathcal{A}$ can approximate the MAX-CUT value within a factor higher than 16/17. □

Known Covariance vs. Identity Covariance. Recall that our hardness result (Theorem 8.7) states that there is no algorithm with sample access to $\mathcal{N}_\pi(\boldsymbol{\mu}^\star) = \mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{I})$, that can compute a mean $\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^d$ in $\mathrm{poly}(d)$ time such that $d_{\mathrm{TV}}(\mathcal{N}_\pi(\widetilde{\boldsymbol{\mu}}), \mathcal{N}_\pi(\boldsymbol{\mu}^\star)) < 1/d^c$ for some absolute constant $c > 1$. In order to prove our hardness result, we assume that there exists such a black-box algorithm $\mathcal{A}$. Hence, to make use of $\mathcal{A}$, one should provide samples generated by a coarse Gaussian with *identity* covariance matrix. However, in our reduction, we show that we can generate samples from a coarse Gaussian (which is associated with the MAX-CUT instance) that has *known* covariance matrix $\boldsymbol{\Sigma}$. Let us consider a sample $S \sim \mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})$. Since $\boldsymbol{\Sigma}$ is known, we can rotate the sets and give as input to the algorithm $\mathcal{A}$ the set

$$\boldsymbol{\Sigma}^{-1/2} \cdot S := \left\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} : \boldsymbol{x} \in S\right\} ,$$

i.e., we can implement the membership oracle $\mathcal{O}_{\boldsymbol{\Sigma}^{-1/2}\cdot S}(\cdot)$, assuming oracle access to $\mathcal{O}_S(\cdot)$. We have that $\mathcal{O}_{\boldsymbol{\Sigma}^{-1/2}\cdot S}(\boldsymbol{x}) = \mathcal{O}_S(\boldsymbol{\Sigma}^{1/2}\boldsymbol{x})$. We continue with a couple of observations.

1. We first observe that, for any partition $\mathcal{S}$ of the d-dimensional Euclidean space, there exists another partition $\boldsymbol{\Sigma}^{-1/2} \cdot \mathcal{S}$ consisting of the sets $\boldsymbol{\Sigma}^{-1/2} \cdot S$, where $S \in \mathcal{S}$. Note that since $\boldsymbol{\Sigma}^{-1/2}$ is full rank, the mapping $\boldsymbol{x} \mapsto \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ is a bijection and so $\boldsymbol{\Sigma}^{-1/2} \cdot \mathcal{S}$ is a partition of the space with $\pi(\boldsymbol{\Sigma}^{-1/2} \cdot \mathcal{S}) = \pi(\mathcal{S})$.

2. We have that $\boldsymbol{x} \in S$ if and only if $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} \in \boldsymbol{\Sigma}^{-1/2} \cdot S$ and so

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathbf{1}\{\boldsymbol{x} \in S\}] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathbf{1}\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} \in \boldsymbol{\Sigma}^{-1/2} \cdot S\}] .$$

   Since it holds that $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if and only if $\boldsymbol{w} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z} + \boldsymbol{\mu}$ with $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, we get for an arbitrary subset $S \subseteq \mathbb{R}^d$ that

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathbf{1}\{\boldsymbol{x} \in S\}] &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\mathbf{1}\left\{\boldsymbol{\Sigma}^{-1/2}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{x} + \boldsymbol{\mu}\right) \in \boldsymbol{\Sigma}^{-1/2} \cdot S\right\}\right] \\ &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu},\boldsymbol{I})}\left[\mathbf{1}\{\boldsymbol{x} \in \boldsymbol{\Sigma}^{-1/2} \cdot S\}\right] .\end{aligned}$$

Let us consider a set $S \subseteq \mathbb{R}^d$ distributed as $\mathcal{N}_\pi(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma})$. This set is the one that the algorithm with the known covariance matrix works with. We are now ready to combine the above two observations in order to understand what is the input to the identity covariance matrix algorithm. We have that

$$\begin{aligned}\mathop{\mathbf{Pr}}_{S\sim\mathcal{N}_\pi(\boldsymbol{\mu}^\star,\boldsymbol{\Sigma})}[S] &= \sum_{\mathcal{S}} \mathbf{1}\{S \in \mathcal{S}\}\pi(\mathcal{S})\mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}; S) \\ &= \sum_{\mathcal{S}} \mathbf{1}\{S \in \mathcal{S}\}\pi(\mathcal{S})\mathcal{N}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}^\star, \boldsymbol{I}; \boldsymbol{\Sigma}^{-1/2} \cdot S) \\ &= \sum_{\boldsymbol{\Sigma}^{-1/2}\cdot\mathcal{S}} \mathbf{1}\{\boldsymbol{\Sigma}^{-1/2} \cdot S \in \boldsymbol{\Sigma}^{-1/2} \cdot \mathcal{S}\}\pi(\boldsymbol{\Sigma}^{-1/2} \cdot \mathcal{S})\mathcal{N}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}^\star, \boldsymbol{I}; \boldsymbol{\Sigma}^{-1/2} \cdot S) \\ &= \mathop{\mathbf{Pr}}_{S'\sim\mathcal{N}_{\pi'}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}^\star,\boldsymbol{I})}[S'] ,\end{aligned}$$

where the set S' is distributed as $\mathcal{N}_{\pi'}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}^\star, \boldsymbol{I})$ where π' is the 'rotated' partition distribution supported on the rotated partitions $\boldsymbol{\Sigma}^{-1/2} \cdot \mathcal{S}$ for each $\mathcal{S}$ with $\pi(\mathcal{S}) > 0$. We remark that the second equation follows from the second observation and the third equation from the first one. Hence, the algorithm $\mathcal{A}$ (the one that works with identity matrix) obtains the rotated sets (i.e., membership oracles) $\boldsymbol{\Sigma}^{-1/2} \cdot S$ and the (unknown) target mean vector is $\boldsymbol{u} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}^\star$.

Efficient Mean Estimation under Convex Partitions

In this section, we formally state and prove Theorem 8.8 which is stated in Section 8.1: we provide an efficient algorithm for Gaussian mean estimation under *convex* partitions. The following definition of information preservation is very similar with the one given in the introduction, see Definition 8.3. The difference is that we only require from π to preserve the distances of Gaussians around the true Gaussian $\mathcal{N}(\boldsymbol{\mu}^\star)$ as opposed to the distance of any pair of Gaussians $\mathcal{N}(\boldsymbol{\mu}^\star)$: this is a somewhat more flexible assumption about the partition distribution π and the true Gaussian $\mathcal{N}(\boldsymbol{\mu}^*)$ as a pair.

Definition 8.23 (Information Preserving Partition Distribution for Gaussians)**.** *Let $\alpha \in [0, 1]$ and consider a d-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^\star)$. We say that π is an α-information preserving partition distribution with respect to the true Gaussian $\mathcal{N}(\boldsymbol{\mu}^\star)$ if for any Gaussian distribution $\mathcal{N}(\boldsymbol{\mu})$, it holds that $d_{\mathrm{TV}}(\mathcal{N}_\pi(\boldsymbol{\mu}), \mathcal{N}_\pi(\boldsymbol{\mu}^\star)) \geq \alpha \cdot d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}), \mathcal{N}(\boldsymbol{\mu}^\star))$.*

We refer to Appendix G.3 for a geometric condition, under which a partition is α-information preserving. In particular, we prove that a partition is α-information preserving if, for any hyperplane, it holds that the mass of the cells of the partition that do not intersect with the hyperplane is at least α. This is true for most natural partitions, see e.g., the Voronoi diagram of Figure 8.2. In this section, we discuss and establish the two structural lemmata required in order to prove Theorem 8.8. Our goal is to maximize the empirical log-likelihood objective

$$\mathcal{L}_N(\boldsymbol{\mu}) = \frac{1}{N} \sum_{i=1}^{N} \log \mathcal{N}(\boldsymbol{\mu}; S_i) \,, \tag{8.7}$$

where the N (convex) sets $S_1, \ldots, S_N$ are drawn from the coarse Gaussian generative process $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$. We first show that the above empirical likelihood is a concave objective with respect to $\boldsymbol{\mu} \in \mathbb{R}^d$. In the following lemma, we show that the log-probability of a convex set S, i.e., the function $\log \mathcal{N}(\boldsymbol{\mu}; S)$ is a concave function of the mean $\boldsymbol{\mu}$.

Lemma 8.24 (Concavity of Log-Likelihood)**.** *Let $S \subseteq \mathbb{R}^d$ be a convex set. The function $\log \mathcal{N}(\boldsymbol{\mu}; S)$ is concave with respect to the mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$.*

In order to prove that the Hessian matrix of this objective is negative semi-definite, we use a variant of the Brascamp-Lieb inequality. Having established the concavity of the empirical log-likelihood, we next have to bound the sample complexity of the empirical log-likelihood. We prove the following lemma.

Lemma 8.25 (Sample Complexity of Empirical Log-Likelihood)**.** *Let $\epsilon, \delta \in (0,1)$ and consider a generative process for coarse d-dimensional Gaussian data $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$ (see Definition 8.5). Also, assume that every $\boldsymbol{\Sigma} \in \operatorname{supp}(\pi)$ is a convex partition of the Euclidean space. Let $N = \widetilde{\Omega}(d/(\epsilon^2\alpha^2)\log(1/\delta))$. Consider the empirical log-likelihood objective*

$$\mathcal{L}_N(\boldsymbol{\mu}) = \frac{1}{N}\sum_{i=1}^{N} \log \mathcal{N}(\boldsymbol{\mu}; S_i)\,.$$

Then, with probability at least $1-\delta$, we have that, for any Gaussian distribution $\mathcal{N}(\boldsymbol{\mu})$ that satisfies $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}), \mathcal{N}(\boldsymbol{\mu}^\star)) \geq \epsilon$, it holds that $\max_{\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^d} \mathcal{L}_N(\widetilde{\boldsymbol{\mu}}) - \mathcal{L}_N(\boldsymbol{\mu}) \geq \Omega(\epsilon^2\alpha^2)$.

The above lemma states that, given roughly $\widetilde{O}(d/(\epsilon^2\alpha^2))$ samples from $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$, we can guarantee that the maximizer $\widetilde{\boldsymbol{\mu}}$ of the empirical log-likelihood achieves a total variation gap at most ϵ against the true mean vector $\boldsymbol{\mu}^\star$, i.e., $d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}}), \mathcal{N}(\boldsymbol{\mu}^\star)) \leq \epsilon$. In fact, thanks to the concavity of the empirical log-likelihood objective, it suffices to show that Gaussian distributions $\mathcal{N}(\boldsymbol{\mu})$, that satisfy $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}), \mathcal{N}(\boldsymbol{\mu}^\star)) > \epsilon$, will also be significantly sub-optimal solutions of the empirical log-likelihood maximization. The key idea in order to attain the desired sample complexity, is that is suffices to focus on guess vectors $\boldsymbol{\mu}$ that lie in a sphere of radius $\Omega(\epsilon)$. Technically, the proof of Lemma 8.25 relies on a concentration result of likelihood ratios and in the observation that, while the empirical log-likelihood objective $\mathcal{L}_N$ is concave (under convex partitions), the regularized objective $\mathcal{L}_N(\boldsymbol{\mu}) + \|\boldsymbol{\mu}\|_2^2$ is convex with respect to the guess mean vector $\boldsymbol{\mu}$.

Concavity of Log-likelihood: Proof of Lemma 8.24

In this section, we show that the log-likelihood is concave when the underlying partitions are convex. The Hessian of the log-likelihood $\mathcal{L}$ for the set S has a notable property. When restricted to a direction $\boldsymbol{v} \in \mathbb{R}^d$, the quadratic $\boldsymbol{v}^T(\nabla^2 \mathcal{L})\boldsymbol{v}$ quantifies the variance reduction, observed between the distributions $\mathcal{N}_S$ (Gaussian conditioned on S) and $\mathcal{N}$ (unrestricted Gaussian, i.e., $S = \mathbb{R}^d$). When the set S is convex (and, hence the indicator function $\mathbf{1}_S$ is log-concave), the variance of the unrestricted Gaussian is always larger than the conditional one. This intriguing result is an application of a variation of the Brascamp-Lieb inequality, due to Hargé (see Lemma 8.26 for the inequality that we utilize). Recall that, both the empirical and the population log-likelihood objectives are convex combinations of the function $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) = \log \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S)$ and, hence, it suffices to show that f is concave with respect to $\boldsymbol{\mu} \in \mathbb{R}^d$, when the set S is convex.

Proof of Lemma 8.24. Without loss of generality, we can take $\boldsymbol{\Sigma} = \boldsymbol{I} \in \mathbb{R}^{d\times d}$. Let $f(\boldsymbol{\mu}; S) = \log \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}; S)$ for an arbitrary convex set $S \subseteq \mathbb{R}^d$. The gradient $\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu})$ of f with respect to $\boldsymbol{\mu}$ is equal to

$$\begin{aligned}
&\nabla_{\mu}\left(\log \int_S \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right) d\boldsymbol{x}\right) \\
&= \frac{\int_S \boldsymbol{x} \exp(-(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})/2) d\boldsymbol{x}}{\int_S \exp(-(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})/2) d\boldsymbol{x}} - \boldsymbol{\mu}\,.
\end{aligned}$$

Hence, we get that

$$\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_S(\boldsymbol{\mu},\boldsymbol{I})}[x] - \boldsymbol{\mu}\,.$$

We continue with the computation of the Hessian of the function f with respect to

$\boldsymbol{\mu}$

$$\nabla^2_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) = -\boldsymbol{I} + \frac{\int_S \boldsymbol{x}(\boldsymbol{x}-\boldsymbol{\mu})^T \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}; \boldsymbol{x}) d\boldsymbol{x}}{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}; S)} - \frac{\left(\int_S \boldsymbol{x} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}; \boldsymbol{x}) d\boldsymbol{x} \right) \left(\int_S (\boldsymbol{x}-\boldsymbol{\mu})^T \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}; \boldsymbol{x}) d\boldsymbol{x} \right)}{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}; S)^2},$$

and, so, we have that

$$\nabla^2_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) = -\boldsymbol{I} + \Big(\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{I})} [\boldsymbol{x}\boldsymbol{x}^T] - \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{I})} [\boldsymbol{x}] \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{I})} [\boldsymbol{x}^T] \Big) = \mathop{\mathbf{Cov}}_{\boldsymbol{x} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{I})} [\boldsymbol{x}] - \boldsymbol{I}\,.$$

Observe that, when $S = \mathbb{R}^d$, we get that both the gradient and the Hessian vanish. In order to show the concavity of f with respect to the mean vector $\boldsymbol{\mu}$, consider an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^d$ in the ball $\|\boldsymbol{v}\|_2 = 1$. We have the quadratic form

$$\boldsymbol{v}^T \nabla^2_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) \boldsymbol{v} = \boldsymbol{v}^T \mathop{\mathbf{Cov}}_{\boldsymbol{x} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{I})} [\boldsymbol{x}] \boldsymbol{v} - 1 = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{I})} \left[(\boldsymbol{v}^T \boldsymbol{x})^2 \right] - \Big(\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{I})} [\boldsymbol{v}^T \boldsymbol{x}] \Big)^2 - 1\,.$$

In order to show the desired inequality, we will apply the following variant of the Brascamp-Lieb inequality.

Lemma 8.26 (Brascamp-Lieb Inequality, Hargé (see Guionnet (2009))). *Let g be convex function on $\mathbb{R}^d$ and let S be a convex set on $\mathbb{R}^d$. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the Gaussian distribution on $\mathbb{R}^d$. It holds that*

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_S} \left[g\left(\boldsymbol{x} + \boldsymbol{\mu} - \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_S} [\boldsymbol{x}] \right) \right] \leq \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}} [g(\boldsymbol{x})]\,. \tag{8.8}$$

We apply the above Lemma with $g(\boldsymbol{x}) = (\boldsymbol{v}^T \boldsymbol{x})^2$. We get that

$$\int_{\mathbb{R}^d} (\boldsymbol{v}^T (\boldsymbol{x} + \boldsymbol{\mu} - \mathop{\mathbf{E}}_{y \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{I})} \boldsymbol{y}))^2 \cdot \frac{\mathbf{1}_S(\boldsymbol{x}) \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}; \boldsymbol{x}) d\boldsymbol{x}}{\int_{\mathbb{R}^d} \mathbf{1}_S(\boldsymbol{x}) \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}; \boldsymbol{x}) d\boldsymbol{x}} \leq \int_{\mathbb{R}^d} (\boldsymbol{v}^T \boldsymbol{x})^2 \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}; \boldsymbol{x}) d\boldsymbol{x}\,.$$

Hence, we get the desired variance reduction in the direction $\boldsymbol{v}$

$$\mathrm{Var}_{\boldsymbol{x}\sim\mathcal{N}_S(\boldsymbol{\mu},\boldsymbol{I})}[\boldsymbol{v}^T\boldsymbol{x}] \leq \mathrm{Var}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I})}[\boldsymbol{v}^T\boldsymbol{x}]\,,$$

that implies the concavity of the function $\log\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};S)$ for convex sets S with respect to the mean vector $\boldsymbol{\mu}\in\mathbb{R}^d$. □

Sample Complexity of Empirical Log-Likelihood: Proof of Lemma 8.25

In this section, we provide the proof of Lemma 8.25. This lemma analyzes the sample complexity of the empirical log-likelihood maximization $\mathcal{L}_N$, whose concavity (in convex partitions) was established in Lemma 8.24. We show that, given roughly $N = \widetilde{O}(d/(\epsilon^2\alpha^2))$ samples from $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$, we can guarantee that Gaussian distributions $\mathcal{N}(\boldsymbol{\mu})$ with mean vectors $\boldsymbol{\mu}$, that are far from the true Gaussian $\mathcal{N}(\boldsymbol{\mu}^\star)$ in total variation distance, will also be sub-optimal solutions of the empirical maximization of the log-likelihood objective, i.e., they are far from being maximizers of the empirical log-likelihood objective. We first give an overview of the proof of Lemma 8.25. In Proposition 8.11 we provided a similar sample complexity bound for an empirical log-likelihood objective. However, in contrast to the analysis of Proposition 8.11, the parameter space is now unbounded – $\boldsymbol{\mu}$ can be any vector of $\mathbb{R}^d$ – and we cannot construct a cover of the whole space with finite size. However, thanks to the concavity of the empirical log-likelihood objective $\mathcal{L}_N$, we can show that it suffices to focus on guess vectors $\boldsymbol{\mu}$ that lie in a sphere $\partial\mathcal{B}$ (i.e., the boundary of a ball $\mathcal{B}$) of radius $\Omega(\epsilon)$. This argument heavily relies on the claim that the maximizer of the empirical log-likelihood $\mathcal{L}_N$ lies inside $\mathcal{B}$, which can be verified by monotonicity properties of the log-likelihood. Afterwards, we consider a discretization $\mathcal{C}$ of the sphere and, for any vector $\boldsymbol{\mu}\in\mathcal{C}$, we can prove that $\mathcal{L}_N(\boldsymbol{\mu}^\star) - \mathcal{L}_N(\boldsymbol{\mu}) \geq \Omega(\alpha^2\epsilon^2)$. The main technical tool for this claim is a concentration result on likelihood ratios and the fact that the partition distribution is α-information preserving. In order to extend this property to the whole sphere, we exploit the convexity (with respect to $\boldsymbol{\mu}$) of a regularized version of the empirical log-likelihood objective $\mathcal{L}_N(\boldsymbol{\mu}) + \|\boldsymbol{\mu}\|_2^2$. The complete proof follows.

Proof of Lemma 8.25. Let $\widetilde{\boldsymbol{\mu}}$ be the maximizer of the empirical log-likelihood objective

$$\widetilde{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} \log \mathcal{N}(\boldsymbol{\mu}; S_i)\,.$$

Since $\widetilde{\boldsymbol{\mu}}$ is the maximizer of the empirical objective, it is sufficient to prove that for any Gaussian $\mathcal{N}(\boldsymbol{\mu})$ whose total variation distance with $\mathcal{N}(\boldsymbol{\mu}^\star)$ is greater than ϵ, it holds that $\mathcal{L}_N(\boldsymbol{\mu}^\star) - \mathcal{L}_N(\boldsymbol{\mu}) \geq \Omega(\alpha^2\epsilon^2)$.

Moreover, we know that when $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2$ is smaller than some sufficiently small absolute constant, it holds $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_1), \mathcal{N}(\boldsymbol{\mu}_2)) \geq \Omega(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2)$. Therefore, any Gaussian whose mean $\boldsymbol{\mu}$ is far from $\boldsymbol{\mu}^\star$, i.e., $\|\boldsymbol{\mu} - \boldsymbol{\mu}^\star\|_2 \geq \Omega(\epsilon)$ will be in total variation distance at least ϵ from $\mathcal{N}(\boldsymbol{\mu}^\star)$ Therefore, to prove the lemma, it suffices to prove it for Gaussians whose means lie outside of a ball $\mathcal{B}$ of radius $\rho := \Omega(\epsilon)$ around $\boldsymbol{\mu}^\star$.

Since all observed sets S_i are convex, the empirical log-likelihood objective $\mathcal{L}_N(\boldsymbol{\mu})$ is concave with respect to $\boldsymbol{\mu}$, see Lemma 8.24. Since $\mathcal{L}_N$ is concave, it suffices to prove that for any $\boldsymbol{\mu}$ that lies exactly on the sphere of radius ρ, i.e., the surface of the ball $\mathcal{B}$ it holds $\mathcal{L}_N(\boldsymbol{\mu}^\star) - \mathcal{L}_N(\boldsymbol{\mu}) \geq \Omega(\alpha^2\epsilon^2)$. To prove this we first show that the maximizer of the empirical objective $\widetilde{\boldsymbol{\mu}}$ has to lie inside the ball $\mathcal{B}$. Assuming that $\widetilde{\boldsymbol{\mu}}$ lies outside of $\mathcal{B}$, let $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ be the antipodal points on the sphere $\partial\mathcal{B}$ that belong to the line $\widetilde{\boldsymbol{\mu}}$ connecting $\widetilde{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}^\star$ and assume that $\boldsymbol{r}_2$ lies between $\boldsymbol{\mu}^\star$ and $\widetilde{\boldsymbol{\mu}}$. In that case the restriction of $\mathcal{L}_N$ on that line cannot be concave, since it has to be increasing from $\boldsymbol{r}_1$ to $\boldsymbol{\mu}^\star$, decreasing from $\boldsymbol{\mu}^\star$ to $\boldsymbol{r}_2$ and then increase again from $\boldsymbol{r}_2$ to $\widetilde{\boldsymbol{\mu}}$. Thus, $\widetilde{\boldsymbol{\mu}}$ lies inside $\mathcal{B}$. Now, by concavity of $\mathcal{L}_N$, we obtain that, by projecting any point $\boldsymbol{\mu}$ that lies outside of the ball $\mathcal{B}$ onto $\mathcal{B}$, we can only increase its empirical likelihood. Therefore, it suffices to consider only points that lie on the sphere $\partial\mathcal{B}$.

We will now show that the claim is true for any $\boldsymbol{\mu} \in \partial\mathcal{B}$. We can create a cover of the sphere of radius $\rho\sqrt{1 + c\alpha^2}$, centered at $\boldsymbol{\mu}^\star$ for some sufficiently small absolute constant $c > 0$, whose convex hull contains $\mathcal{B}$. The following lemma shows that such a cover can be constructed with $(1/(\alpha\epsilon))^{O(d)}$ points.

Lemma 8.27 (see, e.g., Corollary 4.2.13 of Vershynin (2018b))**.** *For any $\epsilon > 0$, there exists an ϵ-cover $\mathcal{C}$ of the unit sphere in $\mathbb{R}^k$, with respect to the ℓ_2-norm, of size $O((1/\epsilon)^k)$. Moreover, the convex hull of the cover $\mathcal{C}$ contains the sphere of radius $1-\epsilon$.*

Since the partition distribution π is α-information preserving we obtain that for any $\boldsymbol{\mu} \in \mathcal{C}$, it holds $d_{\mathrm{TV}}(\mathcal{N}_\pi(\boldsymbol{\mu}), \mathcal{N}_\pi(\boldsymbol{\mu}^\star)) \geq \Omega(\alpha\epsilon)$. Applying Lemma 8.13 with $x = O(\log(|\mathcal{C}|/\delta)) = O(d\log(1/(\epsilon\delta)))$, we get that, with $N = \widetilde{O}(d/(\alpha^2\epsilon^2)\log(1/\delta))$, with probability at least $1-\delta$, it holds that, for any $\boldsymbol{\mu}$ in the cover $\mathcal{C}$, we have

$$\mathcal{L}_N(\boldsymbol{\mu}^\star) - \mathcal{L}_N(\boldsymbol{\mu}) \geq d_{\mathrm{TV}}(\mathcal{N}_\pi(\boldsymbol{\mu}^\star), \mathcal{N}_\pi(\boldsymbol{\mu}))^2 - \alpha^2\epsilon^2/2 \geq \Omega(\alpha^2\epsilon^2)\,. \tag{8.9}$$

Next, we need to extend this bound from the elements of the cover $\mathcal{C}$ to all elements of the sphere $\partial\mathcal{B}$. In what follows, in order to simplify notation, we may assume without loss of generality that $\boldsymbol{\mu}^\star = \mathbf{0}$. We are going to use the fact that $\log(\mathcal{N}(\boldsymbol{\mu}; S_i)) + \|\boldsymbol{\mu}\|_2^2/2$ is convex. To see that, write

$$\log(\mathcal{N}(\boldsymbol{\mu}; S_i)) + \|\boldsymbol{\mu}\|_2^2/2 = \log\left(e^{\|\boldsymbol{\mu}\|_2^2/2}\int_S e^{-\|\boldsymbol{x}-\boldsymbol{\mu}\|_2^2/2}d\boldsymbol{x}\right) = \log\left(\int_S e^{-\|\boldsymbol{x}\|_2^2/2+\boldsymbol{x}^T\boldsymbol{\mu}}d\boldsymbol{x}\right),$$

which is a log-sum-exp function and thus convex (this can also be verified by directly computing the Hessian with respect to $\boldsymbol{\mu}$). This means that $\mathcal{L}_N(\boldsymbol{\mu}) + \|\boldsymbol{\mu}\|_2^2$ is also convex with respect to $\boldsymbol{\mu}$. Let $\boldsymbol{\mu} \in \partial B$. From the construction of the cover $\mathcal{C}$, we have that its convex hull contains the sphere ∂B. Therefore, $\boldsymbol{\mu}$ can be written as a convex combination of points of the cover, i.e., $\boldsymbol{\mu} = \sum_{i=1}^{|\mathcal{C}|}\alpha_i\boldsymbol{\mu}_i$, where $\boldsymbol{\mu}_i \in \mathcal{C}$. The convexity of $\mathcal{L}_N(\boldsymbol{\mu}) + \|\boldsymbol{\mu}\|_2^2$ implies that

$$\mathcal{L}_N(\boldsymbol{\mu}) + \|\boldsymbol{\mu}\|_2^2 \leq \sum_{i=1}^{|\mathcal{C}|}\alpha_i(\mathcal{L}_N(\boldsymbol{\mu}_i) + \|\boldsymbol{\mu}_i\|_2^2) \leq \max_i \mathcal{L}_N(\boldsymbol{\mu}_i) + \rho^2(1+c\alpha^2)\,,$$

where to get the last inequality we used the fact that all points of our cover $\mathcal{C}$ belong to the sphere of radius $\rho\sqrt{1+c\alpha^2}$. Since $\|\boldsymbol{\mu}\|_2^2 = \rho^2$ the above inequality implies that $\mathcal{L}_N(\boldsymbol{\mu}) \leq \max_i \mathcal{L}_N(\boldsymbol{\mu}_i) + c\alpha^2\rho^2$. Combining this inequality with Equation (8.9), we obtain that, since c is sufficiently small and $\rho = \Theta(\epsilon)$, it holds $\mathcal{L}_N(\boldsymbol{\mu}) \leq$

$\mathcal{L}_N(\boldsymbol{\mu}^\star) - \Omega(\epsilon^2\alpha^2)$. □

The Proof of Theorem 8.8

We conclude this section with the proof of Theorem 8.8. Since the likelihood function is concave (and therefore can be efficiently optimized) we focus mainly on bounding the sample complexity of our algorithm.

Proof of Theorem 8.8. Let us assume that the partition distribution π is α-information preserving and that is supported on *convex partitions* of $\mathbb{R}^d$. Our goal is to show that there exists an algorithm, that draws $\widetilde{O}(d/(\epsilon^2\alpha^2)\log(1/\delta))$ samples from $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$ and computes an estimate $\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^d$ so that $d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}}), \mathcal{N}(\boldsymbol{\mu}^\star)) \leq \epsilon$ with probability at least $1-\delta$. The algorithm works as follows: it optimizes the empirical log-likelihood objective

$$\mathcal{L}_N(\boldsymbol{\mu}) = \frac{1}{N}\sum_{i=1}^{N} \log \mathcal{N}(\boldsymbol{\mu}; S_i)\,,$$

where the samples are i.i.d. and $S_i \sim \mathcal{N}_\pi(\boldsymbol{\mu}^\star)$ for any $i \in [N]$. Using Lemma 8.24, we establish that the function $\mathcal{L}_N$ is concave with respect to the mean $\boldsymbol{\mu} \in \mathbb{R}^d$. This follows from the fact that convex combinations of concave functions remain concave. From Lemma 8.25, we obtain that it suffices to compute a point $\boldsymbol{\mu}$ such that $\mathcal{L}_N(\boldsymbol{\mu}) \geq \max_{\boldsymbol{\mu}'} \mathcal{L}_N(\boldsymbol{\mu}') - O(\alpha^2\epsilon^2)$. Specifically, given roughly $\widetilde{O}(d/(\epsilon^2\alpha^2))$ samples from $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$, we can guarantee, with high probability, that the maximizer $\widetilde{\boldsymbol{\mu}}$ of the empirical log-likelihood achieves a total variation gap at most ϵ against the true mean vector $\boldsymbol{\mu}^\star$, i.e., $d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}}), \mathcal{N}(\boldsymbol{\mu}^\star)) \leq \epsilon$. □

We proceed with a discussion about the running time of the above algorithm. Since $\mathcal{L}_N(\boldsymbol{\mu})$ is a concave function with respect to $\boldsymbol{\mu}$, this can be done efficiently. For example, we may perform gradient-ascent: for a fixed convex set $S \subseteq \mathbb{R}^d$ the gradient of the function $f(\boldsymbol{\mu}) = \log \mathcal{N}(\boldsymbol{\mu}; S) = \log \mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu})}[\mathbf{1}\{\boldsymbol{x} \in S\}]$ (see

Lemma 8.24) is equal to

$$\nabla_{\mu} f(\mu) = \underset{x \sim \mathcal{N}_S(\mu)}{\mathbf{E}}[x] - \mu\,.$$

In order to compute the gradient of f, it suffices to approximately compute $\mathbf{E}_{x \sim \mathcal{N}_S(\mu)}[x] = \mathbf{E}_{x \sim \mathcal{N}(\mu)}[x\ \mathbf{1}\{x \in S\}]\ / \mathcal{N}(\mu; S)$. Both terms of this ratio can be estimated using independent samples from the distribution $\mathcal{N}(\mu)$ and access to the oracle $\mathcal{O}_S(\cdot)$, since the mean μ is known (the current guess of the learning algorithm). Hence, the running time will be polynomial in the number of samples using, e.g., the ellipsoid algorithm.

Remark 8.28. *We remark that a precise calculation of the runtime would also depend on the regularity of the concave objective (Lipschitz or smoothness assumptions etc.) which in turn depend on the geometric properties of the sets. We opt not to track such dependencies since our main result is that, in this setting, the likelihood objective is concave and therefore can be efficiently optimized using standard black-box optimization techniques.*

8.4 Further Related Work

Our work is closely related to the literature of learning from censored-truncated data and learning with noise. There has been a large number of recent works dealing inference with truncated data from a Gaussian distribution Daskalakis et al. (2018); Kontonis et al. (2019), mixtures of Gaussians Nagarajan and Panageas (2019), linear regression Daskalakis et al. (2019); Ilyas et al. (2020); Daskalakis et al. (2020), sparse Graphical models Bhattacharyya et al. (2020) or Boolean product distributions Fotakis et al. (2020), and non-parametric estimation Daskalakis et al. (2021). A significant feature of our work is that it can capture the closely related field of censored statistics Cohen (2016); Breen et al. (1996); Wolynetz (1979).

The area of robust statistics Huber (2004) is also very related to our work as it also deals with biased data-sets and aims to identify the distribution that generated the data. Recently, there has been a large volume of theoretical work for computationally-efficient robust estimation of high-dimensional distributions Diakonikolas et al. (2016b); Charikar et al. (2017); Lai et al. (2016b); Diakonikolas et al. (2017a, 2018c); Klivans et al. (2018); Hopkins and Li (2019); Diakonikolas et al. (2019b); Cheng et al. (2020); Bakshi et al. (2020) in the presence of arbitrary corruptions to a small ε fraction of the samples.

The line of research dealing with statistical queries Kearns (1998); Blum et al. (1998); Feldman et al. (2015b,a); Feldman (2017); Feldman et al. (2017); Diakonikolas et al. (2017b, 2020a) is closely related to one of our main results (Theorem 8.4). It is generally believed that SQ algorithms capture all reasonable machine learning algorithms Aslam and Decatur (1998); Blum et al. (1998, 2005); Dunagan and Vempala (2008); Feldman et al. (2017); Balcan and Feldman (2015); Feldman et al. (2015a) and there is a rich line of research indicating SQ lower-bounds for these classes of algorithms Feldman et al. (2017); Diakonikolas et al. (2017b); Shamir (2018); Vempala and Wilmes (2019); Diakonikolas et al. (2020a,d); Goel et al. (2020a,c).

Learning from coarse labels is also referred in the ML literature as Partial Label Learning Cour et al. (2011); Chen et al. (2014); Yu and Zhang (2016) (a weakly supervised learning problem where each training example is associated with a set of candidate labels among which only one is true). We refer to Appendix G.4 for an extensive discussion.

A APPENDIX TO CHAPTER 2

A.1 Omitted Technical Lemmas

Formula for the Gradient

Recall that to simplify notation, we will write $\ell(\boldsymbol{w}, \boldsymbol{x}) = \frac{\boldsymbol{w} \cdot \boldsymbol{x}}{\|\boldsymbol{w}\|_2}$. Note that $\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{w}\|_2} - \boldsymbol{w} \cdot \boldsymbol{x} \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2^3}$. The gradient of the objective $\mathcal{L}_\sigma^{\text{ramp}}(\boldsymbol{w})$ is then

$$\begin{aligned}
&\nabla_{\boldsymbol{w}} \mathcal{L}_\sigma^{\text{ramp}}(\boldsymbol{w}) && \text{(A.1)}\\
&= \mathop{\mathbf{E}}_{(x,y)\sim D}\left[-r'_\sigma\left(-y\ \ell(\boldsymbol{w}, \boldsymbol{x})\right) \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{x})\ y\right]\\
&= \mathop{\mathbf{E}}_{(x,y)\sim D}\left[-r'_\sigma\left(\ell(\boldsymbol{w}, \boldsymbol{x})\right)\ \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{x})\ y\right]\\
&= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[-r'_\sigma\left(\ell(\boldsymbol{w}, \boldsymbol{x})\right)\ \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{x})\ \left(\operatorname{sign}(\boldsymbol{w}^\star \cdot \boldsymbol{x})(1-\eta(\boldsymbol{x})) - \operatorname{sign}(\boldsymbol{w}^\star \cdot \boldsymbol{x})\eta(\boldsymbol{x})\right)\right]\\
&= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[-r'_\sigma\left(\ell(\boldsymbol{w}, \boldsymbol{x})\right)\ \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{x})\ (1-2\eta(\boldsymbol{x}))\operatorname{sign}(\boldsymbol{w}^\star \cdot \boldsymbol{x})\right]\ , && \text{(A.2)}
\end{aligned}$$

where in the second equality we used that the $r'_\sigma(t)$ is an even function.

Proof of Claim 2.1

The following claim relates the angle between two vectors and the zero-one loss between the corresponding halfspaces under bounded distributions.

Claim A.1. *2.1 Let $D_{\boldsymbol{x}}$ be a (U, R)-bounded distribution on $\mathbb{R}^d$. Then for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ we have*

$$(R^2/U)\theta(\boldsymbol{u}, \boldsymbol{v}) \le \operatorname{err}_{0-1}^{D_{\boldsymbol{x}}}(h_{\boldsymbol{u}}, h_{\boldsymbol{v}})\ . \tag{A.3}$$

Moreover, if D is $(U, R, t(\cdot))$-bounded, we have that for any $\epsilon \in (0, 1]$

$$\operatorname{err}_{0-1}^{D_{\boldsymbol{x}}}(h_{\boldsymbol{u}}, h_{\boldsymbol{v}}) \le U t(\epsilon)^2 \theta(\boldsymbol{v}, \boldsymbol{u}) + \epsilon\ . \tag{A.4}$$

Proof. Let V be the subspace spanned by $\boldsymbol{v}, \boldsymbol{u}$, and let $(D_{\boldsymbol{x}})_V$ be the projection of $D_{\boldsymbol{x}}$ onto V. Since $\boldsymbol{v} \cdot \boldsymbol{x} = \boldsymbol{v} \cdot \mathrm{proj}_V(\boldsymbol{x})$ and $\boldsymbol{u} \cdot \boldsymbol{x} = \boldsymbol{u} \cdot \mathrm{proj}_V(\boldsymbol{x})$ we have

$$\mathrm{err}_{0-1}^{D_x}(h_{\boldsymbol{u}}, h_{\boldsymbol{v}}) = \mathrm{err}_{0-1}^{(D_x)_V}(h_{\boldsymbol{u}}, h_{\boldsymbol{v}}) \ .$$

Without loss of generality, we can assume that $V = \mathrm{span}(\boldsymbol{e}_1, \boldsymbol{e}_2)$, where $\boldsymbol{e}_1, \boldsymbol{e}_2$ are orthogonal vectors of $\mathbb{R}^2$. Then from Definition 2.2, using the fact that $1/U \le f_V(\boldsymbol{x})$ for all $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_\infty \le R$, which is also true for all $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_2 \le R$, the above probability is bounded below by $\frac{R^2}{U}\theta(\boldsymbol{u}, \boldsymbol{v})$, which proves (A.3). To prove (A.4), we observe that

$$\begin{aligned}
&\mathrm{err}_{0-1}^{(D_x)_V}(h_{\boldsymbol{u}}, h_{\boldsymbol{v}}) \\
&\le \Pr_{\boldsymbol{x} \sim (D_{\boldsymbol{x}})_V}[\mathrm{sign}(\boldsymbol{u} \cdot \boldsymbol{x}) \neq \mathrm{sign}(\boldsymbol{v} \cdot \boldsymbol{x}) \text{ and } \ \|\boldsymbol{x}\|_2 \le t(\epsilon)] + \Pr_{\boldsymbol{x} \sim (D_{\boldsymbol{x}})_V}[\|\boldsymbol{x}\|_2 \ge t(\epsilon)] \\
&\le Ut(\epsilon)^2\theta + \epsilon.
\end{aligned}$$

□

Relation Between Misclassification Error and Error to Target Halfspace

The following well-known fact relates the misspecification error with respect to D and the zero-one loss with respect to the optimal halfspace. We include a proof for the sake of completeness.

Fact A.2. *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$, $\eta < 1/2$ be an upper bound on the Massart noise rate. Then if $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^\star \cdot \boldsymbol{x})$ and $h(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{u} \cdot \boldsymbol{x})$ we have*

$$\mathrm{err}_{0-1}^{D_x}(h, f) \le \frac{1}{1 - 2\eta}\left(\mathrm{err}_{0-1}^{D}(h) - \mathrm{opt}\right) \ .$$

Proof. We have that

$$\begin{aligned}
\mathrm{err}_{0-1}^{D}(h) &= \mathop{\mathbf{E}}_{(x,y)\sim D}[\mathbb{1}\{h(x) \neq f(x)\}\\
&= \mathop{\mathbf{E}}_{x\sim D_x}[(1-\eta(x))\mathbb{1}\{h(x) \neq f(x)\}] + \mathop{\mathbf{E}}_{x\sim D_x}[\eta(x)\mathbb{1}\{h(x) = f(x)\}]\\
&= \mathop{\mathbf{E}}_{x\sim D_x}[(1-2\eta(x))\mathbb{1}\{h(x) \neq f(x)\}] + \mathop{\mathbf{E}}_{x\sim D_x}[\eta(x)]\\
&\geq \mathop{\mathbf{E}}_{x\sim D_x}[(1-2\eta)\mathbb{1}\{h(x) \neq f(x)\}] + \mathrm{opt}\\
&= (1-2\eta)\,\mathrm{err}_{0-1}^{D_x}(h,f) + \mathrm{opt}\,,
\end{aligned}$$

where in the second inequality we used that $\eta(x) \leq \eta$ and $\mathbf{E}_{x\sim D_x}[\eta(x)] = \mathrm{opt}$. □

Log-concave and s-concave distributions are bounded

Lemma A.3 (Isotropic log-concave density bounds Lovász and Vempala (2007))**.** *Let γ be the density of any isotropic log-concave distribution on $\mathbb{R}^d$. Then $\gamma(x) \geq 2^{-6d}$ for all x such that $0 \leq \|x\|_2 \leq 1/9$. Furthermore, $\gamma(x) \leq \mathrm{e}\, 2^{8d} d^{d/2}$ for all x.*

We are also going to use the following concentration inequality providing sharp bounds on the tail probability of isotropic log-concave distributions.

Lemma A.4 (Paouris' Inequality Paouris (2006))**.** *There exists an absolute constant $c > 0$ such that if D_x is any isotropic log-concave distribution on $\mathbb{R}^d$, then for all $t > 1$ it holds*

$$\mathop{\mathbf{Pr}}_{x\sim D_x}[\|x\|_2 \geq ct\sqrt{d}] \leq \exp(-t\sqrt{d})\,.$$

Fact A.5. *An isotropic log-concave distribution on $\mathbb{R}^d$ is $(e2^{17}, 1/9, c\log(1/\epsilon) + 2c)$-bounded, where $c > 0$ is the absolute constant of Lemma A.4.*

Proof. Follows immediately from Lemma A.3, Lemma A.4, and the fact that the marginals of isotropic log-concave distributions are also isotropic log-concave. □

Now we are going to prove that s-concave are also (U, R, t) bounded for all $s \geq -\frac{1}{2d+3}$. We will require the following lemma:

Lemma A.6 (Theorem 3 Balcan and Zhang (2017a)). *Let $\gamma(\boldsymbol{x})$ be an isotropic s-concave distribution density on $\mathbb{R}^d$, then the marginal on a subspace of $\mathbb{R}^2$ is $\frac{s}{1+(d-2)s}$-concave.*

Lemma A.7 (Theorem 5 Balcan and Zhang (2017a)). *Let $\boldsymbol{x}$ come from an isotropic distribution over $\mathbb{R}^d$, with s-concave density. Then for every $t \geq 16$, we have*

$$\mathbf{Pr}[\|\boldsymbol{x}\|_2 > \sqrt{d}t] \leq \left(1 - \frac{cst}{1+ds}\right)^{(1+ds)/s},$$

where c is an absolute constant.

Lemma A.8 (Theorem 9 Balcan and Zhang (2017a)). *Let $\gamma : \mathbb{R}^d \to \mathbb{R}_+$ be an isotropic s-concave density. Then*

(a) Let $D(s,d) = (1+\alpha)^{-1/\alpha}\frac{1+3\beta}{3+3\beta}$, where $\beta = \frac{s}{1+(d-1)s}$, $\alpha = \frac{\beta}{1+\beta}$ and $\zeta = (1+\alpha)^{-\frac{1}{\alpha}}\frac{1+3\beta}{3+3\beta}$. For any $\boldsymbol{x} \in \mathbb{R}^d$ such that $\|\boldsymbol{x}\| \leq D(s,d)$, we have

$$\gamma(\boldsymbol{x}) \geq \left(\frac{\|\boldsymbol{x}\|}{\zeta}((2-2^{-(d+1)s})^{-1}-1)+1\right)^{1/s}\gamma(0).$$

(b) $\gamma(\boldsymbol{x}) \leq \gamma(0)\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\alpha)^{3/\alpha}}2^{d-1+1/s}\right)^s - 1\right]^{1/s}$ for every $\boldsymbol{x}$.

(c) $(4e\pi)^{-d/2}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\alpha)^{3/\alpha}}2^{d-1+\frac{1}{s}}\right)^s - 1\right]^{-\frac{1}{s}} < \gamma(0) \leq (2-2^{-(d+1)s})^{1/s}\frac{d\Gamma(d/2)}{2\pi^{d/2}\zeta^d}$.

(d) $\gamma(\boldsymbol{x}) \leq (2-2^{-(d+1)s})^{1/s}\frac{d\Gamma(d/2)}{2\pi^{d/2}\zeta^d}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\alpha)^{3/\alpha}}2^{d-1+1/s}\right)^s - 1\right]^{1/s}$ for every $\boldsymbol{x}$.

Lemma A.9. *Any isotropic s-concave distribution on $\mathbb{R}^d$ with $s \geq -\frac{1}{2d+3}$, is $\left(\Theta(1), \Theta(1), c/\epsilon^{1/6}\right)$-bounded where c is an absolute constant.*

Proof. Set $\Gamma = \left(\left(\frac{1+2s}{1+4s}\sqrt{3(1+s/(1+2s))^{(3+6s)/s}}2^{1+1/s}\right)^s - 1\right)^{1/s}$. From Lemma A.8, we have

1. For any $\boldsymbol{x} \in \mathbb{R}^2$ such that $\|\boldsymbol{x}\|_2 \leq (1+\frac{s}{1+2s})^{-\frac{1+2s}{s}}(\frac{1+4s}{3+6s})$, we have $\gamma(\boldsymbol{x}) \geq \frac{1}{4e\pi\Gamma}$.

2. For any $\boldsymbol{x} \in \mathbb{R}^2$, we have: $\gamma(\boldsymbol{x}) \leq \frac{(2^{3s+1}-1)^{1/s}(3+6s)^2\Gamma}{4\pi(1+4s)^2(\frac{1+3s}{1+2s})^{-\frac{1+2s}{s}}}$.

From Lemma A.6, we have that the marginals of an isotropic s-concave distribution on $\mathbb{R}^d$, on a 2-dimensional subspace, are s'-concave where $s' = \frac{s}{1+(d-2)s}$. Using $s \geq -\frac{1}{2d+3}$, for $d \geq 3$, we have $s' > -\frac{1}{8}$ and when $d = 2$, we have $s' = s \geq -1/7$. Thus, the value of s' is lower bounded by $-1/7$. To find the values (U, R), we need to find a lower bound and an upper bound on density. From the expression of Γ, we observe that for $s' \geq -1/7$ it holds $\Gamma < 34 \cdot 10^3$. Therefore, we obtain the following bounds

$$\begin{aligned}
\gamma(\boldsymbol{x}) &\geq \frac{1}{4e\pi\Gamma} > \frac{1}{10^7} \,, \\
R &= \left(1 + \frac{s'}{1+2s'}\right)^{-\frac{1+2s}{s'}} \frac{1+4s'}{3+6s'} \geq 0.065 \,, \\
\gamma(\boldsymbol{x}) &\leq \frac{(2^{3s'+1} - 1)^{1/s'}(3+6s')^2\Gamma}{4\pi(1+4s')^2(\frac{1+3s'}{1+2s'})^{-\frac{1+2s'}{s'}}} < 3.3 \cdot 10^7 \,,
\end{aligned}$$

where we simplified each expression using the bounds of s'. From Lemma A.7 we get tail bounds, by taking the appropriate s' that maximizes the error in the tail bound (which is $s' = -1/7$). This completes the proof. □

A.2 Omitted Proofs from Section 2.4

In Section A.2, we establish the convergence properties of projected SGD that we require. Even though this lemma should be folklore, we did not find an explicit reference. In Section A.2, we establish the smoothness of our non-convex surrogate function.

Proof of Lemma 2.7

For convenience, we restate the lemma here.

Lemma A.10 (PSGD). *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ with $f(\boldsymbol{w}) = \mathbf{E}_{z\sim D}[g(\boldsymbol{z}, \boldsymbol{w})]$ for some function $g : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$. Assume that for any vector $\boldsymbol{w}$, $g(\cdot, \boldsymbol{w})$ is positive homogeneous of* *degree-*0 *on* $\boldsymbol{w}$. *Let* $\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_2 \geq 1\}$ *and assume that* f, g *are continuously differentiable functions on* $\mathcal{W}$. *Moreover, assume that* $|f(\boldsymbol{w})| \leq R$, $\nabla_{\boldsymbol{w}} f(\boldsymbol{w})$ *is L-Lipschitz on* $\mathcal{W}$, $\mathbf{E}_{z\sim D}\left[\|\nabla_{\boldsymbol{w}} g(\boldsymbol{z}, \boldsymbol{w})\|_2^2\right] \leq B$ *for all* $\boldsymbol{w} \in \mathcal{W}$. *After T iterations the output* $(\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)})$ *of Algorithm 1 satisfies*

$$\mathop{\mathbf{E}}_{z^{(1)},\ldots,z^{(T)}\sim D}\left[\frac{1}{T}\sum_{i=1}^{T}\left\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i)})\right\|_2^2\right] \leq \sqrt{\frac{LBR}{2T}}\,.$$

If, additionally, $\|\mathbf{E}_{z\sim D}[\nabla_{\boldsymbol{w}} g(\boldsymbol{z}, \boldsymbol{w})]\|_2^2 \leq C$ *for all* $\boldsymbol{w} \in \mathcal{W}$, *we have that with* $T = (2LBR + 8C^2\log(1/\delta))/\epsilon^4$ *it holds* $\min_{i=1,\ldots,T}\left\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i)})\right\|_2 \leq \epsilon$, *with probability at least* $1-\delta$.

Proof. Consider the update $\boldsymbol{v}^{(i)} = \boldsymbol{w}^{(i-1)} - \beta\nabla g(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i-1)})$ at iteration i of Algorithm 1. The projection step on the unit sphere (line 6 of Algorithm 1) ensures that $\left\|\boldsymbol{w}^{(i-1)}\right\|_2 = 1$. Observe that, since $g(\boldsymbol{z}, \boldsymbol{w})$ is constant in the direction of $\boldsymbol{w}$, we have that $\nabla_{\boldsymbol{w}} g(\boldsymbol{z}, \boldsymbol{w}^{(i-1)})$ is perpendicular to $\boldsymbol{w}^{(i-1)}$. Therefore, by the Pythagorean theorem, $\left\|\boldsymbol{v}^{(i)}\right\|_2^2 = \left\|\boldsymbol{w}^{(i-1)}\right\|_2^2 + \beta^2\left\|\nabla g(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i-1)})\right\|_2^2 > 1$ which implies that $\boldsymbol{v}^{(i)} \in \mathcal{W}$. Observe that the line that connects $\boldsymbol{v}^{(i)}$ and $\boldsymbol{w}^{(i-1)}$ is also contained in $\mathcal{W}$. Therefore, we have

$$\begin{aligned}
&f(\boldsymbol{v}^{(i)}) - f(\boldsymbol{w}^{(i-1)}) \\
&= \nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i-1)}) \cdot \boldsymbol{v}^{(i)} - \boldsymbol{w}^{(i-1)} \\
&+ \int_0^1 \nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i-1)} + t(\boldsymbol{v}^{(i)} - \boldsymbol{w}^{(i-1)})) - \nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i-1)}) \cdot (\boldsymbol{v}^{(i)} - \boldsymbol{w}^{(i-1)})\mathrm{d}t \\
&\leq -\beta\nabla f(\boldsymbol{w}^{(i-1)}) \cdot \nabla_{\boldsymbol{w}} g(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i-1)}) + \frac{\beta^2 L}{2}\left\|\nabla_{\boldsymbol{w}} g(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i-1)})\right\|_2^2.
\end{aligned}$$

Observe now that, since f does not depend on the length of its argument, we have $f(\boldsymbol{v}^{(i)}) = f(\boldsymbol{w}^{(i)})$ and therefore

$$f(\boldsymbol{w}^{(i)}) - f(\boldsymbol{w}^{(i-1)}) \leq -\beta\nabla f(\boldsymbol{w}^{(i-1)}) \cdot \nabla_{\boldsymbol{w}} g(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i-1)}) + \frac{\beta^2 L}{2}\left\|\nabla_{\boldsymbol{w}} g(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i-1)})\right\|_2^2.$$

Conditioning on the previous samples $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(i-1)}$ we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{z}^{(i)}}[f(\boldsymbol{w}^{(i)}) - f(\boldsymbol{w}^{(i-1)}) | \boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(i-1)}] \\
&\leq \beta \left\| \nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i-1)}) \right\|_2^2 + \frac{\beta^2 L}{2} \mathop{\mathbf{E}}_{\boldsymbol{z}^{(i)}} \left[\left\| \nabla_{\boldsymbol{w}} g(\boldsymbol{z}^{(i)}, \boldsymbol{w}^{(i-1)}) \right\|_2^2 \right] \\
&\leq -\beta \left\| \nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(i-1)}) \right\|_2^2 + \frac{\beta^2 L B}{2}.
\end{aligned}$$

Rearranging the above inequality, taking the average over T iterations and using the law of total expectation, we obtain that by setting $\beta = \sqrt{2R/(LBT)}$. To get the high-probability version, we set

$$S_T(\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)}) = (1/T) \sum_{i=1}^{T} \left\| \nabla f(\boldsymbol{w}^{(i)}) \right\|_2^2 .$$

Notice that with $T = 2LBR/\epsilon^4$ from the previous argument we obtain that $\mathbf{E}[S_T(\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)})] \leq \epsilon^2/2$. Observe that

$$\begin{aligned}
&\left| S_T(\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(i)}, \ldots, \boldsymbol{w}^{(T)}) - S_T(\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(i)'}, \ldots, \boldsymbol{w}^{(T)}) \right| \\
&\leq \frac{\left| \left\| \nabla f(\boldsymbol{w}^{(i)}) \right\|_2^2 - \left\| \nabla f(\boldsymbol{w}^{(i)'}) \right\|_2^2 \right|}{T} \\
&\leq \frac{2C}{T} \, .
\end{aligned}$$

Lemma A.11 (Theorem 2.2 of Devroye and Lugosi (2001)). *Suppose that $X_1, \ldots X_d \in \mathcal{X}$ are independent random variables, and let $f : \mathcal{X}^d \mapsto \mathbb{R}$. Let $c_1, \ldots, c_n$ satisfy*

$$\sup_{x_1, \ldots, x_d, x_i'} |f(x_1, \ldots x_i, \ldots x_d) - f(x_1, \ldots x_i', \ldots x_d)| \leq c_i$$

for $i \in [d]$. Then

$$\mathbf{Pr}[f(X) - \mathbf{E}[f(X)] \geq t] \leq \exp\left(- 2t^2 / \sum_{i=1}^{d} c_i^2 \right).$$

Now using Lemma A.11, we obtain that

$$\mathbf{Pr}[S_T(\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)}) - \mathbf{E}[S_T(\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)})] > t] \leq \exp(-t^2 T/(2C^2)).$$

Choosing $T \geq 2LBR/\epsilon^4 + 8C^2 \log(1/\delta)/\epsilon^4$ and combining the above bounds, gives us that with probability at least $1 - \delta$, it holds $S_T(\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(T)}) \leq \epsilon^2$. Since the minimum element is at most the average, we obtain that with probability at least $1 - \delta$ it holds

$$\min_{i \in [T]} \left\| \nabla f(\boldsymbol{w}^{(i)}) \right\|_2 \leq \epsilon \ .$$

This completes the proof. □

Proof of Lemma 2.8

We start with the following more general lemma from which we can deduce Lemma 2.8.

Lemma A.12 (Objective Properties)**.** *Let D be a distribution on $\mathbb{R}^d \times \{-1, +1\}$ such that the marginal $D_{\boldsymbol{x}}$ on $\mathbb{R}^d$ is in isotropic position. Let $g(\boldsymbol{x}, y, \boldsymbol{w}) = f(-y\boldsymbol{w} \cdot \boldsymbol{x})$ and*

$$\mathcal{L}_\sigma(\boldsymbol{w}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},y) \sim D} [g(\boldsymbol{x}, y, \boldsymbol{w})] \ .$$

Assume that f is a twice differentiable function on $\mathbb{R}$ such that $|f(t)| \leq R$, $|f'(t)| \leq B$, and $f''(t) \leq K$ for all $t \in \mathbb{R}$. Then $\mathcal{L}_\sigma(\boldsymbol{w})$ is continuously differentiable, $|\mathcal{L}_\sigma(\boldsymbol{w})| \leq R$ for all $\boldsymbol{w}$ in $\mathcal{W} = \{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \geq 1\}$, $\mathbf{E}_{(\boldsymbol{x},y) \sim D}[\|\nabla_{\boldsymbol{w}} g(\boldsymbol{x}, y, \boldsymbol{w})\|_2^2] \leq 4B^2 d$,

$$\left\| \mathop{\mathbf{E}}_{(\boldsymbol{x},y) \sim D} [\nabla_{\boldsymbol{w}} g(\boldsymbol{x}, y, \boldsymbol{w})] \right\|_2^2 \leq 3B^2$$

, and $\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\boldsymbol{w})$ is $(6B+4K)$-Lipschitz.

Proof. Write $g(\boldsymbol{x},y,\boldsymbol{w}) = f(\ell(\boldsymbol{w},\boldsymbol{x})y)$, where $\ell(\boldsymbol{w},\boldsymbol{x}) = \boldsymbol{w}\cdot\boldsymbol{x}/\left\|\boldsymbol{w}\right\|_2$. Note that $|g(\boldsymbol{x},y,\boldsymbol{w})| \leq R$. Therefore, $|\mathcal{L}_\sigma(\boldsymbol{w})| \leq R$.

We now deal with the function $\ell(\boldsymbol{w},\boldsymbol{x}) = \boldsymbol{w}\cdot\boldsymbol{x}/\left\|\boldsymbol{w}\right\|_2$. We have that $\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{w}\|_2} - \boldsymbol{w}\cdot\boldsymbol{x}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2^3}$. Observe that $\left\|\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{x})\right\|_2 \leq 2\left\|\boldsymbol{x}\right\|_2/\left\|\boldsymbol{w}\right\|_2 \leq 2\left\|\boldsymbol{x}\right\|_2$. Therefore, since $D_{\boldsymbol{x}}$ is isotropic, we get that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[\left\|\nabla_{\boldsymbol{w}}g(\boldsymbol{x},y,\boldsymbol{w})\right\|_2^2] \leq 4B^2\,\mathbf{E}_{(\boldsymbol{x},y)\sim D}[\left\|\boldsymbol{x}\right\|_2^2] = 4B^2 d$. Moreover, we have

$$\begin{aligned}\left\|\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[\nabla_{\boldsymbol{w}}g(\boldsymbol{x},y,\boldsymbol{w})]\right\|_2^2 &= \left(\sup_{\|\boldsymbol{v}\|_2=1}\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[\nabla_{\boldsymbol{w}}g(\boldsymbol{x},y,\boldsymbol{w})\cdot\boldsymbol{v}]\right)^2\\ &\leq B^2\left(\sup_{\|\boldsymbol{v}\|_2=1}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{x})\cdot\boldsymbol{v}]\right)^2\\ &\leq B^2\left(\sup_{\|\boldsymbol{v}\|_2=1}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\frac{|\boldsymbol{x}\cdot\boldsymbol{v}|}{\|\boldsymbol{w}\|_2} + |\boldsymbol{w}\cdot\boldsymbol{x}|\frac{|\boldsymbol{w}\cdot\boldsymbol{v}|}{\|\boldsymbol{w}\|_2^3}\right]\right)^2\\ &\leq B^2\left(2\sup_{\|\boldsymbol{v}\|_2=1}\sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|\boldsymbol{x}\cdot\boldsymbol{v}|^2]}\right)^2 \leq 4B^2\,,\end{aligned}$$

where in the first inequality we used $f'(t) \leq B$ and in the third we used the Cauchy-Swartz inequality and that $\left\|\boldsymbol{w}\right\|_2 \geq 1$.

We finally prove that the gradient of $\mathcal{L}_\sigma$ is Lipschitz. We have that

$$\nabla^2_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{x}) = -\frac{\boldsymbol{x}\boldsymbol{w}^T}{\|\boldsymbol{w}\|_2^3} - \frac{\boldsymbol{w}\boldsymbol{x}^T}{\|\boldsymbol{w}\|_2^3} - \frac{\boldsymbol{x}\cdot\boldsymbol{w}}{\|\boldsymbol{w}\|_2^3}\boldsymbol{I} + 3\boldsymbol{x}\cdot\boldsymbol{w}\frac{\boldsymbol{w}\boldsymbol{w}^T}{\|\boldsymbol{w}\|_2^5}\,.$$

Therefore,

$$
\begin{aligned}
\nabla_w^2 g(x,y,w) &= f''(y\ell(w,x))\nabla_w \ell(w,x)\nabla_w \ell(w,x)^T + f'(\ell(w,x))\nabla_w^2 \ell(w,x) \\
&= f''(y\ell(w,x))\left(\frac{xx^T}{\|w\|_2^2} - \frac{w\cdot x}{\|w\|_2^4} wx^T - \frac{w\cdot x}{\|w\|_2^4} xw^T + \frac{w\cdot x^2}{\|w\|_2^6} ww^T\right) \\
&\quad + f'(y\ell(w,x))y\ \nabla_w^2 \ell(w,x).
\end{aligned}
$$

To prove that $\mathcal{L}_\sigma(w)$ has Lipschitz gradient, we will bound $\left\|\nabla_w^2 \mathcal{L}_\sigma(w)\right\|_2$. Let $v \in \mathbb{S}^{d-1}$. We have

$$
\begin{aligned}
\left| v \cdot \mathop{\mathbf{E}}_{(x,y)\sim D}\left[\frac{f''(y\ell(w,x))}{\|w\|_2^2} xx^T\right] v \right| &\le \mathop{\mathbf{E}}_{(x,y)\sim D}\left[\frac{|f''(y\ell(w,x))|}{\|w\|_2^2} x\cdot v^2\right] \\
&\le \frac{K}{\|w\|_2^2} \mathop{\mathbf{E}}_{(x,y)\sim D}\left[x\cdot v^2\right] \le \frac{K}{\|w\|_2^2}\,,
\end{aligned}
$$

where we used the fact that $|f''(t)| \le K$ for all t. To get the last equality, we used the fact that the marginal distribution on x is isotropic. Similarly, we have

$$
\begin{aligned}
&\left| v \cdot \mathop{\mathbf{E}}_{(x,y)\sim D}\left[\frac{f''(y\ell(w,x))}{\|w\|_2^4} w\cdot x w x^T\right] v \right| \le \mathop{\mathbf{E}}_{(x,y)\sim D}\left[\frac{|f''(y\ell(w,x))|}{\|w\|_2^4} |w\cdot x||v\cdot w||x\cdot v|\right] \\
&\le \frac{K}{\|w\|_2^3} \mathop{\mathbf{E}}_{(x,y)\sim D}[|w\cdot x||x\cdot v|] \le \frac{K}{\|w\|_2^3}\sqrt{\mathop{\mathbf{E}}_{(x,y)\sim D}[w\cdot x^2]}\sqrt{\mathop{\mathbf{E}}_{(x,y)\sim D}[x\cdot v^2]} \le \frac{K}{\|w\|_2^3}\,,
\end{aligned}
$$

where the last step follows because the distribution D_x is isotropic. Similarly, we can bound the rest of the terms of $|v^T \nabla_w^2 \mathcal{L}_\sigma(w) v|$ to obtain

$$
|v^T \nabla_w^2 \mathcal{L}_\sigma(w) v| \le B\left(\frac{2}{\|w\|_2^2} + \frac{4}{\|w\|_2^3}\right) + K\left(\frac{1}{\|w\|_2^2} + \frac{2}{\|w\|_2^3} + \frac{1}{\|w\|_2^4}\right) \le 6B + 4K\,,
$$

where we used the fact that $\|w\|_2 \ge 1$. □

Our desired lemma now follows as a corollary.

Lemma A.13 (Sigmoid Smoothness). *Let* $S_\sigma(t) = 1/(1+e^{-t/\sigma})$ *and* $\mathcal{L}_\sigma(w) =$

$\mathbf{E}_{(\boldsymbol{x},y)\sim D}\left[S_\sigma\left(-y\frac{\boldsymbol{w}\cdot\boldsymbol{x}}{\|\boldsymbol{w}\|_2}\right)\right]$, for $\boldsymbol{w} \in \mathcal{W}$, where $\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_2 \geq 1\}$. We have that $\mathcal{L}_\sigma(\boldsymbol{w})$ is continuously differentiable in $\mathcal{W}$, $|\mathcal{L}_\sigma(\boldsymbol{w})| \leq 1$, $\mathbf{E}_{(\boldsymbol{x},y)\sim D}[\|\nabla_{\boldsymbol{w}} S_\sigma(\boldsymbol{w},\boldsymbol{x},y)\|_2^2] \leq 4d/\sigma^2$, $\|\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\boldsymbol{w})\|_2^2 \leq 4/\sigma^2$, and $\nabla_{\boldsymbol{w}}\mathcal{L}_\sigma(\boldsymbol{w})$ is $(6/\sigma + 12/\sigma^2)$-Lipschitz.

Proof. We first observe that $|S_\sigma(t)| \leq 1$ for all t in $\mathbb{R}$. Moreover, S_σ is continuously differentiable. The first and the second derivative of S_σ with respect to t is

$$S'_\sigma(t) = S^2_\sigma(t)\frac{e^{-t/\sigma}}{\sigma} \quad \text{and} \quad S''_\sigma(t) = S^3_\sigma(t)\frac{2e^{-2t/\sigma}}{\sigma^2} - S^2_\sigma(t)\frac{e^{-t/\sigma}}{\sigma^2}\ .$$

We have that $S'_\sigma(t) \leq S'_\sigma(0) = 1/\sigma$ and $S''_\sigma(t) \leq 3/\sigma^2$. The result follows by applying Lemma A.12. □

B Appendix to Chapter 3

B.1 Omitted Proofs

Proof of Lemma 3.5

Lemma B.1. *3.5 Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the Tsybakov noise condition with parameters (α, A). Then for every measurable set $S \subseteq \mathbb{R}^d$ it holds $\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})(1-2\eta(\boldsymbol{x}))] \geq C_\alpha^A \left(\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})]\right)^{\frac{1}{\alpha}}$, where $C_\alpha^A = \alpha\left(\frac{1-\alpha}{A}\right)^{\frac{1-\alpha}{\alpha}}$.*

Proof. We have

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})(1-2\eta(\boldsymbol{x}))] &\geq t \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})\mathbb{1}\{1-2\eta(\boldsymbol{x}) \geq t\}] \\ &\geq t \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})] - t \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})\mathbb{1}\{1-2\eta(\boldsymbol{x}) \leq t\}] \\ &\geq t \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})] - A\, t^{\frac{1}{1-\alpha}} .\end{aligned}$$

Let $G = \mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})]$ and set $t = \left(\frac{(1-\alpha)G}{A}\right)^{\frac{1-\alpha}{\alpha}}$. Then we have

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})(1-2\eta(\boldsymbol{x}))] \geq G^{1/\alpha}\alpha\left(\frac{1-\alpha}{A}\right)^{\frac{1-\alpha}{\alpha}} .$$

□

Proof of Fact 3.7 and Lemma 3.8

Fact B.2. *3.7 We denote by $T_k(t)$ the degree-k Chebyshev polynomial of the first kind. It holds*

$$T_k(t) = \begin{cases} \cos(k \arccos t)\,, & |t| \leq 1 \\ \frac{1}{2}\left(\left(t-\sqrt{t^2-1}\right)^k + \left(t+\sqrt{t^2-1}\right)^k\right), & |t| \geq 1\,. \end{cases}$$

Moreover, it holds $\|T_k\|_2^2 \le 2^{6k+2\log k+4}$.

Proof. Using that $\|T_k\|_2^2 \le \|T_k\|_1^2$, we are going to show that $\|T_k\|_1^2 \le 2^{6k+2\log k+4}$. We have that

$$\|T_k(t)\|_1 = \frac{k}{2}\sum_{i=1}^{\lfloor \frac{k}{2} \rfloor} 2^{k-2i}\binom{k-i}{i}\frac{1}{k-i}x^i \le Fib(k+1)2^k\frac{k}{2} \le \left(1+\sqrt{5}\right)^{k+1} 2^k k \,,$$

where we used that $\sum_{i=1}^{\lfloor \frac{k}{2} \rfloor}\binom{k-i}{i} = Fib(k+1)$. Thus, $\|T_k\|_1^2 \le 2^{6k+2\log k+4}$. □

Lemma B.3. *3.8 Let* $p(t) = \sum_{i=0}^{k} c_i t^i$ *be a degree-k univariate polynomial. Given* $\boldsymbol{w} \in \mathbb{R}^d$ *with* $\|\boldsymbol{w}\|_2 \le 1$, *define the multivariate polynomial* $q(\boldsymbol{x}) = p(\boldsymbol{w}\cdot\boldsymbol{x}) = \sum_{S:|S|\le k} C_S \boldsymbol{x}^S$. *It holds,* $\sum_{S:|S|\le k} C_S^2 \le d^{2k}\sum_{i=0}^{k} c_i^2$. *Moreover, let* $r(t) = p(at+b) = \sum_{i=0}^{k} d_i t^i$ *for some* $a, b \in \mathbb{R}$. *Then* $\|r\|_2^2 \le (2\max(1,a)\max(1,b))^{2k}\|p\|_2^2$.

Proof. We write

$$q(\boldsymbol{x}) = \sum_{i=0}^{k} c_i \boldsymbol{w}\cdot\boldsymbol{x}^i = \sum_{i=0}^{k} c_i \sum_{S:|S|=i}\frac{i!}{S!}\prod_{i=1}^{d}(x_i w_i)^{S_i} = \sum_{i=0}^{k} c_i \sum_{S:|S|=i}\frac{i!}{S!}\boldsymbol{w}^S\boldsymbol{x}^S \,.$$

We have

$$\sum_{i=0}^{k}\sum_{S:|S|=i} c_i^2\left(\frac{i!}{S!}\right)^2 \boldsymbol{w}^{2S} \le \sum_{i=0}^{k} c_i^2\left(\sum_{S:|S|=i}\frac{i!}{S!}\right)^2 \le d^{2k}\sum_{i=0}^{k} c_i^2 \,,$$

where we used the fact that $|\boldsymbol{w}_i| \le 1$ for all i. To prove the second claim, we work similarly. We have

$$r(x) = \sum_{i=0}^{k} c_i \sum_{j=0}^{i}\binom{i}{j} a^j b^{i-j} x^j = \sum_{i=0}^{k} c_i \sum_{j=0}^{i}\binom{i}{j} a^j b^{i-j} x^j.$$

We have

$$\sum_{i=0}^{k} c_i^2 \sum_{j=0}^{i} \left(\binom{i}{j} a^j b^{i-j} \right)^2 \leq (2\max(1,a)\max(1,b))^{2k} \sum_{i=0}^{k} c_i^2 \,.$$

□

Proof of Lemma 3.17

Lemma B.4. *3.17 Let $p_t(\boldsymbol{x})$ be the non-negative function, given from the SDP* (3.3)*. Then taking $d^{O(k)}\log(1/\delta)$ samples, where $k = O\left(\frac{1}{\alpha^2 R\beta}\log^2\left(\frac{BA}{\epsilon LR}\right)\right)$, we can efficiently compute a function $\hat{\ell}_t(\boldsymbol{w})$ such that with probability at least $1-\delta$, the following conditions hold*

- $|\hat{\ell}_t(\boldsymbol{w}) - \mathbf{E}_{(\boldsymbol{x},y)\sim D}[(p_t(\boldsymbol{x})+\lambda)y\boldsymbol{w}\cdot\boldsymbol{x}]| \leq \epsilon$*, for any $\lambda > 0$ and $\boldsymbol{w} \in \mathcal{V}$,*
- $\left\|\nabla_{\boldsymbol{w}}\hat{\ell}_t\right\|_2 \leq d^{O(k)}$ *.*

Proof. For convenience, let $g_t(\boldsymbol{x}) = p_t(\boldsymbol{x}) + \lambda$. The proof is similar to Lemma 3.10. Let $\hat{\ell}_t(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} g_t(\boldsymbol{x}^{(i)})y^{(i)}\boldsymbol{x}^{(i)}\cdot\boldsymbol{w}$ and $\ell_t(\boldsymbol{w}) = \mathbf{E}_{(\boldsymbol{x},y)\sim D}[g_t(\boldsymbol{x})y\boldsymbol{x}\cdot\boldsymbol{w}]$. Then from Cauchy-Schwarz we have

$$|\hat{\ell}_t(\boldsymbol{w}) - \ell_t(\boldsymbol{w})| \leq \left\| \frac{1}{N}\sum_{i=1}^{N} g_t(\boldsymbol{x}^{(i)})y^{(i)}\boldsymbol{x}^{(i)} - \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[g_t(\boldsymbol{x})y\boldsymbol{x}] \right\|_2 \|\boldsymbol{w}\|_2 \,.$$

We have that $\|\boldsymbol{w}\|_2 \leq 1$, thus we need to prove that

$$\mathbf{Pr}\left[\left\| \frac{1}{N}\sum_{i=1}^{N} g_t(\boldsymbol{x}^{(i)})y^{(i)}\boldsymbol{x}^{(i)} - \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[g_t(\boldsymbol{x})y\boldsymbol{x}] \right\|_2 > \epsilon \right] \leq \delta \,. \tag{B.1}$$

Let $\boldsymbol{M}_j = \mathbf{E}_{(\boldsymbol{x},y)\sim D}[m(\boldsymbol{x})m(\boldsymbol{x})^T \mathbb{1}_B(\boldsymbol{x})]\boldsymbol{x}_j$ and $\widetilde{\boldsymbol{M}_j} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}(\boldsymbol{x}^{(i)})\boldsymbol{m}(\boldsymbol{x}^{(i)})^T \mathbb{1}_B(\boldsymbol{x}^{(i)})\boldsymbol{x}_j^{(i)}$, and then $\boldsymbol{A}$ be a matrix such that $\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{M}_j\right) = \mathbf{E}_{(\boldsymbol{x},y)\sim D}[p_t(\boldsymbol{x})y\boldsymbol{x}_j]$, i.e., the matrix of the coefficients of the polynomial and assume that $\|\boldsymbol{A}\|_F \leq Q$, where $Q = d^{O(k)}$.

Using the same proof ideas as in Lemma 3.10, we get

$$\operatorname{tr}\left(\boldsymbol{A}(\boldsymbol{M}_j - \widetilde{\boldsymbol{M}_j})\right) \leq \|\boldsymbol{A}\|_F \left\|\boldsymbol{M}_j - \widetilde{\boldsymbol{M}_j}\right\|_F .$$

Therefore, it suffices to bound the probability that $\left\|\boldsymbol{M}_j - \widetilde{\boldsymbol{M}_j}\right\|_F \geq \epsilon/(2dQ)$. From Markov's inequality, we have

$$\mathbf{Pr}\left[\left\|\boldsymbol{M}_j - \widetilde{\boldsymbol{M}_j}\right\|_F \geq \epsilon/(2dQ)\right] \leq \frac{4d^2Q^2}{\epsilon^2}\,\mathbf{E}\left[\left\|\boldsymbol{M}_j - \widetilde{\boldsymbol{M}_j}\right\|_F^2\right] .$$

Using Equation (3.6) (which holds in our case as well and is proved the same way by setting $\boldsymbol{w} = \boldsymbol{e}_j$), we get

$$\mathbf{Pr}\left[\left\|\boldsymbol{M}_j - \widetilde{\boldsymbol{M}_j}\right\|_F \geq \epsilon/(2dQ)\right] \leq \frac{4d^2Q^2}{\epsilon^2}\frac{1}{N}B(\beta/2)^{-2k}(d+k)^{3k+1} .$$

Then, for $N \geq Bd^3Q^2(\beta/2)^{-2k}(d+k)^{3k+1}/(4\epsilon^2)$ samples we can estimate $\boldsymbol{M}_j$ within the target accuracy with probability at least $1 - 1/(8d)$. Now we are going to give a loose bound for the

$$\mathbf{Pr}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\lambda y^{(i)}\boldsymbol{x}^{(i)} - \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[\lambda y\boldsymbol{x}]\right\|_2 > \epsilon\right] \leq \delta .$$

Using the same argument as before, we have from Markov's inequality, that

$$\mathbf{Pr}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} y^{(i)}\boldsymbol{x}^{(i)} - \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y\boldsymbol{x}]\right\|_2 \geq \epsilon/(2d\lambda)\right] \leq \frac{4d^2\lambda^2}{\epsilon^2}\,\mathbf{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} y^{(i)}\boldsymbol{x}^{(i)} - \mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[y\boldsymbol{x}]\right\|_2^2\right] .$$

Using the linearity of expectation, we have

$$\mathbf{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} y^{(i)}\boldsymbol{x}^{(i)} - \mathop{\mathbf{E}}_{(x,y)\sim D}[y\boldsymbol{x}]\right\|_2^2\right] \le \sum_{j=1}^{d}\mathbf{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} y^{(i)}\boldsymbol{x}_j^{(i)} - \mathop{\mathbf{E}}_{(x,y)\sim D}[y\boldsymbol{x}_j]\right)^2\right]$$
$$\le \sum_{j=1}^{d}\mathbf{Var}\left[\frac{1}{N}\sum_{i=1}^{N} y^{(i)}\boldsymbol{x}_j^{(i)}\right] .$$

Then, using the fact that $\boldsymbol{x}$ is in isotropic position, we have

$$\mathbf{Var}\left[\frac{1}{N}\sum_{i=1}^{N} y^{(i)}\boldsymbol{x}_i^{(i)}\right] \le \frac{1}{N}\mathop{\mathbf{E}}_{(x,y)\sim D}[(\boldsymbol{x}_i^{(i)}y)^2] = 1/N .$$

Thus, for $N > 4d^3\lambda^2/\epsilon^2$, with probability at least $1 - 1/8$, we have that

$$\left\|\frac{1}{N}\sum_{i=1}^{N}\lambda y^{(i)}\boldsymbol{x}^{(i)} - \mathop{\mathbf{E}}_{(x,y)\sim D}[\lambda y\boldsymbol{x}]\right\|_2 \le \epsilon/2 .$$

Putting everything together and by the union bound, we have that for $N > \max(Bd^3Q^2(\beta/2)^{-2k}(d+k)^{3k+1}/(4\epsilon^2), 4d^3\lambda^2/\epsilon^2)$, with probability $3/4$, we have that

$$\left\|\frac{1}{N}\sum_{i=1}^{N} g_t(\boldsymbol{x}^{(i)})y^{(i)}\boldsymbol{x}^{(i)} - \mathop{\mathbf{E}}_{(x,y)\sim D}[p_t(\boldsymbol{x})y\boldsymbol{x}]\right\|_2 \le \left\|\frac{1}{N}\sum_{i=1}^{N} p_t(\boldsymbol{x}^{(i)})y^{(i)}\boldsymbol{x}^{(i)} - \mathop{\mathbf{E}}_{(x,y)\sim D}[g_t(\boldsymbol{x})y\boldsymbol{x}]\right\|$$
$$+ \left\|\frac{1}{N}\sum_{i=1}^{N}\lambda y^{(i)}\boldsymbol{x}^{(i)} - \mathop{\mathbf{E}}_{(x,y)\sim D}[\lambda y\boldsymbol{x}]\right\|_2 \le \epsilon/2 + \epsilon/2 = \epsilon .$$

To amplify the confidence probability to $1 - \delta$, we can use the above empirical estimate ℓ times to obtain estimates $\widetilde{\boldsymbol{M}_j}^{(1)}, \ldots, \widetilde{\boldsymbol{M}_j}^{(\ell)}$ for all $j \in [d]$ and keep the median as our final estimate. It follows that $\ell = O(\log(d/\delta))$ repetitions suffice to guarantee confidence probability at least $1 - \delta$.

To prove the second statement, from Equation (B.1), we have that with probability $1 - \delta$

$$\left\|\nabla_{w}\hat{\ell}_t\right\|_2 \le \|\nabla_{w}\ell_t\|_2 + \epsilon \le d^{O(k)} + \epsilon = d^{O(k)} \ ,$$

where we used Theorem 3.4. This completes the proof. □

C Appendix to Chapter 4

Proof of Claim 4.15

Proof of Claim 4.15. To bound from below the expectation $I_{2,2}$, we use the fact that the distribution is $(2, L, R, \beta)$-well-behaved. For $I_{1,2}^{R/2}$, we have

$$
\begin{aligned}
I_{2,2} = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\mathbb{1}_{B_3^{R/2}}(\boldsymbol{x})\zeta(\boldsymbol{x})|\boldsymbol{x}_1|\right] &= \int_{B_3^{R/2}} |\boldsymbol{x}_1|\zeta(\boldsymbol{x})\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\geq \int_0^{R/\sqrt{2}}\int_{R/(2\sqrt{2})}^{R/\sqrt{2}} \boldsymbol{x}_1\zeta(\boldsymbol{x}_1,\boldsymbol{x}_2)\gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{x}_1\boldsymbol{x}_2 \\
&\geq \frac{R}{2\sqrt{2}}\int_{R/2}^{R/\sqrt{2}}\int_{R/(2\sqrt{2})}^{R/\sqrt{2}} \zeta(\boldsymbol{x}_1,\boldsymbol{x}_2)\gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{x}_1\boldsymbol{x}_2 \\
&\geq \frac{R}{2\sqrt{2}}C_\alpha^A\left(\int_{R/2}^{R/\sqrt{2}}\int_{R/(2\sqrt{2})}^{R/\sqrt{2}} \gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{x}_1\boldsymbol{x}_2\right)^{1/\alpha} \\
&\geq \frac{R}{4}C_\alpha^A\left(\frac{R^2L}{16}\right)^{1/\alpha} ,
\end{aligned}
$$

where we used Lemma C.7, and we bound from below the integral by a smaller square region, i.e., $[R/2, R/\sqrt{2}] \times [R/(2\sqrt{2}), R/\sqrt{2}]$. For $I_{2,2}$, we have

$$
\begin{aligned}
I_{1,2}^{R/2} = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\mathbb{1}_{B_3^{R/2}}(\boldsymbol{x})\zeta(\boldsymbol{x})\right] &= \int_{B_3^{R/2}} \zeta(\boldsymbol{x})\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\geq C_\alpha^A\left(\int_{B_3^{R/2}}\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^{1/\alpha} \\
&\geq C_\alpha^A\left(\int_0^{R/\sqrt{2}}\int_{R/(2\sqrt{2})}^{R/\sqrt{2}} \gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{x}_1\boldsymbol{x}_2\right)^{1/\alpha} \\
&\geq C_\alpha^A\left(\frac{R^2L}{4}\right)^{1/\alpha} ,
\end{aligned}
$$

where we used Lemma C.7. Thus,

$$I_{1,2}^{R/2} \geq C_\alpha^A \left(\frac{R^2 L}{4}\right)^{1/\alpha} = (RL/A)^{O(1/\alpha)} \quad \text{and} \quad I_{2,2} \geq (RL/A)^{O(1/\alpha)} .$$

This completes the proof of Claim 4.15. □

Proof of Claim 4.16

Proof of Claim 4.16. Recall that $\xi(\boldsymbol{x}_2) = \boldsymbol{x}_2/\tan\theta + b/\sin\theta$. We have that

$$I_1^{R/2} \leq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\mathbb{1}_{B_1^{R/2}}(\boldsymbol{x})\zeta(\boldsymbol{x})\right] - I_{1,2}^{R/2} \leq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\mathbb{1}_{B_1^{R/2}}(\boldsymbol{x})\zeta(\boldsymbol{x})\right] - \Gamma/2 .$$

We can bound from below the first term as follows

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\mathbb{1}_{B_1^{R/2}}(\boldsymbol{x})\zeta(\boldsymbol{x})\right] &\leq \int_{-R}^{-R/2}\int_{-\infty}^{\xi(\boldsymbol{x}_2)} \gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2 \\ &\leq \int_{-R}^{-R/2}\int_{-\infty}^{\xi(-R)} \gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2 \\ &\leq \mathbf{Pr}[\boldsymbol{x}_2 \geq |\xi(-R)|] \\ &\leq \exp(1-|\xi(-R)|/\beta) .\end{aligned}$$

Note that $|\xi(-R)| = (R\cos\theta - b)/\sin\theta \geq 3b/\sin\theta$, thus using the assumption $\theta < b\Gamma/(4\beta)$, we obtain $\exp(1-|\xi(-R)|/\beta) \leq \Gamma/4$, and therefore $I_1^{R/2} \leq -\Gamma/4$, completing the proof of Claim 4.16. □

Proof of Lemma 4.20

We start with a useful fact about the sub-exponential random variables.

Fact C.1 (see, e.g., Corollary of Proposition 2.7.1 in Vershynin (2018a))**.** *Let X be sub-exponential random variable with tail parameter β. For any function $f : \mathbb{R} \mapsto \mathbb{R}$, the random variable $Xf(X) - \mathbf{E}[Xf(X)]$ is zero mean sub-exponential with tail parameter $O(\beta \sup|f|)$.*

Using Fact C.1, we can bound from above the sample complexity needed to construct $\widehat{D}$.

Proof of Lemma 4.20. Let $\hat{\boldsymbol{g}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{B^{R/2}}\left(\boldsymbol{x}^{(i)}\right) y^{(i)}\boldsymbol{x}^{(i)}$. For any $\boldsymbol{u} \in \mathbb{R}^d$, we have that

$$|\boldsymbol{u} \cdot \boldsymbol{g}| \leq \mathop{\mathbf{E}}_{x\sim D_x}[|\boldsymbol{u} \cdot \boldsymbol{x}|] = \int_0^\infty \mathop{\mathbf{Pr}}_{x\sim D_x}[|\boldsymbol{u} \cdot \boldsymbol{x}| \geq t]\mathrm{d}t \leq \int_0^\infty \exp(1 - t/\beta)\mathrm{d}t = e\beta \,, \tag{C.1}$$

thus $\|\boldsymbol{g}\|_2 \leq e\beta$. Next we prove that the random variable $X = \mathbb{1}_{B^{R/2}}(\boldsymbol{x})y\boldsymbol{x} - \boldsymbol{g}$ is zero-mean with sub-exponential tails. First, we clearly have that $\mathbf{E}[X] = 0$. Using Fact C.1, it follows that X is sub-exponential with tail parameter $\beta' = O(\beta)$. We will now use the following Bernstein-type inequality.

Fact C.2. *Let $X_1, X_2, \ldots, X_N$ be independent zero-mean sub-exponential random variables with tail parameter $\beta \geq 1$. There exists an absolute constant $c > 0$ such that for every $\epsilon > 0$ we have*

$$\mathbf{Pr}\left[\left|\sum_{i=1}^{N} X_i\right| \geq \epsilon N\right] \leq 2\exp\left(-cN\epsilon^2/\beta^2\right) \,.$$

Using Fact C.2, we have that for every $1 \leq j \leq d$ it holds

$$\mathbf{Pr}\left[|\hat{g}_j - g_j| \geq \epsilon/\sqrt{d}\right] \leq 2\exp\left(-cN\epsilon^2/\left(d\beta'^2\right)\right) \,.$$

Thus, taking $N = O\left((d\beta^2/\epsilon^2)\log(d/\delta)\right)$, we get that $\|\hat{\boldsymbol{g}} - \boldsymbol{g}\|_2 \leq \epsilon$ with probability $1 - \delta$. For the second statement, using the triangle inequality and Equation (C.1) the result follows. □

Proof of Lemma 4.21

The proof requires a couple of known probabilistic facts. The first one is the bounded-difference inequality.

Fact C.3 (see, e.g., Theorem 2.2 of Devroye and Lugosi (2001)). *Let $X_1, \ldots, X_d \in \mathcal{X}$ be independent random variables and let $f : \mathcal{X}^d \mapsto \mathbb{R}$. Let $c_1, \ldots, c_d$ satisfy*

$$\sup_{x_1,\ldots,x_d,x_i'} \left| f(x_1, \ldots, x_i, \ldots, x_d) - f(x_1, \ldots, x_i', \ldots, x_d) \right| \le c_i$$

for $i \in [d]$. Then we have that $\mathbf{Pr}\left[f(X) - \mathbf{E}[f(X)] \ge t\right] \le \exp\left(-2t^2 / \sum_{i=1}^{d} c_i^2 \right)$.

We additionally require the symmetrization of the empirical distribution.

Fact C.4 (see, e.g., Exercise 8.3.24 of Vershynin (2018a)). *Let $\mathcal{F}$ be a class of measurable real-valued functions. Let $X_1, \ldots, X_N$ be N i.i.d. samples from a distribution D. Then*

$$\mathbf{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} f(X_i) - \mathbf{E}[f(X)] \right|\right] \le 2\,\mathbf{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i f(X_i) \right|\right] ,$$

where the ϵ_i's are independent Rademacher random variables.

The last fact we need connects the symmetrization with the VC dimension.

Definition C.5 (VC dimension). *A collection of sets $\mathcal{F}$ is said to* shatter *a set S if for all $S' \subseteq S$, there is an $F \in \mathcal{F}$ so that $F \cap S = S'$. The VC dimension of $\mathcal{F}$, denoted $\mathrm{VC}(\mathcal{F})$, is the largest n for which there exists an S with $|S| = n$ such that $\mathcal{F}$ shatters S.*

We note that a collection of sets $\mathcal{F}$ over a ground set is equivalent to a class of Boolean-valued functions on the same ground set. With this terminology, we have the following fact.

Fact C.6 (VC Inequality, see, e.g., Devroye and Lugosi (2001) or Theorem 8.3.3 in Vershynin (2018a)). *Let $\mathcal{F}$ be a class of Boolean-valued functions with $\mathrm{VC}(\mathcal{F}) \ge 1$. Let $X_1, \ldots, X_N$ be N i.i.d. samples from a distribution D. Then*

$$\mathop{\mathbf{E}}_{\epsilon_i}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i f(X_i) \right|\right] \le C\sqrt{\mathrm{VC}(\mathcal{F})/N} ,$$

where $C > 0$ is an absolute constant and the ϵ_i's are independent Rademacher random variables.

We are ready to bound the sample complexity required to check if Algorithm 5 finds a certificate.

Proof of Lemma 4.21. The proof is a simple application of the VC inequality. In more detail, we first use the bounded-difference inequality and then, using the symmetrization, we can apply the VC inequality to obtain the desired result.

For $N = O(\log(1/\delta)/\epsilon^2)$, we apply Fact C.3 for the function

$$f((X_1, Y_1), \dots, (X_N, Y_N)) = \sup_{t \in \mathbb{R}_+} \left| \mathbf{E}_{(x,y)\sim D}[\mathbb{1}_{B^t}(\boldsymbol{x})\, y] - \frac{1}{N}\sum_{i=1}^{N}[\mathbb{1}_{B^t}(X_i)\, Y_i] \right| ,$$

noting that $c_i = 2/N$ for all $i \leq N$. Therefore, with probability at least $1 - \delta$, we have that

$$\begin{aligned} &\sup_{t \in \mathbb{R}_+} \left| \mathbf{E}_{(x,y)\sim D}[\mathbb{1}_{B^t}(\boldsymbol{x})\, y] - \frac{1}{N}\sum_{i=1}^{N}[\mathbb{1}_{B^t}(X_i)\, Y_i] \right| \\ &\leq \mathbf{E}\left[\sup_{t \in \mathbb{R}_+} \left| \mathbf{E}_{(x,y)\sim D}[\mathbb{1}_{B^t}(\boldsymbol{x})\, y] - \frac{1}{N}\sum_{i=1}^{N}[\mathbb{1}_{B^t}(X_i)\, Y_i] \right| \right] + \epsilon \, . \end{aligned}$$

Then, by Fact C.4, we have that

$$\begin{aligned} \mathbf{E}\left[\sup_{t \in \mathbb{R}_+} \left| \mathbf{E}_{(x,y)\sim D}[\mathbb{1}_{B^t}(\boldsymbol{x})\, y] - \frac{1}{N}\sum_{i=1}^{N}[\mathbb{1}_{B^t}(X_i)\, Y_i] \right| \right] &\leq 2 \mathbf{E}_{\epsilon_i}\left[\sup_{t \in \mathbb{R}_+} \left| \frac{1}{N}\sum_{i=1}^{N} \epsilon_i Y_i \mathbb{1}_{B^t}(X_i) \right| \right] \\ &= 2 \mathbf{E}_{\epsilon_i}\left[\sup_{t \in \mathbb{R}_+} \left| \frac{1}{N}\sum_{i=1}^{N} \epsilon_i \mathbb{1}_{B^t}(X_i) \right| \right] , \end{aligned}$$

where the last inequality follows from the fact that $Y_i\epsilon_i$ and ϵ_i have the same distribution (because ϵ_i and Y_i are independent). Finally, using the fact that the class of indicators of the form $\mathbb{1}\{x \leq t\}$ has VC dimension 1, Fact C.6 implies that

$$\mathop{\mathbf{E}}_{\epsilon_i}\left[\sup_{t\in\mathbb{R}_+}\left|\frac{1}{N}\sum_{i=1}^{N}\epsilon_i \mathbb{1}_{B^t}(X_i)\right|\right] = O(\sqrt{1/N}) = O(\epsilon)\ .$$

Putting everything together completes the proof. □

Useful Technical Lemma

We are going to use the following simple fact about Tsybakov noise that shows that large probability regions will also have large integral even if we weight the integral with the noise function $1-2\eta(\boldsymbol{x}) > 0$. Notice that larger noise $\eta(\boldsymbol{x})$ makes $1-2\eta(\boldsymbol{x})$ closer to 0, and therefore tends to reduce the probability mass of the regions where $\eta(\boldsymbol{x})$ is large. A similar lemma can be found in Tsybakov (2004).

Lemma C.7. *Let D be a distribution on $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (α, A)-Tsybakov noise condition. Then for every measurable set $S \subseteq \mathbb{R}^d$ it holds $\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})(1-2\eta(\boldsymbol{x}))] \geq C_\alpha^A \left(\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\mathbb{1}_S(\boldsymbol{x})]\right)^{\frac{1}{\alpha}}$, where $C_\alpha^A = \alpha\left(\frac{1-\alpha}{A}\right)^{\frac{1-\alpha}{\alpha}}$.*

See Diakonikolas et al. (2021b) for the simple proof.

C.1 Omitted Proofs from Section 4.4

Proof of Lemma 4.29

Proof of Lemma 4.29. For the first condition, the lemma follows from Lemma 4.20. For the second condition, let $\boldsymbol{X} = \mathbf{E}_{(x,y)\sim D}[y(\boldsymbol{x}\boldsymbol{x}^\intercal - \boldsymbol{I})]$ and $\widehat{\boldsymbol{X}} = \mathbf{E}_{(x,y)\sim \widehat{D}}[y(\boldsymbol{x}\boldsymbol{x}^\intercal - \boldsymbol{I})]$. We are going to bound the variance, so we can apply Chebyshev's inequality.

For $0 < i, j \leq d$, we have

$$\mathbf{Var}_{(x,y)\sim D}[\widehat{X}_{ij}] = \frac{1}{N}\mathbf{Var}_{(x,y)\sim D}[X_{ij}] \leq \frac{1}{N}\mathop{\mathbf{E}}_{(x,y)\sim D}[X_{ij}^2] = \frac{1}{N}\mathop{\mathbf{E}}_{(x,y)\sim D}[y^2(x_i x_j - 1)^2]$$
$$\leq \frac{2}{N}\left(\mathop{\mathbf{E}}_{x\sim D_x}[x_i^2 x_j^2] + 1\right) \leq \frac{2}{N}\left(\sqrt{\mathop{\mathbf{E}}_{x\sim D_x}[x_i^4]\mathop{\mathbf{E}}_{x\sim D_x}[x_j^4]} + 1\right) = O(1/N) ,$$

where the last inequality follows from the fact that the marginals of a log-concave density have sub-exponential tails. Thus, from Chebyshev's inequality, for $0 < i, j \leq d$, we have that

$$\mathop{\mathbf{Pr}}_{(x,y)\sim D}[|\widehat{X}_{ij} - X_{ij}| \geq \epsilon/d] = O\left(\frac{d^2}{\epsilon^2 N}\right) .$$

Choosing $N = O(d^4/\epsilon^2)$, we have that $\left\|X - \widehat{X}\right\|_F \leq \epsilon$ with high constant probability. This completes the proof. □

Proof of Claim 4.34

Proof of Claim 4.34. For notational convenience, let $D^\perp = D_{B_{x_0}}^{\mathrm{proj}_{w^\perp}}$. Fix any unit vector $u \in w^\perp$. Without loss of generality, we may assume that $w = e_1$ and $u = e_2$. Denote by $\gamma(x_1, x_2)$ the marginal density of D on the first two coordinates. We have that

$$\mathop{\mathbf{E}}_{x\sim D^\perp}[|x^\intercal u|] = \frac{1}{\mathbf{Pr}_D[B_{x_0}]}\int |x_2|\mathbb{1}\{x_0 \leq x_1 \leq x_0 + s'\}\gamma(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 .$$

From Fact 4.27, we have that $\gamma(x_1, x_2) \leq (1/c)\exp(-|x_2|/c)$, for some absolute constant $c > 0$. Therefore,

$$\frac{1}{\mathbf{Pr}_D[B_{x_0}]}\int_{-\infty}^{\infty}\int_{x_0}^{x_0+s'} |x_2|\gamma(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \leq \frac{s'}{c\,\mathbf{Pr}_D[B_{x_0}]}\int_{-\infty}^{\infty}|x_2|e^{-|x_2|/c}\mathrm{d}x_2 = O(1) ,$$

where we used that $x_0, x_0 + s'$ are sufficiently small and it holds $\mathbf{Pr}_D[B_{x_0}] = \Theta(s')$, see Fact 4.27.

We next bound the covariance. Pick a unit vector $\boldsymbol{u} \in \boldsymbol{w}^\perp$. Without loss of generality, we may assume that $\boldsymbol{u} = \boldsymbol{e}_2$. Let $\theta = \boldsymbol{e}_2^\intercal \mathbf{E}_{\boldsymbol{x}\sim D^\perp}[\boldsymbol{x}]$ be the projection of the mean of $D^\perp$ on the direction $\boldsymbol{e}_2$. To bound the maximum and minimum eigenvalues of the covariance matrix of $D^\perp$, we need to bound from above and below the following expectation:

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D^\perp}[(\boldsymbol{x}_2 - \theta)^2] = \frac{1}{\mathbf{Pr}_D[B_{x_0}]} \int_{-\infty}^{\infty} \int_{x_0}^{x_0+s'} (\boldsymbol{x}_2 - \theta)^2 \gamma(\boldsymbol{x}_1, \boldsymbol{x}_2) \mathrm{d}\boldsymbol{x}_1 \mathrm{d}\boldsymbol{x}_2 \,.$$

We first bound it from below. Using again Fact 4.27 we know that, for the same absolute constant c as above, it holds that $\gamma(\boldsymbol{x}_1, \boldsymbol{x}_2) \geq c$ for points with distance smaller than c from the origin. Therefore,

$$\begin{aligned} &\frac{1}{\mathbf{Pr}_D[B_{x_0}]} \int_{-\infty}^{\infty} \int_{x_0}^{x_0+s'} (\boldsymbol{x}_2 - \theta)^2 \gamma(\boldsymbol{x}_1, \boldsymbol{x}_2) \mathrm{d}\boldsymbol{x}_1 \mathrm{d}\boldsymbol{x}_2 \\ &\geq \frac{c}{\mathbf{Pr}_D[B_{x_0}]} \int_{-c/\sqrt{2}}^{c/\sqrt{2}} (\boldsymbol{x}_2 - \theta)^2 \mathrm{d}\boldsymbol{x}_2 \int_{x_0}^{x_0+s'} \mathrm{d}\boldsymbol{x}_1 = \Omega(1) \,, \end{aligned}$$

where we used again the fact that $\mathbf{Pr}_D[B_{x_0}] = \Theta(s')$ and also picked the worst case θ to minimize the above expression, i.e., $\theta = 0$. We next bound the covariance eigenvalues from above. Using again the fact that $\gamma(\boldsymbol{x}_1, \boldsymbol{x}_2) \leq c \exp(-c|\boldsymbol{x}_2|)$ for some absolute constant $c > 0$, we compute

$$\begin{aligned} &\frac{1}{\mathbf{Pr}_D[B_{x_0}]} \int_{-\infty}^{\infty} \int_{x_0}^{x_0+s'} (\boldsymbol{x}_2 - \theta)^2 \gamma(\boldsymbol{x}_1, \boldsymbol{x}_2) \mathrm{d}\boldsymbol{x}_1 \mathrm{d}\boldsymbol{x}_2 \\ &\leq \frac{1}{c\,\mathbf{Pr}_D[B_{x_0}]} \int_{-\infty}^{\infty} \int_{x_0}^{x_0+s'} (\boldsymbol{x}_2 - \theta)^2 e^{-|\boldsymbol{x}_2|/c} \mathrm{d}\boldsymbol{x}_1 \mathrm{d}\boldsymbol{x}_2 = O(1) \,, \end{aligned}$$

where we used the fact that $\theta = O(1)$, as already shown above, and that $\mathbf{Pr}_D[B_{x_0}] = \Theta(s')$. This completes the proof. □

Proof of Claim 4.38

Proof of Claim 4.38. To prove that F is L-smooth, we need to show that $\sup_{\|\boldsymbol{r}\|_2 \le R} \left\|\nabla^2 F(\boldsymbol{r})\right\|_2 \le L$, for some $L > 0$. We have

$$\begin{aligned} G(\boldsymbol{r}) := \nabla F(\boldsymbol{r}) &= -2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x}\boldsymbol{x}^\intercal \mathbb{1}\{\boldsymbol{r}\cdot\boldsymbol{x} \ge 0\} e^{-\boldsymbol{r}\cdot\boldsymbol{x}}] \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x} \min(1, e^{-\boldsymbol{r}\cdot\boldsymbol{x}})] \\ &= -2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x}\boldsymbol{x}^\intercal g_1(\boldsymbol{r}^\intercal\boldsymbol{x})] \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x} g_2(\boldsymbol{r}^\intercal\boldsymbol{x})] \,, \end{aligned}$$

where $g_1(t) = \mathbb{1}\{t \ge 0\}e^{-t}$ and $g_2(t) = \min(1, e^{-t})$. Using the product rule, we obtain that the derivative of $G(\boldsymbol{r})$ at $\boldsymbol{r}$, $DG|_{\boldsymbol{r}}$, is the following linear function from $\mathbb{R}^d$ to $\mathbb{R}^d$:

$$DG|_{\boldsymbol{r}}\boldsymbol{h} = -2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x}\boldsymbol{x}^\intercal g_1'(\boldsymbol{x}^\intercal\boldsymbol{r})\boldsymbol{x}^\intercal\boldsymbol{h}] \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x} g_2(\boldsymbol{x}^\intercal\boldsymbol{r})] - 2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x}\boldsymbol{x}^\intercal g_1(\boldsymbol{x}^\intercal\boldsymbol{r})] \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x} g_2'(\boldsymbol{x}^\intercal\boldsymbol{r})\boldsymbol{x}^\intercal\boldsymbol{h}] \,,$$

where $g_1'(t) = \delta(t)e^{-t} - \mathbb{1}\{t \ge 0\}e^{-t}$ (here by δ we denote the Dirac delta function), and $g_2'(t) = -\mathbb{1}\{t \ge 0\}e^{-t}$. To show that F is smooth, we need to bound the operator norm of $DG|_{\boldsymbol{r}}$, i.e.,

$$\sup_{\boldsymbol{h}:\|\boldsymbol{h}\|_2=1} \|DG|_{\boldsymbol{r}}\boldsymbol{h}\|_2 \,.$$

Using the triangle and Cauchy-Schwarz inequalities, we can bound the first term as follows:

$$\begin{aligned} &\left\| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x}\boldsymbol{x}^\intercal g_1'(\boldsymbol{x}^\intercal\boldsymbol{r})\boldsymbol{x}^\intercal\boldsymbol{h}] \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x} g_2(\boldsymbol{x}^\intercal\boldsymbol{r})] \right\|_2 \\ &\le \left\| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x}\boldsymbol{x}^\intercal g_1'(\boldsymbol{x}^\intercal\boldsymbol{r})\boldsymbol{x}^\intercal\boldsymbol{h}] \right\|_2 \left\| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x} g_2(\boldsymbol{x}^\intercal\boldsymbol{r})] \right\|_2 \\ &\le \left\| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x}\boldsymbol{x}^\intercal\boldsymbol{x}^\intercal\boldsymbol{h}\, \delta(\boldsymbol{x}^\intercal\boldsymbol{r}) e^{-\boldsymbol{x}^\intercal\boldsymbol{r}}] \right\|_2 + \left\| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x}\boldsymbol{x}^\intercal\boldsymbol{x}^\intercal\boldsymbol{h}] \right\|_2 \left\| \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}} [\boldsymbol{x}] \right\|_2 \,. \end{aligned}$$

We will first handle the term $\|\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\boldsymbol{x}\boldsymbol{x}^\intercal \boldsymbol{x}^\intercal \boldsymbol{h}\,\delta(\boldsymbol{x}^\intercal \boldsymbol{r})]\|_2$. To simplify notation, we may set without loss of generality $\boldsymbol{r} = \boldsymbol{e}_1$. We have

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\boldsymbol{x}\boldsymbol{x}^\intercal \boldsymbol{x}^\intercal \boldsymbol{h}\,\delta(\boldsymbol{x}^\intercal \boldsymbol{r})e^{-\boldsymbol{x}^\intercal \boldsymbol{r}}] = \mathop{\mathbf{E}}_{\boldsymbol{x}'\sim D'_{\boldsymbol{x}}}[\boldsymbol{x}'(\boldsymbol{x}')^\intercal(\boldsymbol{x}')^\intercal \boldsymbol{h}\gamma(0)]\ ,$$

where $D'_{\boldsymbol{x}}$ is the distribution $D_{\boldsymbol{x}}$ conditioned on $\boldsymbol{x}_1 = 0$, and $\gamma(0)$ is the one-dimensional p.d.f. at point 0 (which is bounded by a universal constant for log-concave distributions). Note that $D'_{\boldsymbol{x}}$ is still log-concave.

Since $D_{\boldsymbol{x}}$ is $(O(1), O(1))$-isotropic, it holds

$$\left\|\mathop{\mathbf{E}}_{\boldsymbol{x}'\sim D'_{\boldsymbol{x}}}[\boldsymbol{x}'(\boldsymbol{x}')^\intercal(\boldsymbol{x}')^\intercal \boldsymbol{h}\right\|_2 \le \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D'_{\boldsymbol{x}}}[\|\boldsymbol{x}'(\boldsymbol{x}')^\intercal(\boldsymbol{x}')^\intercal \boldsymbol{h}\|_2] \le \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D'_{\boldsymbol{x}}}[\|\boldsymbol{x}'\|_2^3] \le \mathrm{poly}(d)\ ,$$

where we used that $\|\boldsymbol{A}\boldsymbol{B}\|_2 \le \|\boldsymbol{A}\|_2\,\|\boldsymbol{B}\|_2$, and that $\|\boldsymbol{h}\|_2 = 1$. Similarly, $\|\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\boldsymbol{x}\boldsymbol{x}^\intercal \boldsymbol{x}^\intercal \boldsymbol{h}]\|_2 \le \mathrm{poly}(d)$. Finally,

$$\left\|\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\boldsymbol{x}]\right\|_2 \le \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[\|\boldsymbol{x}\|_2] \le \mathrm{poly}(d)\ .$$

Putting everything together, we get that $L = \mathrm{poly}(d)$, which completes the proof.

□

C.2 Omitted Proofs from Section 4.5

Proof of Proposition 4.44

We will require the following standard regret bound from online convex optimization.

Lemma C.8 (see, e.g., Theorem 3.1 of Hazan (2016)). *Let $\mathcal{V} \subseteq \mathbb{R}^n$ be a non-empty closed convex set with diameter K. Let $\ell_1, \ldots, \ell_T$ be a sequence of T convex functions $\ell_t : \mathcal{V} \mapsto \mathbb{R}$ differentiable in open sets containing $\mathcal{V}$, and let $G = \max_{t\in[T]} \|\nabla_{\boldsymbol{w}}\ell_t\|_2$. Pick any $\boldsymbol{w}^{(1)} \in \mathcal{V}$ and set $\eta_t = \frac{K}{G\sqrt{t}}$ for $t \in [T]$. Then, for all $\boldsymbol{u} \in \mathcal{V}$, we have that $\sum_{t=1}^T(\ell_t(\boldsymbol{w}^{(t)}) - \ell_t(\boldsymbol{u})) \le \frac{3}{2}GK\sqrt{T}$.*

For the set $\mathcal{B}$, i.e., the unit ball with respect the $\|\cdot\|_2$, the diameter K equals to 2. We will show that the optimal vector $\boldsymbol{w}^\star$ and our current candidate vector $\boldsymbol{w}^{(t)}$ have a separation in the value of ℓ_t. Since we do not have access to ℓ_t precisely, we need a function $\hat{\ell}_t$, which is close to ℓ_t with high probability. The following simple lemma gives us an efficient way to compute an approximation $\hat{\ell}_t$ of ℓ_t.

Lemma C.9 (Estimating the function ℓ_t)**.** *Let D be a $(3, L, R, \beta)$-well-behaved distribution and $T_{\boldsymbol{w}}(\boldsymbol{x})$ be the non-negative function given by a ρ-certificate oracle. Then after drawing $O(d\beta^2/\epsilon^2 \log(d/\delta))$ samples from D, with probability at least $1-\delta$, the empirical distribution $\widehat{D}$ satisfies the following conditions:*

- $\left|\mathbf{E}_{(\boldsymbol{x},y)\sim\widehat{D}}[(T_{\boldsymbol{w}}(\boldsymbol{x}) + \frac{\rho}{2})y\boldsymbol{u}\cdot\boldsymbol{x}] - \mathbf{E}_{(\boldsymbol{x},y)\sim D}[(T_{\boldsymbol{w}}(\boldsymbol{x}) + \frac{\rho}{2})y\boldsymbol{u}\cdot\boldsymbol{x}]\right| \leq \epsilon$*, for any* $\boldsymbol{u} \in \mathcal{B}$.
- $\left\|\mathbf{E}_{(\boldsymbol{x},y)\sim\widehat{D}}[(T_{\boldsymbol{w}}(\boldsymbol{x}) + \frac{\rho}{2})y\boldsymbol{x}]\right\|_2 \leq 1 + \frac{\rho}{2} + \epsilon$.

Proof. The proof of this lemma is similar to the proof of Lemma 4.20. Let $\hat{\boldsymbol{g}} = \mathbf{E}_{(\boldsymbol{x},y)\sim\widehat{D}}[(T_{\boldsymbol{w}}(\boldsymbol{x}) + \frac{\rho}{2})y\boldsymbol{x}]$ and $\boldsymbol{g} = \mathbf{E}_{(\boldsymbol{x},y)\sim D}[(T_{\boldsymbol{w}}(\boldsymbol{x}) + \frac{\rho}{2})y\boldsymbol{x}]$. For any unit vector $\boldsymbol{u}$, we have

$$|\boldsymbol{u}\cdot\boldsymbol{g}| \leq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|T_{\boldsymbol{w}}(\boldsymbol{x})||\boldsymbol{u}\cdot\boldsymbol{x}|] + \frac{\rho}{2}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|\boldsymbol{u}\cdot\boldsymbol{x}|] \leq 1 + \frac{\rho}{2}\,,$$

where we used that $|T(\boldsymbol{x})| \leq 1$ and that the distribution $D_{\boldsymbol{x}}$ is in isotropic position. Moreover, from Fact C.1, the random variable $X = (T_{\boldsymbol{w}}(\boldsymbol{x}) + \frac{\rho}{2})y\boldsymbol{x} - \boldsymbol{g}$ is sub-exponential with tail bound $\beta' = O(\beta)$. Thus, the rest of proof follows as in Lemma 4.20. □

The last item we need to proceed with our main proof is to establish that when the oracle $\mathcal{C}$ in Step 13 of Algorithm 8 returns a function $T_{\boldsymbol{w}^{(t)}}$, then there exists a function ℓ_t for which our current candidate vector $\boldsymbol{w}^{(t)}$ and the optimal vector $\boldsymbol{w}^\star$ are not close.

Lemma C.10 (Error of ℓ_t)**.** *Let $\boldsymbol{w}^{(t)} \in \mathcal{B}$ and $\boldsymbol{w}^\star$ be the optimal weight vector. For $g_t(\boldsymbol{x}) = -(T_{\boldsymbol{w}^{(t)}}(\boldsymbol{x}) + \frac{\rho}{2})$ and $\ell_t(\boldsymbol{w}) = \mathbf{E}_{(\boldsymbol{x},y)\sim D}[g_t(\boldsymbol{x})y\boldsymbol{x}\cdot\boldsymbol{w}]$, where $T_{\boldsymbol{w}^{(t)}}(\boldsymbol{x})$ is the* *function given by a ρ-certificate oracle, we have that*

$$\ell_t\left(\boldsymbol{w}^\star\right) \le -\rho\alpha\left(\frac{R\,L}{A}\right)^{O(1/\alpha)} \quad \text{and} \quad \ell_t(\boldsymbol{w}^{(t)}) \ge \left\|\boldsymbol{w}^{(t)}\right\|_2 \frac{\rho}{2}\,.$$

Proof. Without loss of generality, let $\boldsymbol{w}^\star = \boldsymbol{e}_1$. From Fact 3.3 and the definition of $\eta(\boldsymbol{x})$, we have that for every $t \in [T]$, it holds $\ell_t(\boldsymbol{w}^\star) \le -\lambda\,\mathbf{E}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|\boldsymbol{w}^\star\cdot\boldsymbol{x}|(1-2\eta(\boldsymbol{x}))]$. To bound from above this expectation, we use the (L,R,B,β)-bound properties. We have that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}[|\boldsymbol{w}^\star\cdot\boldsymbol{x}|(1-2\eta(\boldsymbol{x}))] \ge \frac{R}{4}C_\alpha^A\left(\frac{R^3\,L}{2}\right)^{1/\alpha}\,,$$

where in the last inequality we used Lemma C.7. Therefore, $\ell_t\left(\boldsymbol{w}^\star\right) \le -\frac{\rho}{2}\,\frac{R}{4}C_\alpha^A\left(\frac{R^3\,L}{2}\right)^{1/\alpha}$. Then we bound from below $\ell_t(\boldsymbol{w}^{(t)})$ as follows

$$\begin{aligned}\ell_t(\boldsymbol{w}^{(t)}) &= -\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}\left[(T_{\boldsymbol{w}^{(t)}}\,(\boldsymbol{x})+\lambda)\boldsymbol{w}^{(t)}\cdot\boldsymbol{x}y\right] \ge \left\|\boldsymbol{w}^{(t)}\right\|_2\rho - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\frac{\rho}{2}\boldsymbol{w}^{(t)}\cdot\boldsymbol{x}y\right]\\ &\ge \left\|\boldsymbol{w}^{(t)}\right\|_2\rho - \frac{\rho}{2}\sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim D_{\boldsymbol{x}}}\left[\boldsymbol{w}^{(t)}\cdot\boldsymbol{x}^2\right]} \ge \left\|\boldsymbol{w}^{(t)}\right\|_2\frac{\rho}{2}\,,\end{aligned}$$

where we used the Cauchy-Schwarz inequality and the fact that $\boldsymbol{x}$ is in isotropic position. □

We are ready to prove Proposition 4.44.

Proof of Proposition 4.44. Let $G = \alpha\left(\frac{R\,L}{A}\right)^{O(1/\alpha)}$. Assume, in order to reach a contradiction, that for all steps $t \in [T]$ it holds that $\theta\left(\boldsymbol{w}^{(t)},\boldsymbol{w}^\star\right) \ge \epsilon$. For each step t, let $T_{\boldsymbol{w}^{(t)}}(\boldsymbol{x})$ be the non-negative function output by the oracle $\mathcal{C}(\boldsymbol{w}^{(t)},\epsilon,\delta/T)$. Note that

$$\mathop{\mathbf{E}}_{(\boldsymbol{x},y)\sim D}[T_{\boldsymbol{w}^{(t)}}(\boldsymbol{x})y\boldsymbol{w}^{(t)}\cdot\boldsymbol{x}] \le -\left\|\boldsymbol{w}^{(t)}\right\|_2\frac{\rho}{2}\,.$$

Let $\hat{\ell}_t(\boldsymbol{w})$ be the empirical estimator of $\ell_t\left(\boldsymbol{w}\right) = \mathbf{E}[\hat{\ell}_t(\boldsymbol{w})] = -\mathbf{E}_{(\boldsymbol{x},y)\sim D}[\left(T_{\boldsymbol{w}^{(t)}}(\boldsymbol{x})+\frac{\rho}{2}\right)y\boldsymbol{x}\cdot$

$\boldsymbol{w}]$. Using Lemma C.9, for $N = O\left(\frac{d\beta^2}{\rho^2 G^2}\log\left(\frac{T}{\delta}\right)\right)$ samples, we have that

$$\mathbf{Pr}\left[|\hat{\ell}_t(\boldsymbol{w}^{(t)}) - \ell_t(\boldsymbol{w}^{(t)})| \geq \frac{1}{4}G\rho\right] \leq \frac{\delta}{2T}$$

and $\mathbf{Pr}\left[|\hat{\ell}_t(\boldsymbol{w}^\star) - \ell_t(\boldsymbol{w}^\star)| \geq \frac{1}{4}G\rho\right] \leq \frac{\delta}{2T}$.

From Lemma C.10, for every step t, we have that $\ell_t(\boldsymbol{w}^{(t)}) \geq \frac{1}{2}\left\|\boldsymbol{w}^{(t)}\right\|_2 \rho \geq 0$ and $\ell_t\left(\boldsymbol{w}^\star\right) \leq -\rho G$, thus, with probability at least $1 - \frac{\delta}{T}$, $\hat{\ell}_t(\boldsymbol{w}^{(t)}) \geq -\frac{1}{4}G\rho$ and $\hat{\ell}_t\left(\boldsymbol{w}^\star\right) \leq -\frac{3}{4}G\rho$. Using Lemma C.8, we get

$$\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\ell}_t\left(\boldsymbol{w}^{(t)}\right) - \hat{\ell}_t\left(\boldsymbol{w}^*\right)\right) \leq \frac{1 + \frac{\rho}{2} + \frac{1}{4}\rho G}{\sqrt{T}} \,.$$

By the union bound, it follows that with probability at least $1 - \delta$, we have that

$$\frac{1}{2}G\rho \leq \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\ell}_t\left(\boldsymbol{w}^{(t)}\right) - \hat{\ell}_t\left(\boldsymbol{w}^\star\right)\right) \leq \frac{4}{\sqrt{T}} \,,$$

which leads to a contradiction for $T = \frac{16}{(\rho G)^2}$.

Thus, either there exists $t \in [T]$ such that $\theta\left(\boldsymbol{w}^{(t)}, \boldsymbol{w}^\star\right) < \epsilon$, which the algorithm returns in Step 15, or the oracle $\mathcal{C}$ did not provide a correct certificate, which happens with probability at most δ. Moreover, the algorithm calls the certificate T times and the number of samples needed to construct the empirical distribution $\widehat{D}$ is

$$O(T\,N) = \frac{d\beta^2}{\rho^4}\log\left(\frac{1}{\delta\rho}\right)\frac{1}{\alpha}\left(\frac{A}{R\,L}\right)^{O(1/\alpha)} \,.$$

This completes the proof. □

Using Proposition 4.44 and our certificate algorithms, we obtain the following parameter estimation result for halfspaces with Tsybakov noise.

Theorem C.11 (Parameter Estimation of Tsybakov Halfspaces Under Well-Behaved Distributions). *Let D be a $(3, L, R, \beta)$-well-behaved isotropic distribution on*

$\mathbb{R}^d \times \{\pm 1\}$ *that satisfies the* (α, A)*-Tsybakov noise condition with respect to an unknown halfspace* $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^\star \cdot \boldsymbol{x})$*. There exists an algorithm that draws* $N = \beta^4 \left(\frac{dA}{RL\epsilon}\right)^{O(1/\alpha)} \log(1/\delta)$ *samples from* D*, runs in* $\mathrm{poly}(N, d)$ *time, and computes a vector* $\widehat{\boldsymbol{w}}$ *such that with probability* $1 - \delta$ *we have* $\theta(\widehat{\boldsymbol{w}}, \boldsymbol{w}^\star) \leq \epsilon$*.*

We note here that Theorem C.11 does not require the "U bounded" condition of the underlying distribution on examples that is required in our Theorem 4.39. Recall that this condition corresponds to an anti-concentration property of the data distribution. With this additional property, Theorem 4.39 follows easily from Theorem C.11, since it allows us to translate the small angle guarantee of Theorem C.11 to the zero-one loss.

Proof of Theorem C.11. We start by noting how to obtain a ρ-certificate oracle for $(3, L, R, \beta)$-well-behaved distributions. The algorithm of Theorem 4.5, returns a function $T_{\boldsymbol{w}}$ such that $\mathbf{E}_{(\boldsymbol{x},y)\sim D}\left[T_{\boldsymbol{w}}(\boldsymbol{x}) y \boldsymbol{w} \cdot \boldsymbol{x}\right] \leq -\frac{1}{\beta}\left(\theta L R/(dA)\right)^{O(1/\alpha)}$. By definition, the function $T_{\boldsymbol{w}}$ (i.e., Equation (4.3)) is bounded, namely $\|T_{\boldsymbol{w}}\|_\infty \leq \frac{1}{\min_{x \in B} |\boldsymbol{w} \cdot \boldsymbol{x}|} \leq O\left(\frac{d}{\theta}\right)$, where B is the band from Equation (4.3). Therefore, the function $T_{\boldsymbol{w}} / \|T_{\boldsymbol{w}}\|_\infty$ satisfies the conditions of a ρ-certificate oracle. Thus, by scaling the output of the algorithm of Theorem 4.23, we obtain a $\frac{1}{\beta}\left(\theta L R/(dA)\right)^{O(1/\alpha)}$-certificate oracle. From Proposition 4.44, this gives us an algorithm that returns a vector $\widehat{\boldsymbol{w}}$ such that $\theta(\widehat{\boldsymbol{w}}, \boldsymbol{w}^\star) \leq \epsilon$ with probability $1 - \delta$. □

To prove Theorem 4.39, we need the following claim for (L, R, U, B, β)-well-behaved distributions.

Claim C.12 (see, e.g., Claim 2.1 of Diakonikolas et al. (2020c))**.** *Let* $D_{\boldsymbol{x}}$ *be an* (L, R, U, B, β)*-well-behaved distribution on* $\mathbb{R}^d$*. Then, for any* $0 < \epsilon \leq 1$*, we have that* $\mathrm{err}_{0-1}^{D_x}(h_{\boldsymbol{u}}, h_{\boldsymbol{v}}) \leq U\beta^2 \log^2(1/\epsilon) \cdot \theta(\boldsymbol{v}, \boldsymbol{u}) + \epsilon$.

Proof of Theorem 4.39. Running Algorithm 8 for $\epsilon' = \frac{\epsilon}{2U\beta^2} \frac{1}{\log^2(2/\epsilon)}$, by Theorem C.11, Algorithm 8 outputs a $\widehat{\boldsymbol{w}}$ such that $\theta(\widehat{\boldsymbol{w}}, \boldsymbol{w}^\star) \leq \frac{\epsilon}{2U\beta^2} \frac{1}{2\log^2(1/\epsilon)}$, then from Claim C.12, we have $\mathrm{err}_{0-1}(h_{\widehat{\boldsymbol{w}}}, f) \leq \epsilon$. □

D Appendix to Chapter 5

D.1 Learning LSFs with Bounded Noise in Kendall's Tau distance

Improperly Learning LSFs with Bounded Noise

We provide an improper learner for LSFs in the presence of bounded noise. We first restate the main result of this section, whose proof relies on a connection between noisy linear label ranking distributions and the Massart noise model.

Theorem D.1 (Non-Proper Learning Algorithm)**.** *Fix $\eta \in [0, 1/2)$ and $\epsilon, \delta \in (0,1)$. Let $\mathcal{D}$ be an η-noisy linear label ranking distribution satisfying the assumptions of Definition 1.10.* `ImproperLSF` *(Algorithm 9) draws $N = \widetilde{O}\left(\frac{d}{\epsilon(1-2\eta)^6} \log(k/\delta)\right)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(d, k, 1/\epsilon, \log(1/\delta))$ time and, with probability at least $1-\delta$, outputs a hypothesis $h : \mathbb{R}^d \to \mathbb{S}_k$ that is ϵ-close in KT distance to the target.*

Proof. Assume that the target function is $\sigma^\star(\boldsymbol{x}) = \sigma_{\boldsymbol{W}^\star}(\boldsymbol{x}) = \mathrm{argsort}(\boldsymbol{W}^\star \boldsymbol{x})$ for some unknown matrix $\boldsymbol{W}^\star \in \mathbb{R}^{k \times d}$. Consider a collection of N i.i.d. samples from an η-noisy linear label ranking distribution $\mathcal{D}$ (see Definition 1.10) and let T be the associated training set. For each example $(\boldsymbol{x}, \pi) \in T$, we create a list of $\binom{k}{2}$ binary examples $(\boldsymbol{x}, y_{ij})$ with $y_{ij} = \mathrm{sign}(\pi(i) - \pi(j))$ for any $1 \le i < j \le k$, where $\pi(i)$ denotes the position of the element i. Hence, we create the datasets T_{ij} consisting of the binary labeled examples $(\boldsymbol{x}, y_{ij})$. We have that

$$\mathop{\mathbf{Pr}}_{(\boldsymbol{x},\pi)\sim\mathcal{D}}\left[y_{ij} \cdot \mathrm{sign}((\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \cdot \boldsymbol{x}) < 0 \mid \boldsymbol{x}\right] = \mathop{\mathbf{Pr}}_{\pi\sim\mathcal{M}(\sigma^\star(\boldsymbol{x}))}\left[\pi(i) < \pi(j) \mid \boldsymbol{W}_i^\star \cdot \boldsymbol{x} < \boldsymbol{W}_j^\star \cdot \boldsymbol{x}\right] .$$

Since $\mathcal{M}(\sigma^\star(\boldsymbol{x}))$ is an η-bounded noise ranking distribution (see Definition 1.9), we get that

$$\mathop{\mathbf{Pr}}_{\pi\sim\mathcal{M}(\sigma^\star(\boldsymbol{x}))}\left[\pi(i) < \pi(j) \mid \sigma^\star(\boldsymbol{x})(i) > \sigma^\star(\boldsymbol{x})(j)\right] \le \eta < 1/2 ,$$

where $\sigma^\star(\boldsymbol{x})(i)$ denotes the position of the element i in the ranking $\sigma^\star(\boldsymbol{x})$. Focusing on the training set T_{ij}, we have that the sign y_{ij} is flipped with probability at most η. So, we have reduced the problem to $\binom{k}{2}$ sub-problems concerning the learnability of halfspaces in the presence of Massart noise. The Massart noise model is a special case of Definition 1.10 where $k = 2$. Note also that for each training set T_{ij}, the features $\boldsymbol{x}$ have the same distribution. We can now apply the following result for LTFs with Massart noise for the standard Gaussian distribution. Recall that the concept class of homogeneous halfspaces (or linear threshold functions) is $\mathcal{C}_{\mathrm{LTF}} = \{h_{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x}) : \boldsymbol{w} \in \mathbb{R}^d\}$.

Lemma D.2 (Learning Halfspaces with Massart noise Zhang et al. (2020b))**.** *Fix $\eta \in [0, 1/2)$ and let $\epsilon, \delta \in (0, 1)$. Let $\mathcal{D}$ be an η-noisy linear label ranking distribution satisfying the assumptions of Definition 1.10 with $k = 2$ (where $\mathcal{C}_{\mathrm{LSF}} = \mathcal{C}_{\mathrm{LTF}}$). There is a computationally efficient algorithm `MassartLTF` that draws $m = O(\frac{d\ \mathrm{polylog}(d)}{\epsilon(1-2\eta)^6} \cdot \log(1/\delta))$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(m)$ time and outputs a linear threshold function h that is ϵ-close to the target linear threshold function $h^\star$ with probability at least $1 - \delta$, i.e., it holds $\mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{N}_d}[h(\boldsymbol{x}) \neq h^\star(\boldsymbol{x})] \leq \epsilon$.*

We can invoke the algorithm of Lemma D.2 for any alternatives $1 \leq i < j \leq k$ with accuracy $\epsilon' = O(\epsilon)$, $\delta' = O(\delta/k^2)$ and error rate $\eta < 1/2$[1]. We remark that Lemma D.2 returns a halfspace. Each one of the $\binom{k}{2}$ calls will provide a vector $\boldsymbol{v}_{ij} \in \mathbb{R}^d$ such that, with probability at least $1 - \delta'$, it satisfies

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\mathrm{sign}(\boldsymbol{v}_{ij} \cdot \boldsymbol{x}) \neq \mathrm{sign}((\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \cdot \boldsymbol{x})] \leq \epsilon' ,$$

where the true target halfspace has normal vector $\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star$. Moreover, for any $i < j$, the algorithm requires that the training set T_{ij} is of size

$$|T_{ij}| = \Omega\left(\frac{d}{\epsilon'} \cdot \frac{1}{(1-2\eta)^6} \cdot \log(1/\delta')\right) ,$$

[1] We can assume that η is known without loss of generality.

and, so, a total number of

$$N = \Omega\left(\frac{d}{\epsilon} \cdot \frac{1}{(1-2\eta)^6} \cdot \log(k/\delta)\right) ,$$

samples (x, π) is required from the distribution $\mathcal{D}$. Given a collection of linear classifiers with normal vectors v_{ij} for any $i < j$, it remains to aggregate them and compute a sorting function $h : \mathbb{R}^d \to \mathbb{S}_k$. To this end, the estimator h, given an example x, creates the directed complete graph G with k nodes with directed edge $i \to j$ if $v_{ij} \cdot x > 0$. If all the linear classifiers are correct (which occurs with probability $1 - O(\epsilon k^2)$ over $\mathcal{D}_x$ due to the union bound), the graph G is acyclic (since it will match the true directions induced by $W^\star$) and the estimator h outputs the induced permutation. Observe that the KT distance is

$$\frac{1}{\binom{k}{2}} \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{N}_d}\left[\sum_{1 \le i < j \le k} \mathbb{1}\{\mathrm{sign}(v_{ij} \cdot x) \neq \mathrm{sign}((W_i^\star - W_j^\star) \cdot x)\}\right] \le \epsilon' .$$

Otherwise, the classifiers are inconsistent and G contains cycles. So, the expected number of mistakes in the graph G is ϵk^2. The estimator in order to output a ranking uses a deterministic constant approximation algorithm for the minimum Feedback Arc Set Ailon et al. (2008) in order to remove the cycles. For an overview of this fundamental line of research, we refer to Ailon et al. (2008); Van Zuylen and Williamson (2009); Kenyon-Mathieu and Schudy (2006).

Lemma D.3 (3-Approximation Algorithm for mimimum FAS (see Van Zuylen and Williamson (2009); Ailon et al. (2008))). *There is a deterministic algorithm* `MFAS` *for the minimum Feedback Arc Set on unweighted tournaments with k vertices that outputs orderings with cost less than* $3 \cdot \mathrm{OPT}$. *The running time is* $\mathrm{poly}(k)$.

In the above, OPT is the minimum number of flips the algorithm should perform. With input the cyclic directed graph G induced by the estimated linear classifiers, the algorithm of Lemma D.3 computes, in $\mathrm{poly}(k)$ time, a 3-approximation of the optimal solution (i.e., instead of correcting ϵ_0 directed edges, the algorithm will provide a directed acyclic graph with $3\epsilon_0$ changed edges). Hence, for the hypothesis $h : \mathbb{R}^d \to \mathbb{S}_k$, where $h(\boldsymbol{x})$ is the output of the minimum FAS approximation algorithm with input G (G depends on the input $\boldsymbol{x}$, the randomness of the samples and the internal randomness of the $\binom{k}{2}$ calls of the Massart linear classifiers), and the target function $\sigma^\star(\boldsymbol{x})$, we have that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_d}[\Delta_{KT}(h(\boldsymbol{x}), \sigma^\star(\boldsymbol{x}))] \leq (\epsilon' + 3\epsilon') = 4\epsilon' ,$$

which completes the proof, by setting $\epsilon' = \epsilon/4$. □

Remark D.4. *Consider the following variant of the above procedure: compute the $O(k^2)$ linear classifiers with accuracy $\epsilon' = \epsilon/k^2$: If the induced directed graph is acyclic, output the ranking; otherwise, output a random permutation. With probability ϵ, the KT distance will be of order k^2. Hence, one has to draw in total $O(k^4 d/\epsilon)$ samples to make the expected KT distance roughly $O(\epsilon)$. The algorithm of Theorem D.1 improves on this approach.*

The Proof of Theorem 5.1: Properly Learning LSFs with Bounded Noise

We first restate the main result of this section.

Theorem D.5 (Proper Learning Algorithm)**.** *Fix $\eta \in [0, 1/2)$ and $\epsilon, \delta \in (0, 1)$. Let $\mathcal{D}$ be an η-noisy linear label ranking distribution satisfying the assumptions of Definition 1.10.* `ProperLSF` *(Algorithm 10) draws $N = \widetilde{O}\left(\frac{d}{\epsilon(1-2\eta)^6}\log(k/\delta)\right)$ samples from $\mathcal{D}$, runs in $\mathrm{poly}(d, k, 1/\epsilon, \log(1/\delta))$ time and, with probability at least $1 - \delta$, outputs a Linear Sorting function $h : \mathbb{R}^d \to \mathbb{S}_k$ that is ϵ-close in KT distance to the target.*

We are now ready to provide the proof of our efficient proper learning algorithm for the class of Linear Sorting functions in the presence of bounded noise with respect to the standard Gaussian probability measure.

Proof. As a first step, the algorithm calls the improper learning algorithm `ImproperLSF` (Algorithm 9) with parameters ϵ, δ and $\eta < 1/2$ and obtains a list of linear classifiers with normal vectors $\boldsymbol{v}_{ij}$ for $i < j$. The utility of this step implies that, with probability at least $1 - \delta$, each one of the classifiers ϵ-learns the associated true halfspace, i.e., it holds

$$\mathop{\mathbf{Pr}}_{x \sim \mathcal{N}_d}[\mathrm{sign}(\boldsymbol{v}_{ij} \cdot \boldsymbol{x}) \neq \mathrm{sign}((\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \cdot \boldsymbol{x})] \leq \epsilon\,,$$

where $\boldsymbol{W}^\star$ is the matrix of the target Linear Sorting function. Without loss of generality, assume that $\|\boldsymbol{v}_{ij}\|_2 = 1$. In order to make the learner proper, it suffices to solve the following convex program on $\boldsymbol{W}$:

$$\begin{aligned} &\text{Find} && \boldsymbol{W} \in \mathbb{R}^{k \times d}, && \text{(D.1)} \\ &\text{such that} && (\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{v}_{ij} \geq (1 - \phi) \cdot \|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2 \quad \text{for any } 1 \leq i < j \leq k\,, && (\texttt{CP}) \\ &&&&& \text{(D.2)} \\ &&& \|\boldsymbol{W}\|_F \leq 1\,, && \text{(D.3)} \end{aligned}$$

for some $\phi \in (0, 1)$ to be decided. The main key ideas are summarized in the next claim.

Claim D.6. *The following properties hold true for $\phi = O(\epsilon^2)$ with probability at least $1 - \delta$.*

1. *The convex program D.1 is feasible.*
2. *Any solution of the convex program D.1 induces an LSF that is ϵ-close in KT distance to the true target $\sigma_{\boldsymbol{W}^\star}(\cdot)$.*
3. *The feasible set of the convex program D.1 contains a ball of radius $r = 2^{-\mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))}$ and is contained in a ball of radius 1. Both balls are with respect to the Frobenius norm.*
4. *The convex program D.1 can be solved in time $\mathrm{poly}(d, k, 1/\epsilon, \log(1/\delta))$ using the ellipsoid algorithm.*

Proof of Item 1. First, we can choose the error ϕ so that this convex program is feasible. Let us set $\boldsymbol{W} = \boldsymbol{W}^\star$, where $\boldsymbol{W}^\star$ is the underlying matrix of the target Linear

Sorting function $\sigma^\star$ with $\sigma^\star(\boldsymbol{x}) = \mathrm{argsort}(\boldsymbol{W}^\star \boldsymbol{x})$. Recall that, by the guarantees of the improper learning algorithm, for the pair $1 \leq i < j \leq k$, it holds

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\mathrm{sign}(\boldsymbol{v}_{ij} \cdot \boldsymbol{x}) \neq \mathrm{sign}((\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \cdot \boldsymbol{x})] \leq \epsilon \,. \tag{D.4}$$

Since the standard Gaussian is rotationally symmetric, the angle $\theta(\boldsymbol{u}, \boldsymbol{v})$ between two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ is equal to $\pi \cdot \mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{N}_d}[\mathrm{sign}(\boldsymbol{u} \cdot \boldsymbol{x}) \neq \mathrm{sign}(\boldsymbol{v} \cdot \boldsymbol{x})]$. Hence, using this observation and Equation (D.4), we get that the angle between the guess vector $\boldsymbol{v}_{ij}$ and the true normal vector $\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star$ is

$$\theta(\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star, \boldsymbol{v}_{ij}) \leq c \cdot \epsilon \,,$$

for some constant $c > 0$. For sufficiently small ϵ, this bound implies that the cosine of the above angle is of order $1 - (c\epsilon)^2$ and so the following inequality will hold

$$(\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \cdot \boldsymbol{v}_{ij} \geq (1 - 2(c\epsilon)^2) \cdot \|\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star\|_2 \,,$$

since $\boldsymbol{v}_{ij}$ is unit. Hence, by setting $\phi = 2(c\epsilon)^2$, the convex program with variables $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ will be feasible; $\boldsymbol{W}^\star$ will be a solution with probability $1 - \delta$, where the randomness is over the output of the algorithm dealing with the Massart linear classifiers. Note that we can assume that $\|\boldsymbol{W}^\star\|_F \leq 1$ without loss of generality, since we can divide each row with the Frobenius norm.

Proof of Item 2. Let $\widetilde{\boldsymbol{W}}$ be a solution of the convex program.We will make use of the observation that the angle between two vectors is equal to the disagreement of the associated linear threshold functions with respect to the standard normal times π. Observe that any solution $\widetilde{\boldsymbol{W}}$ to the convex program will satisfy that

$$(\forall i, j) \quad \theta(\boldsymbol{v}_{ij}, \widetilde{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_j) \leq O(\sqrt{\phi}) = c\epsilon \,.$$

and

$$(\forall i, j) \quad \theta(\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star, \boldsymbol{v}_{ij}) \leq \epsilon \,.$$

This implies that

$$d_{\text{angle}}(\boldsymbol{W}^\star, \widetilde{\boldsymbol{W}}) \leq c' \, \epsilon$$

Claim D.7. *For the matrices $\boldsymbol{W}, \boldsymbol{W}^\star \in \mathbb{R}^{k \times d}$, it holds that*

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\Delta_{\text{KT}}(\sigma_{\boldsymbol{W}}(\boldsymbol{x}), \sigma_{\boldsymbol{W}^\star}(\boldsymbol{x}))] \leq d_{\text{angle}}(\boldsymbol{W}, \boldsymbol{W}^\star)\,.$$

Proof. We have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\Delta_{\text{KT}}(\sigma_{\boldsymbol{W}}(\boldsymbol{x}), \sigma_{\boldsymbol{W}^\star}(\boldsymbol{x}))] \\
&= \frac{1}{\binom{k}{2}} \cdot \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\sum_{1 \leq i < j \leq k} \mathbb{1}\{((\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{x})\,((\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \cdot \boldsymbol{x}) < 0\} \\
&= \frac{1}{\binom{k}{2}} \cdot \sum_{1 \leq i < j \leq k} \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\text{sign}(\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{x}) \neq \text{sign}((\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \cdot \boldsymbol{x})] \\
&= \frac{1}{\pi} \max_{i,j} \theta(\boldsymbol{W}_i - \boldsymbol{W}_j, \boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \\
&\leq d_{\text{angle}}(\boldsymbol{W}, \boldsymbol{W}^\star)\,.
\end{aligned}$$

□

Using the above claim, we get an expected KT distance bound of order $O(\epsilon)$. This gives the desired result.

Proof of Item 3. We will make use of the next lemma.

Lemma D.8. *Fix $\epsilon, \delta \in (0,1)$. Let $\boldsymbol{W}^\star \in \mathbb{R}^{k \times d}$ be the true parameter matrix. There exists a matrix $\widetilde{\boldsymbol{W}}^\star \in \mathbb{R}^{k \times d}$ such that, with probability at least $1 - \delta$:*

- $\mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{N}_d}[\text{sign}((\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star) \cdot \boldsymbol{x}) \neq \text{sign}((\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star) \cdot \boldsymbol{x})] \leq \epsilon$ *for all $i \neq j$, and,*
- $\|\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star\|_2 \geq 2^{-\text{poly}(d,k,1/\epsilon,\log(1/\delta))}$ *for any $i \neq j$.*

Proof of Lemma D.8. The above lemma is a result of the next Appendix D.1. In particular, it is a direct implication of Lemma D.11 and Corollary D.17. □

Note that the above lemma implies that

$$(\forall i, j) \quad \mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{N}_d}[\mathrm{sign}(\boldsymbol{v}_{ij} \cdot \boldsymbol{x}) \neq \mathrm{sign}((\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star) \cdot \boldsymbol{x})] \leq 2\epsilon\,,$$

with probability at least $1 - 2\delta$. Hence, up to constants, the analysis concerning the feasibility of the true matrix $\boldsymbol{W}^\star$ (see Item 1) will still hold for $\widetilde{\boldsymbol{W}}^\star$. From now on we can work with this matrix $\widetilde{\boldsymbol{W}}^\star$ which enjoys the "well-conditionedness" property of the second item of the lemma.

We will use the above lemma in order to prove Item 3 which controls the volume of the feasible region: it states that there exist $0 < r < R$ so that the feasible region of the convex program contains a ball of radius r and is contained in a ball of radius R (where the balls are with respect to the Frobenius norm). Moreover, $r = 2^{-\mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))}$ and $R = 1$.

For the chosen $\phi \in (0, 1)$, the feasible set contains matrices $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ that satisfy $\|\boldsymbol{W} - \widetilde{\boldsymbol{W}}^\star\|_F \leq 2r$, r to be decided. For any $i \neq j$, we have that the following properties hold:

1. $\|\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star\|_2 \geq 2^{-\mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))}$ (well-conditionedness).

2. $(\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star) \cdot \boldsymbol{v}_{ij} \geq (1 - \phi)\ \|\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star\|_2$ (feasibility).

3. $\|\boldsymbol{W} - \widetilde{\boldsymbol{W}}^\star\|_F \leq 2r$ which implies that $\|\boldsymbol{W}_i - \widetilde{\boldsymbol{W}}_i^\star\|_2 \leq 2r$ for any $i \in [k]$ (ball around feasible point).

4. $\|\boldsymbol{v}_{ij}\|_2 = 1$.

Our goal is to prove that for a matrix in the above ball it holds $(\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{v}_{ij} \geq (1 - \phi)\ \|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2$.

We have that

$$\begin{aligned}(\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star) \cdot \boldsymbol{v}_{ij} &= (\widetilde{\boldsymbol{W}}_i^\star - \boldsymbol{W}_i) \cdot \boldsymbol{v}_{ij} + (\boldsymbol{W}_j - \widetilde{\boldsymbol{W}}_j^\star) \cdot \boldsymbol{v}_{ij} + (\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{v}_{ij} \\ &\leq \|\widetilde{\boldsymbol{W}}_i^\star - \boldsymbol{W}_i\|_2 + \|\boldsymbol{W}_j - \widetilde{\boldsymbol{W}}_j^\star\|_2 + (\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{v}_{ij} \\ &\leq 4r + (\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{v}_{ij}\,.\end{aligned}$$

More to that

$$\begin{aligned} \|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2 &= \|\boldsymbol{W}_i - \widetilde{\boldsymbol{W}}_i^\star + \widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star + \widetilde{\boldsymbol{W}}_j^\star - \boldsymbol{W}_j\|_2 \\ &\leq \|\boldsymbol{W}_i - \widetilde{\boldsymbol{W}}_i^\star\|_2 + \|\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star\|_2 + \|\widetilde{\boldsymbol{W}}_j^\star - \boldsymbol{W}_j\|_2 \\ &\leq 4r + \|\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star\|_2 , \end{aligned}$$

and similarly: $\|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2 \geq \|\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star\|_2 - 4r$.

Combining the above inequalities, we get that

$$\begin{aligned} (\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \boldsymbol{v}_{ij} &\geq (\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star) \cdot \boldsymbol{v}_{ij} - 4r \\ &\geq (1-\phi)\ \|\widetilde{\boldsymbol{W}}_i^\star - \widetilde{\boldsymbol{W}}_j^\star\|_2 - 4r \\ &\geq (1-\phi)\ (\|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2 - 4r) - 4r \\ &= (1-\phi)\ \|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2 - 8r\,. \end{aligned}$$

We pick r sufficiently small and of order $2^{-\mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))}$ and get that $\boldsymbol{W}$ is a feasible solution of the convex program. Moreover, we can select $R = 1$ since $\|\widetilde{\boldsymbol{W}}^\star\|_F = 1$ without loss of generality, since we can normalize the row differences of $\widetilde{\boldsymbol{W}}^\star$ with the norm $\|\widetilde{\boldsymbol{W}}^\star\|_F$.

Proof of Item 4. We apply the ellipsoid algorithm in order to solve the convex program D.1 and compute a matrix $\widetilde{\boldsymbol{W}} \in \mathbb{R}^{k\times d}$. The algorithm `ProperLSF` outputs the linear sorting function $h(\cdot) = \sigma_{\widetilde{\boldsymbol{W}}}(\cdot)$.

Lemma D.9 (Efficiency of the Ellipsoid Algorithm Vishnoi (2021)). *Suppose that $P \subseteq \mathbb{R}^d$ is a full-dimensional polytope that is contained in a d-dimensional Euclidean ball of radius $R > 0$ and contains a d-dimensional Euclidean ball of radius $r > 0$. Then, the ellipsoid method outputs a point $\widetilde{\boldsymbol{x}} \in P$ after $O(d^2 \log(R/r))$ iterations. Moreover, every iteration can be implemented in $O(d^2 + T_{\mathrm{sep}})$ time, where T_{sep} is the time required to answer a single query by the separation oracle.*

Assume that Item 3 holds true. Then the algorithm can be used with $r = 2^{-\mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))}$ and $R = 1$. Hence, the ellipsoid algorithm will provide in time

$\text{poly}(d, k, 1/\epsilon, \log(1/\delta))$ a point $\widetilde{\boldsymbol{W}}$ that lies in the feasible region of the convex program D.1[2].

□

Remark D.10. *We remark that both the improper (Algorithm 9) and the proper (Algorithm 10) learning algorithms hold for the more general case where the* $\boldsymbol{x}$*-marginal lies in the class of isotropic log-concave distributions Lovász and Vempala (2007): A distribution* $\mathcal{D}_x$ *lies inside the class of isotropic log-concave distributions* $\mathcal{F}_{\mathrm{LC}}$ *over* $\mathbb{R}^d$ *if* $\mathcal{D}_x$ *has a probability density function* f *over* $\mathbb{R}^d$ *such that* $\log f$ *is concave, its mean is zero, and its covariance is identity, i.e.,* $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{D}_x}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}$.

The proof of Lemma D.8

We provide the following result.

Lemma D.11. *Fix* $\epsilon, \delta \in (0, 1)$*. Let* $\boldsymbol{W}^\star \in \mathbb{R}^{k\times d}$ *be the true parameter matrix. There exists a matrix* $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ *such that, with probability at least* $1 - \delta$*:*

- $\mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}_d}[\text{sign}((\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star)\cdot\boldsymbol{x}) \neq \text{sign}((\boldsymbol{W}_i - \boldsymbol{W}_j)\cdot\boldsymbol{x})] \leq \epsilon$ *for all* $i \neq j$*, and,*
- *The bit complexity of* $\boldsymbol{W}$ *is* $\text{poly}(k, d, 1/\epsilon, \log(1/\delta))$

Proof. The matrix $\boldsymbol{W}$ will be the output of a linear program that can be used to learn the LSF $\sigma_{\boldsymbol{W}^\star}(\cdot)$ in the noiseless setting.

Consider the unit sphere $\mathcal{S}^{d-1}$ and a δ_0-cover of the unit sphere with parameter $\delta_0 > 0$ to be decided. For any sample $(\boldsymbol{x}, \pi) \sim \mathcal{D}$ of the 0-noisy linear label ranking distribution, i.e., $\boldsymbol{x} \sim \mathcal{N}_d$ and $\pi = \sigma_{\boldsymbol{W}^\star}(\boldsymbol{x})$, we consider the rounded sample $(\widetilde{\boldsymbol{x}}, \pi)$ where $\widetilde{\boldsymbol{x}}$ is obtained by first projecting $\boldsymbol{x} \in \mathbb{R}^d$ to $\mathcal{S}^{d-1}$ and then by obtaining the closest point of $\widehat{x}$ in the cover. The cover's size is $O(1/\delta_0)^d$.

Let us fix $1 \leq i < j \leq k$ and set $y_{ij} = \text{sign}(\pi(i) - \pi(j))$. For a training set $\{(\boldsymbol{x}^{(t)}, \pi^{(t)})\}_{t\in[N]}$ of size N, we create the following linear system L_{ij} with variables

[2] We remark that the runtime will also depend on the time required to answer a single query by the separation oracle. We assume that this time is polynomial in the parameters of our problem and we opt not to track these details in this work.

$\boldsymbol{W} \in \mathbb{R}^{k \times d}$:

$$y_{ij}^{(t)} \, (\boldsymbol{W}_i - \boldsymbol{W}_j) \cdot \widetilde{\boldsymbol{x}}^{(t)} \geq 0 \, , \; t \in [N] \qquad (\mathrm{L}_{ij}) \, .$$

Consider the concatenation of the linear systems $\mathrm{L} = \cup_{i<j} \mathrm{L}_{ij}$. The number of equations in the linear system of equations L is $N \cdot \binom{k}{2}$.

We first have to show that, with high probability, the system L is feasible, i.e., there exists $\boldsymbol{W}$ that satisfies the system's equations. Note that if we replace $\widetilde{\boldsymbol{x}}^{(t)}$ with the original points $\boldsymbol{x}^{(t)}$, the true matrix $\boldsymbol{W}^\star$ is a solution to the system. We now have to study the rounded linear system.

Claim D.12. *The (rounded) linear system* L *is feasible with high probability.*

Proof. In order to show the feasibility of L, we will use the anti-concentration properties of the Gaussian.

Fact D.13 (Dasgupta et al. (2005))**.** *Let $\mathcal{P}$ be the standard normal distribution over $\mathbb{R}^d$. For any fixed unit vector $\boldsymbol{a} \in \mathbb{R}^d$ and any $\gamma \leq 1$,*

$$\gamma/4 \leq \mathop{\mathbf{Pr}}_{x \sim \mathcal{P}} \left[|\boldsymbol{a} \cdot \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}| \leq \frac{\gamma}{\sqrt{d}} \right] \leq \gamma \, .$$

Let us focus on the pair $1 \leq i < j \leq k$. We first observe that scaling all samples to lie on the unit sphere does not affect the feasibility of the system. It suffices to focus on that single halfspace with normal vector $\boldsymbol{v}_{ij} = \boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star \in \mathbb{R}^d$ and consider the probability of the event that the collection of the N rounded points $\{\widetilde{\boldsymbol{x}}^{(t)}\}_t$ with labels $\{y_{ij}^{(t)}\}_t$, that come from N Gaussian vectors $\{\boldsymbol{x}^{(t)}\}_t$ which are linearly separable (with labels $\{y_{ij}^{(t)}\}_t$), becomes non-linearly separable. For this it suffices to control the probability that the rounding procedure flips the label of the data point. Using the union bound, we have that, if the rounding has accuracy δ_0, the described bad event has probability

$$\begin{aligned} &\mathop{\mathbf{Pr}}_{\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(N)} \sim \mathcal{N}_d} [\exists t \in [N] : \mathrm{sign}(\boldsymbol{v}_{ij} \cdot \widetilde{\boldsymbol{x}}^{(t)}) \neq \mathrm{sign}(\boldsymbol{v}_{ij} \cdot \boldsymbol{x}^{(t)})] \\ &\leq N \cdot \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d} \left[|\boldsymbol{v}_{ij} \cdot \boldsymbol{x} / \|\boldsymbol{x}\|_2| \leq 2\delta_0 \right] \leq N \cdot O(\delta_0 \sqrt{d}) \, , \end{aligned}$$

where we remark that the first event is scale invariant and so we can assume that the normal vector is unit, the first inequality follows from the fact that it suffices to control the mass assigned to a strip of width $2\delta_0$ (due to the discretization) and the second inequality follows from Fact D.13. We now have to select the discretization. Let $\delta \in (0,1)$. By choosing $\delta_0 = O(\frac{\delta}{N\sqrt{d}k^2})$, the bad event for all the pairs $i < j$ occurs with probability at most δ, i.e., with probability at least $1-\delta$, each one of the N drawn i.i.d. samples does not fall in any one of the $\binom{k}{2}$ "bad" strips. □

We can now consider the case that the system L is feasible (with the target matrix $\boldsymbol{W}^\star$ being a feasible point) that occurs with probability $1-\delta$. The class of homogenous halfspaces in d dimensions has VC dimension d; therefore, the sample complexity of learning halfspaces using ERM is $O((d+\log(1/\delta))/\epsilon)$. Moreover, in the realizable case, we can implement the ERM using e.g., linear programming and find a solution in $\mathrm{poly}(d, 1/\epsilon, \log(1/\delta))$ time. We next focus on the quality of the solution which will give the desired sample complexity.

Claim D.14. *Assume that the algorithm draws $N = \widetilde{O}(\frac{d+\log(k/\delta)}{\epsilon})$ i.i.d. samples of the form $(\boldsymbol{x}, \pi)$ with $\boldsymbol{x} \sim \mathcal{N}_d$ and $\pi = \sigma_{\boldsymbol{W}^\star}(\boldsymbol{x})$. For any $i \neq j$ and with probability at least $1-2\delta$, the solution $\boldsymbol{W}$ of the linear system* L *satisfies*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[\mathrm{sign}((\boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star)\cdot\boldsymbol{x}) \neq \mathrm{sign}((\boldsymbol{W}_i - \boldsymbol{W}_j)\cdot\boldsymbol{x})] \leq \epsilon\,.$$

Proof. Since the matrix $\boldsymbol{W}$ satisfies the sub-system L_{ij}, the result follows using a union bound on the events that (i) the linear system is feasible and (ii) the ERM is a successful PAC learner. □

Claim D.15. *Consider the solution $\boldsymbol{W}$ of the linear system. Then, $\boldsymbol{W}$ has bounded bit complexity of order* $\mathrm{poly}(d, k, 1/\epsilon, \log(1/\delta))$.

Proof. We will make use of the following result that relates the size of the input and the output of a linear program using Cramer's rule.

Lemma D.16 (Schrijver (1998); Papadimitriou (1981))**.** *Let $\boldsymbol{A} \in \mathbb{Z}^{m\times n}, \boldsymbol{b} \in \mathbb{Z}^m, \boldsymbol{c} \in \mathbb{Z}^n$. Consider a linear program* $\min \boldsymbol{c}\cdot\boldsymbol{x}$ *subject to $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ and $\boldsymbol{x} \geq \boldsymbol{0}$. Let U be* *the maximum size of A_{ij}, b_i, c_j. The output of the linear program has size $O(m(nU + n\log(n)))$ bits.*

We will apply the above lemma (which holds even by dropping the constraint $\boldsymbol{x} \geq \mathbf{0}$) to our setting where $\boldsymbol{A}\boldsymbol{w} \geq 0$ where $\boldsymbol{w} = (\boldsymbol{W}_i)_{i\in[k]} \in \mathbb{Q}^{kd}$, i.e., $\boldsymbol{w}$ is the vectorization of the matrix $\boldsymbol{W}$. Moreover, $\boldsymbol{A}$ is the matrix containing the N (rounded) Gaussian samples $\widetilde{\boldsymbol{x}}^{(t)}$. We have that the matrix $\boldsymbol{A}$ has dimension $N\binom{k}{2} \times kd$ and each entry A_{ij} is an integer and has size at most $U = \mathrm{poly}(d,k)$ (since the samples are rounded on the δ_0-cover of the sphere. Recall that the labels $y_{ij}^{(t)} \in \{-1,+1\}$ and $\widetilde{\boldsymbol{x}}^{(t)}$ lie in the unit sphere. In particular, each row of the matrix $\boldsymbol{A}$ has $2d$ non-zero entries and is associated with a tuple (i,j,t) for $1 \leq i < j \leq k$ and $t \in [N]$. Then, it holds that the output has size at most $O(Nk^2(dU + dk\log(dk)))$ bits. So, we get that the output $\boldsymbol{W}$ can be described using at most $\mathrm{poly}(d,k,1/\epsilon,U,\log(1/\delta)) = \mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))$ bits (due to the size of the entries of the matrix $\boldsymbol{A}$). □

Combining the above claims, we conclude the proof. □

As a corollary of the bounded bit complexity, we obtain the following key result.

Corollary D.17. *Let $\epsilon > 0$. Assume that $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ has bit complexity at most $\mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))$. Then, for any $i,j \in [k]$ with $i \neq j$, it holds that $\|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2 > 2^{-\mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))}$.*

Proof. First, we can assume that $\boldsymbol{W}_i \neq \boldsymbol{W}_j$ for any $i \neq j$; in case of equal rows, we obtain a low-dimensional instance. Then, since any vector $\boldsymbol{W}_i$ has bounded bit complexity, we have that the difference of any two such vectors, provided that it is non-zero, has a lower bound in its norm, i.e., $\|\boldsymbol{W}_i - \boldsymbol{W}_j\|_2 > 2^{-\mathrm{poly}(d,k,1/\epsilon,\log(1/\delta))}$ for any $i,j \in [k]$. □

D.2 Learning in Top-1 Disagreement from Label Rankings

Let us set $\sigma_1(\boldsymbol{W}\boldsymbol{x}) = \text{argmax}_{i \in [k]} \boldsymbol{W}_i \cdot \boldsymbol{x}$ for $\boldsymbol{x} \in \mathbb{R}^d$. The main result of this section follows.

Theorem D.18 (Proper Top-1 Learning Algorithm)**.** *Fix $\eta \in [0, 1/2)$ and $\epsilon, \delta \in (0, 1)$. Let $\mathcal{D}$ be an η-noisy linear label ranking distribution satisfying the assumptions of Definition 1.10. There exists an algorithm that draws $N = O\left(\frac{dk\sqrt{\log k}}{\epsilon(1-2\eta)^6} \log(k/\delta)\right)$ samples from $\mathcal{D}$, runs in* $\text{poly}(N)$ *time and, with probability at least $1 - \delta$, outputs a Linear Sorting function $h : \mathbb{R}^d \to \mathbb{S}_k$ that is ϵ-close in top-1 disagreement to the target.*

Proof. Note that the `MassartLTF` algorithm (see Lemma D.2) has the guarantee that it returns a vector $\boldsymbol{w}$ so that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\text{sign}(\boldsymbol{w} \cdot \boldsymbol{x}) \neq \text{sign}(\boldsymbol{w}^\star \cdot \boldsymbol{x})] \leq \epsilon\,,$$

with probability $1 - \delta$, where $\boldsymbol{w}^\star$ is the target normal vector. Since the above misclassification probability with respect to $\mathcal{N}_d$ is directly connected with the angle $\theta(\boldsymbol{w}, \boldsymbol{w}^\star)$, we get that we can control the angle between $\boldsymbol{w}$ and $\boldsymbol{w}^\star$ efficiently. Moreover, in our setting, for a matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$, there exist $\binom{k}{2}$ homogeneous halfspaces with normal vectors $\boldsymbol{W}_i - \boldsymbol{W}_j$ and so we can control the angles $\theta(\boldsymbol{W}_i - \boldsymbol{W}_j, \boldsymbol{W}_i^\star - \boldsymbol{W}_j^\star)$. In order to deduce the sample complexity bound of Theorem D.18, we show the next lemma which essentially bounds the top-1 misclassification error using the angles of these $O(k^2)$ halfspaces. We apply Lemma D.19 with $\boldsymbol{U} = \boldsymbol{W}$ and $\boldsymbol{V} = \boldsymbol{W}^\star$ and so we can take $\epsilon' = \epsilon/(k\sqrt{\log k})$ and invoke the proper learning algorithm of Algorithm 10. This completes the proof. □

We continue with the proof of our key lemma.

Lemma D.19 (Misclassification Error)**.** *Consider two matrices $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{k \times d}$ and let $\mathcal{N}_d$ be the standard Gaussian in d dimensions. We have that*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\sigma_1(\boldsymbol{U}\boldsymbol{x}) \neq \sigma_1(\boldsymbol{V}\boldsymbol{x})] \leq c \cdot k \cdot \sqrt{\log k} \cdot \max_{i \neq j} \theta(\boldsymbol{U}_i - \boldsymbol{U}_j, \boldsymbol{V}_i - \boldsymbol{V}_j)\,,$$

where $c > 0$ is some universal constant.

Proof. We have that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\sigma_1(\boldsymbol{U}\boldsymbol{x}) \neq \sigma_1(\boldsymbol{V}\boldsymbol{x})] = \sum_{i \in [k]} \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\sigma_1(\boldsymbol{U}\boldsymbol{x}) = i, \sigma_1(\boldsymbol{V}\boldsymbol{x}) \neq i]\,.$$

We have that $\mathcal{C}_{\boldsymbol{U}}^{(i)} = \mathbb{1}\{\boldsymbol{x} : \sigma_1(\boldsymbol{U}\boldsymbol{x}) = i\} = \prod_{j \neq i} \mathbb{1}\{(\boldsymbol{U}_i - \boldsymbol{U}_j) \cdot \boldsymbol{x} \geq 0\}$ is the set indicator of a homogeneous polyhedral cone as the intersection of $k - 1$ homogeneous halfspaces. Similarly, we consider the cone $\mathcal{C}_{\boldsymbol{V}}^{(i)} = \{\boldsymbol{x} : \sigma_1(\boldsymbol{V}\boldsymbol{x}) = i\}$. Hence, we have that $\{\boldsymbol{x} : \sigma_1(\boldsymbol{V}\boldsymbol{x}) \neq i\}$ is the complement of a homogeneous polyhedral cone. Let us define $C_{\boldsymbol{U}}^{(i)} : \mathbb{R}^d \mapsto \{0, 1\}$ and $C_{\boldsymbol{V}}^{(i)} : \mathbb{R}^d \mapsto \{0, 1\}$ be the associated indicator functions of the two cones. We have that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\sigma_1(\boldsymbol{U}\boldsymbol{x}) = i, \sigma_1(\boldsymbol{V}\boldsymbol{x}) \neq i] = \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[C_{\boldsymbol{U}}^{(i)}(\boldsymbol{x}) = 1, C_{\boldsymbol{V}}^{(i)}(\boldsymbol{x}) = 0]\,.$$

Finally, we have that

$$\mathcal{C}_{\boldsymbol{U}}^{(i)} \cap \left(\mathcal{C}_{\boldsymbol{V}}^{(i)}\right)^c = \mathcal{C}_{\boldsymbol{U}}^{(i)} \setminus \mathcal{C}_{\boldsymbol{V}}^{(i)} \subseteq \mathcal{C}_{\boldsymbol{U}}^{(i)} \setminus \mathcal{C}_{\boldsymbol{V}}^{(i)} \cup \mathcal{C}_{\boldsymbol{V}}^{(i)} \setminus \mathcal{C}_{\boldsymbol{U}}^{(i)}\,.$$

We can hence apply Lemma D.20 for the cones $\mathcal{C}_{\boldsymbol{U}}^{(i)}, \mathcal{C}_{\boldsymbol{V}}^{(i)}$ for each $i \in [k]$. □

Lemma D.20 (Cone Disagreement)**.** *Let $C_1 : \mathbb{R}^d \mapsto \{0, 1\}$ be the indicator function of the homogeneous polyhedral cone defined by the k unit vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \mathbb{R}^d$, i.e., $C_1(\boldsymbol{x}) = \prod_{i=1}^k \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \geq 0\}$. Similarly, define $C_2 : \mathbb{R}^d \mapsto \{0, 1\}$ to be the homogeneous polyhedral cone with normal vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$. It holds that*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathbb{N}_d}[C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})] \leq c\sqrt{\log(k)}\, \max_{i \in [k]} \theta(\boldsymbol{v}_i, \boldsymbol{u}_i)\,,$$

where $c > 0$ is some universal constant.

Proof. To simplify notation, denote $\theta = \max_{i \in [k]} \theta(\boldsymbol{v}_i, \boldsymbol{u}_i)$. We first observe that it suffices to prove the upper bound on the probability of $C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})$ for sufficiently small values of θ. Indeed, if we have that the bound is true for θ smaller than some θ_0 we can then form a path of sufficiently large length N (in particular we need $\theta/N \leq \theta_0$) starting from the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ to the final vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$, where at each step we only rotate the vectors by at most $\theta/N \leq \theta_0$. By the triangle inequality, we immediately obtain that the probability that $C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})$ is at most equal to the sum of the probabilities of the intermediate steps which is at most $\sum_{i=1}^{N} c\sqrt{\log(k)} \frac{\theta}{N} = c\sqrt{\log(k)}\theta$. Notice in the above argument the constant θ_0 can be arbitrarily small and may also depend on k and d.

We define the indicator of the positive orthant in k dimensions to be $R(\boldsymbol{t}) = \prod_{i=1}^{k} \mathbb{1}\{\boldsymbol{t}_i \geq 0\}$. Using this notation, we have that the cone indicator can be written as $C_1(\boldsymbol{x}) = R(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}) = R(\boldsymbol{V}\boldsymbol{x})$, where $\boldsymbol{V}$ is the $k \times d$ matrix whose i-th row is the vector $\boldsymbol{v}_i$. Moreover, we define the i-th face of the cone $R(\boldsymbol{V}\boldsymbol{x})$ to be

$$F_i(\boldsymbol{V}\boldsymbol{x}) = R(\boldsymbol{V}\boldsymbol{x})\, \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} = 0\}\,.$$

We will first handle the case where only one of the normal vectors $\boldsymbol{v}_i$ changes. We show the following claim.

Claim D.21. *Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \mathbb{R}^d$ and $\boldsymbol{r} \in \mathbb{R}^d$ with $\theta(\boldsymbol{v}_1, \boldsymbol{r}) \leq \theta$ for some sufficiently small $\theta \in (0, \pi/2)$. It holds that*

$$\mathop{\mathbf{Pr}}_{x \sim \mathbb{N}_d}\left[R(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \neq R(\boldsymbol{r} \cdot \boldsymbol{x}, \boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x})\right] \leq c \cdot \theta \cdot \Gamma(F_1)\sqrt{\log\left(\frac{1}{\Gamma(F_1)} + 1\right)},$$

where F_1 is the face with $\boldsymbol{v}_1 \cdot \boldsymbol{x} = 0$ of the cone $R(\boldsymbol{V}\boldsymbol{x})$ and c is some universal constant.

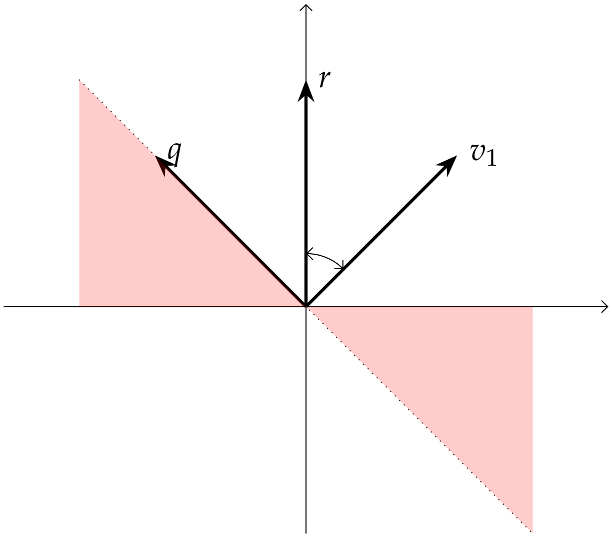


Figure D.1: The vectors $\boldsymbol{r}, \boldsymbol{v}_1$ and $\boldsymbol{q}$ and the disagreement region of the halfspaces with normal vectors $\boldsymbol{r}$ and $\boldsymbol{v}_1$.

Proof. We have

$$\begin{aligned}
&\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathbb{N}_d}[R(\boldsymbol{v}_1\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x}) \neq R(\boldsymbol{r}\cdot\boldsymbol{x},\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x})] \\
&\qquad = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_d}[|R(\boldsymbol{v}_1\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x}) - R(\boldsymbol{r}\cdot\boldsymbol{x},\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x})|] \\
&\qquad = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_d}[R(\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x})\,|\mathbb{1}\{\boldsymbol{v}_1\cdot\boldsymbol{x}\geq 0\} - \mathbb{1}\{\boldsymbol{r}\cdot\boldsymbol{x}\geq 0\}|]\,.
\end{aligned}$$

We have that $|\mathbb{1}\{\boldsymbol{v}_1\cdot\boldsymbol{x}\geq 0\} - \mathbb{1}\{\boldsymbol{r}\cdot\boldsymbol{x}\geq 0\}| = \mathbb{1}\{(\boldsymbol{v}_1\cdot\boldsymbol{x})(\boldsymbol{r}\cdot\boldsymbol{x}) < 0\}$, i.e., this is the event that the halfspaces $\mathbb{1}\{\boldsymbol{v}_1\cdot\boldsymbol{x}\geq 0\}$ and $\mathbb{1}\{\boldsymbol{r}\cdot\boldsymbol{x}\geq 0\}$ disagree. Let $\boldsymbol{q}$ be the normalized projection of $\boldsymbol{r}$ onto the orthogonal complement of $\boldsymbol{v}_1$, i.e., $\boldsymbol{q} = \mathrm{proj}_{\boldsymbol{v}_1^\perp}\boldsymbol{r}/\|\mathrm{proj}_{\boldsymbol{v}_1^\perp}\boldsymbol{r}\|_2$. We have that $\boldsymbol{v}_1$ and $\boldsymbol{q}$ is an orthonormal basis of the subspace spanned by the vectors $\boldsymbol{v}_1$ and $\boldsymbol{r}$. We have that $\boldsymbol{r} = \cos\theta(\boldsymbol{v}_1,\boldsymbol{r})\boldsymbol{v}_1 + \sin\theta(\boldsymbol{v}_1,\boldsymbol{r})\boldsymbol{q}$. Moreover, we have that the region $(\boldsymbol{v}_1\cdot\boldsymbol{x})(\boldsymbol{r}\cdot\boldsymbol{x}) < 0$ is equal to

$$\{0 < \boldsymbol{v}_1\cdot\boldsymbol{x} < -(\boldsymbol{q}\cdot\boldsymbol{x})\tan\theta(\boldsymbol{v}_1,\boldsymbol{r})\} \cup \{-(\boldsymbol{q}\cdot\boldsymbol{x})\tan\theta(\boldsymbol{v}_1,\boldsymbol{r}) < \boldsymbol{v}_1\cdot\boldsymbol{x} < 0\}\,.$$

Thus, we have that the disagreement region $(\boldsymbol{v}_1\cdot\boldsymbol{x})(\boldsymbol{r}\cdot\boldsymbol{x}) < 0$ is a subset of the region $\{|\boldsymbol{v}_1 \cdot \boldsymbol{x}| \leq |\boldsymbol{q} \cdot \boldsymbol{x}| \tan \theta(\boldsymbol{v}_1, \boldsymbol{r})\}$. Since $\tan \theta(\boldsymbol{v}_1, \boldsymbol{r}) \leq \theta$ and we have that θ is sufficiently small we can also replace the above region by the larger region: $\{|\boldsymbol{v}_1 \cdot \boldsymbol{x}| \leq 2\theta |\boldsymbol{q} \cdot \boldsymbol{x}|\}$. Therefore, we have

$$\begin{aligned}&\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{N}_d} \left[R(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \, \mathbb{1}\{(\boldsymbol{v}_1 \cdot \boldsymbol{x})(\boldsymbol{r} \cdot \boldsymbol{x}) < 0\}\right] \\ &\qquad \leq \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{N}_d} \left[R(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \, \mathbb{1}\{|\boldsymbol{v}_1 \cdot \boldsymbol{x}| \leq 2\theta |\boldsymbol{q} \cdot \boldsymbol{x}|\}\right] .\end{aligned}$$

The derivative of the above expression with respect to θ is equal to

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{N}_d} \left[R(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \, \delta\left(\frac{|\boldsymbol{v}_1 \cdot \boldsymbol{x}|}{2|\boldsymbol{q} \cdot \boldsymbol{x}|} - \theta\right)\right] ,$$

where $\delta(t)$ is the Dirac delta function. At $\theta = 0$ and using the property that $\delta(t/a) = a\delta(t)$, we have that the above derivative is equal to

$$2 \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{N}_d} \left[R(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \, |\boldsymbol{q} \cdot \boldsymbol{x}| \, \delta(|\boldsymbol{v}_1 \cdot \boldsymbol{x}|)\right] .$$

Notice that, if we did not have the term $|\boldsymbol{q} \cdot \boldsymbol{x}|$, the above expression would be exactly equal to two times the Gaussian surface area of the face with $\boldsymbol{v}_1 \cdot \boldsymbol{x} = 0$, i.e., it would be equal to $2\Gamma(F_1)$. We now show that this extra term of $|\boldsymbol{q} \cdot \boldsymbol{x}|$ can only increase the above surface integral by at most a logarithmic factor. We have that

$$\begin{aligned}&\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{N}_d} \left[R(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \, |\boldsymbol{q} \cdot \boldsymbol{x}| \, \delta(|\boldsymbol{v}_1 \cdot \boldsymbol{x}|)\right] = \int_{\boldsymbol{x} \in F_1} \phi_d(\boldsymbol{x}) |\boldsymbol{q} \cdot \boldsymbol{x}| d\mu(\boldsymbol{x}) \\ &\leq \int_{\boldsymbol{x} \in F_1} \phi_d(\boldsymbol{x}) |\boldsymbol{q} \cdot \boldsymbol{x}| \mathbb{1}\{|\boldsymbol{q} \cdot \boldsymbol{x}| \leq \xi\} d\mu(\boldsymbol{x}) + \int_{\boldsymbol{x} \in F_1} \phi_d(\boldsymbol{x}) |\boldsymbol{q} \cdot \boldsymbol{x}| \mathbb{1}\{|\boldsymbol{q} \cdot \boldsymbol{x}| \geq \xi\} d\mu(\boldsymbol{x}) \\ &\leq \xi \int_{\boldsymbol{x} \in F_1} \phi_d(\boldsymbol{x}) d\mu(\boldsymbol{x}) + \int_{\boldsymbol{x} \in F_1} \phi_d(\boldsymbol{x}) |\boldsymbol{q} \cdot \boldsymbol{x}| \mathbb{1}\{|\boldsymbol{q} \cdot \boldsymbol{x}| \geq \xi\} d\mu(\boldsymbol{x}) ,\end{aligned}$$

where $d\mu(\boldsymbol{x})$ is the standard surface measure in $\mathbb{R}^d$. The first term above is exactly equal to the Gaussian surface area of the face F_1. To bound from above the second term we can use the fact that the face F_1 is a subset of the hyperplane $\boldsymbol{v}_1 \cdot \boldsymbol{x} = 0$, i.e., it holds that $F_1 \subseteq \{\boldsymbol{x} : |\boldsymbol{v}_1 \cdot \boldsymbol{x}| = 0\}$. To simplify notation we may assume that

$\boldsymbol{v}_1 = \boldsymbol{e}_1$ and $\boldsymbol{q} = \boldsymbol{e}_2$ (recall that $\boldsymbol{v}_1$ and $\boldsymbol{q}$ are orthogonal unit vectors), and in this case we obtain

$$\begin{aligned}\int_{\boldsymbol{x}\in F_1} \phi_d(\boldsymbol{x})|\boldsymbol{q}\cdot\boldsymbol{x}|\mathbb{1}\{|\boldsymbol{q}\cdot\boldsymbol{x}| \geq \xi\}d\mu(\boldsymbol{x}) &\leq \int_{x_1=0} \phi_d(\boldsymbol{x})|x_2|\mathbb{1}\{|x_2| \geq \xi\}d\mu(\boldsymbol{x}) \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} |x_2|\mathbb{1}\{|x_2| \geq \xi\}\frac{e^{-x_2^2/2}}{\sqrt{2\pi}}dx_2 \\ &= \frac{1}{\pi}e^{-\xi^2/2}\,.\end{aligned}$$

Combining the above bounds we obtain that the derivative with respect to θ of the expression $\mathbf{E}_{\boldsymbol{x}\sim\mathbb{N}_d}\left[R(\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x})\ \mathbb{1}\{|\boldsymbol{v}_1\cdot\boldsymbol{x}| \leq 2\theta|\boldsymbol{q}\cdot\boldsymbol{x}|\}\right]$ is equal to

$$\frac{d}{d\theta}\Big(\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_d}\left[R(\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x})\ \mathbb{1}\{|\boldsymbol{v}_1\cdot\boldsymbol{x}| \leq 2\theta|\boldsymbol{q}\cdot\boldsymbol{x}|\}\right]\Big)\Big|_{\theta=0} \leq 2\xi\Gamma(F_1) + \frac{2e^{-\xi^2/2}}{\pi}\,.$$

By picking $\xi = \sqrt{2\log(1+1/\Gamma(F_1))}$, the result follows since up to introducing $o(\theta)$ error we can bound the term $\mathbf{Pr}_{\boldsymbol{x}\sim\mathbb{N}_d}\left[R(\boldsymbol{v}_1\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x}) \neq R(\boldsymbol{r}\cdot\boldsymbol{x},\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x})\right]$ by its derivative with respect to θ (evaluated at 0) times θ. □

We can complete the proof of Lemma D.20 using Claim D.21. In order to bound the disagreement of the cones C_1 and C_2 we can start from C_1 and change one of its vectors at a time so that we can use Claim D.21 that can handle this case. For example, at the first step, we can swap $\boldsymbol{v}_1$ for $\boldsymbol{u}_1$ and use the triangle inequality to obtain that

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathbb{N}_d}[C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})] &\leq \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathbb{N}_d}[R(\boldsymbol{v}_1\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x}) \neq R(\boldsymbol{u}_1\cdot\boldsymbol{x},\boldsymbol{v}_2\cdot\boldsymbol{x}\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x})] \\ &\quad+ \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathbb{N}_d}[R(\boldsymbol{u}_1\cdot\boldsymbol{x},\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x}) \neq R(\boldsymbol{u}_1\cdot\boldsymbol{x},\boldsymbol{u}_2\cdot\boldsymbol{x}\ldots,\boldsymbol{u}_k\cdot\boldsymbol{x})] \\ &\leq c\cdot\theta\,\Gamma(F_1)\sqrt{\log(1/\Gamma(F_1)+1)} \\ &\quad+ \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathbb{N}_d}[R(\boldsymbol{u}_1\cdot\boldsymbol{x},\boldsymbol{v}_2\cdot\boldsymbol{x},\ldots,\boldsymbol{v}_k\cdot\boldsymbol{x}) \neq R(\boldsymbol{u}_1\cdot\boldsymbol{x},\boldsymbol{u}_2\cdot\boldsymbol{x}\ldots,\boldsymbol{u}_k\cdot\boldsymbol{x})]\,,\end{aligned}$$

where $F_1 = F_1(\boldsymbol{V}\boldsymbol{x})$ is the face with $\boldsymbol{v}_1\cdot\boldsymbol{x} = 0$ of the cone C_1. Notice that we have replaced $\boldsymbol{v}_1$ by $\boldsymbol{u}_1$ in the above bound. Our plan is to use the triangle inequality and continue replacing the vectors of C_1 by the vectors of C_2 sequentially. To make this formal we define the matrix $\boldsymbol{A}^{(i)} \in \mathbb{R}^{k \times d}$ whose first $i-1$ rows are the vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{i-1}$ and its last $k-i+1$ rows are the vectors $\boldsymbol{v}_i, \ldots, \boldsymbol{v}_k$, i.e.,

$$\boldsymbol{A}^{(i)}_j = \begin{cases} \boldsymbol{u}_j & \text{if} \quad 1 \le j \le i-1, \\ \boldsymbol{v}_j & \text{if} \quad i \le j \le k\,. \end{cases}$$

Notice that $\boldsymbol{A}^{(1)} = \boldsymbol{V}$ and $\boldsymbol{A}^{(k+1)} = \boldsymbol{U}$. Using the triangle inequality we obtain that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathbb{N}_d}[C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})] \le \sum_{i=1}^{k} \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathbb{N}_d}[R(\boldsymbol{A}^{(i)}\boldsymbol{x}) \neq R(\boldsymbol{A}^{(i+1)}\boldsymbol{x})].$$

Since the matrices $\boldsymbol{A}^{(i)}$ and $\boldsymbol{A}^{(i+1)}$ only differ on one row, we can use Claim D.21 to obtain the following bound:

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathbb{N}_d}[C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})] \le c \cdot \theta \cdot \sum_{i=1}^{k} \Gamma(F_i(\boldsymbol{A}^{(i)}\boldsymbol{x}))\sqrt{\frac{1}{\Gamma(F_i(\boldsymbol{A}^{(i)}\boldsymbol{x}))} + 1}\,.$$

We now observe that the Gaussian surface area $\Gamma(F_i(\boldsymbol{A}^{(i)}\boldsymbol{x}))$ is a continuous function of the matrix $\boldsymbol{A}^{(i)}$. By flattening the matrix $\boldsymbol{A}^{(i)}$ (since it is isomorphic to a vector $\boldsymbol{z} \in \mathbb{R}^{n^2}$) and letting $S_{\boldsymbol{z}}$ be the induced surface $\{\boldsymbol{x} : R(\boldsymbol{A}^{(i)}\boldsymbol{x}) = 1 \wedge \boldsymbol{v}_i \cdot \boldsymbol{x} = 0\}$, it suffices to show that

$$\lim_{\boldsymbol{w} \to \boldsymbol{z}} \int \phi_n(\boldsymbol{x}) \mathbb{1}\{\boldsymbol{x} \in S_{\boldsymbol{w}}\} d\mu(\boldsymbol{x}) = \int \phi_n(\boldsymbol{x}) \mathbb{1}\{\boldsymbol{x} \in S_{\boldsymbol{z}}\} d\mu(\boldsymbol{x})\,,$$

by the smoothness of the surface $S_{\boldsymbol{z}}$. Consider a sequence of functions (g_m) and vectors $(\boldsymbol{w}_m)$ so that $g_m(\boldsymbol{x}) = \phi_n(\boldsymbol{x})\mathbb{1}\{\boldsymbol{x} \in S_{\boldsymbol{w}_m}\}$ and $\lim_{m \to \infty} \boldsymbol{w}_m = \boldsymbol{z}$. Note that $|g_m(\boldsymbol{x})| \le 1$ everywhere. Hence, by the dominated convergence theorem, we have that

$$\lim_{m \to \infty} \int g_m(\boldsymbol{x}) d\mu(\boldsymbol{x}) = \int \lim_{m \to \infty} g_m(\boldsymbol{x}) d\mu(\boldsymbol{x}) = \int \phi_n(\boldsymbol{x}) \lim_{m \to \infty} \mathbb{1}\{\boldsymbol{x} \in S_{\boldsymbol{w}_m}\} d\mu(\boldsymbol{x})\,.$$

Since the sequence consists of smooth surfaces, we have that $\lim_{m\to\infty} \mathbb{1}\{\boldsymbol{x} \in S_{\boldsymbol{w}_m}\} = \mathbb{1}\{\boldsymbol{x} \in S_{\boldsymbol{z}}\}$ and so the Gaussian surface area is continuous with respect to the matrix $\boldsymbol{A}^{(i)}$ for any $i \in [k]$.

Also, as $\theta \to 0$, we have that $\boldsymbol{A}^{(i)} \to \boldsymbol{V}$. This is because the sequence of matrices $\boldsymbol{A}^{(i)}$ depends only on the vectors $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ for $j \in [k]$ and the following two properties hold true: $\theta = \max_{j\in[k]} \theta(\boldsymbol{v}_j, \boldsymbol{u}_j)$ and all the vectors are unit. Hence, as θ tends to zero, they tend to become the same vectors and so any matrix $\boldsymbol{A}^{(i)}$ tends to become $\boldsymbol{V}$. Therefore, taking this limit we obtain that for $\theta \to 0$ it holds that

$$\lim_{\theta\to 0} \frac{\mathbf{Pr}_{\boldsymbol{x}\sim\mathbb{N}_d}[C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})]}{\theta} \leq c \cdot \sum_{i=1}^{k} \Gamma(F_i(\boldsymbol{V}\boldsymbol{x}))\sqrt{\log\left(1/\Gamma(F_i(\boldsymbol{V}\boldsymbol{x})) + 1\right)}\,. \quad \text{(D.5)}$$

We will now use the following lemma that shows that the surface area of any homogeneous polyhedral cone is independent of the number of faces k and in fact is at most 1 for all k.

Lemma D.22 (Gaussian Surface Area of Homogeneous Cones Nazarov (2003b))**.** *Let C be a cone with apex at the origin (i.e., an intersection of arbitrarily many halfspaces all of whose boundaries contain the origin). Then C has Gaussian surface area $\Gamma(C)$ at most* 1.

Using Lemma D.22 we obtain that $\sum_{i=1}^k \Gamma(F_i(\boldsymbol{V}\boldsymbol{x})) \leq 1$. Next, we observe that, when the positive numbers $a_1, \ldots, a_k$ satisfy $\sum_{i=1}^k a_i \leq 1$, it holds that $\sum_{i=1}^k a_i \sqrt{\log(1/a_i)} \leq \sqrt{\sum_{i=1}^k a_i \log(1/a_i)} \leq \sqrt{\log(k)}$ (using the fact that the uniform distribution maximizes the entropy). Using this fact and Equation (D.5), we obtain

$$\lim_{\theta\to 0} \frac{\mathbf{Pr}_{\boldsymbol{x}\sim\mathbb{N}_d}[C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})]}{\theta} \leq c\sqrt{\log(k)}\,.$$

Thus, we have shown that, for sufficiently small θ, it holds that $\mathbf{Pr}_{\boldsymbol{x}\sim\mathbb{N}_d}[C_1(\boldsymbol{x}) \neq C_2(\boldsymbol{x})] \leq c\sqrt{\log(k)}\theta$, but, as we discussed in the start of the proof, the general bound follows directly from the bound for sufficiently small values of $\theta > 0$. □

D.3 Learning in Top-r Disagreement from Label Rankings

We prove the next result which corresponds to a proper learning algorithm for LSF in the presence of bounded noise with respect to the top-r disagreement.

Theorem D.23 (Proper Top-r Learning Algorithm)**.** *Fix $\eta \in [0, 1/2)$, $r \in [k]$ and $\epsilon, \delta \in (0, 1)$. Let $\mathcal{D}$ be an η-noisy linear label ranking distribution satisfying the assumptions of Definition 1.10. There exists an algorithm that draws $N = \widetilde{O}\left(\frac{d\ rk}{\epsilon(1-2\eta)^6}\log(1/\delta)\right)$ samples from $\mathcal{D}$, runs in* $\mathrm{poly}(N)$ *time and, with probability at least $1-\delta$, outputs a Linear Sorting function $h : \mathbb{R}^d \to \mathbb{S}_k$ that is ϵ-close in top-r disagreement to the target.*

The main result of this section is the next lemma, which directly implies the above theorem (using the same steps as the proof of Theorem D.18).

Lemma D.24 (Top-r Misclassification)**.** *Let $r \in [k]$. Consider two matrices $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{k \times d}$ and let $\mathcal{N}_d$ be the standard Gaussian in d dimensions. We have that*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\sigma_{1..r}(\boldsymbol{Ux}) \neq \sigma_{1..r}(\boldsymbol{Vx})] \leq c \cdot k \cdot r \cdot \sqrt{\log(kr)} \cdot \max_{i \neq j} \theta(\boldsymbol{U}_i - \boldsymbol{U}_j, \boldsymbol{V}_i - \boldsymbol{V}_j)\,,$$

where $c > 0$ is some universal constant.

Proof. Let us set $\sigma_{1..r}(\boldsymbol{Wx})$ denote the ordering of the top-r alternatives in the ranking $\sigma(\boldsymbol{Wx})$. Moreover, recall that $\sigma_\ell(\boldsymbol{Wx})$ denotes the alternative in the ℓ-th position of the ranking $\sigma(\boldsymbol{Wx})$. For two matrices $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{k \times d}$, we have that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[\sigma_{1..r}(\boldsymbol{Ux}) \neq \sigma_{1..r}(\boldsymbol{Vx})] = \sum_{j=1}^{k} \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}\left[\bigcup_{\ell=1}^{r} \{j = \sigma_\ell(\boldsymbol{Ux}), j \neq \sigma_\ell(\boldsymbol{Vx})\}\right].$$

The first step is to understand the geometry of the set $\bigcup_{\ell=1}^{r}\{\boldsymbol{x} : j = \sigma_\ell(\boldsymbol{Ux})\} = \{\boldsymbol{x} : j \in \sigma_{1..r}(\boldsymbol{Ux})\}$ for $j \in [k]$. We have that this set is equal to

$$\mathcal{T}_{\boldsymbol{U}}^{(j)} = \bigcup_{S \subseteq [k] : |S| \leq r-1} \bigcap_{i \in S} \{\boldsymbol{x} : (\boldsymbol{U}_i - \boldsymbol{U}_j) \cdot \boldsymbol{x} \geq 0\} \cap \bigcap_{i \notin S} \{\boldsymbol{x} : (\boldsymbol{U}_i - \boldsymbol{U}_j) \cdot \boldsymbol{x} \leq 0\}\,.$$

In words, $\mathcal{T}_{\boldsymbol{U}}^{(j)}$ iterates over any possible collection of alternatives that can win the element j (they lie in the set of top elements S) and the remaining elements lose when compared with j (they lie in the complement set $[k] \setminus S$). Overloading the notation, let us define the mapping $T(\boldsymbol{t}) = T(t_1, ..., t_k) = \sum_{S \subseteq [k]:|S| \leq r-1} \prod_{i \in S} \mathbb{1}\{t_i \geq 0\} \prod_{i \notin S} \mathbb{1}\{t_i \leq 0\}$. Using this mapping, we can define the indicator of the set $T_{\boldsymbol{U}}^{(j)}$ as $T((\boldsymbol{U}_1 - \boldsymbol{U}_j) \cdot \boldsymbol{x}, \ldots, (\boldsymbol{U}_k - \boldsymbol{U}_j) \cdot \boldsymbol{x})$. The top-$r$ disagreement $\mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{N}_d}[j \in \sigma_{1..r}(\boldsymbol{U}\boldsymbol{x}), j \notin \sigma_{1..r}(\boldsymbol{V}\boldsymbol{x})]$ is equal to:

$$\underset{\boldsymbol{x} \sim \mathcal{N}_d}{\mathbf{Pr}}[T((\boldsymbol{U}_1 - \boldsymbol{U}_j) \cdot \boldsymbol{x}, ..., (\boldsymbol{U}_k - \boldsymbol{U}_j) \cdot \boldsymbol{x}) = 1, T((\boldsymbol{V}_1 - \boldsymbol{V}_j) \cdot \boldsymbol{x}, ..., (\boldsymbol{V}_k - \boldsymbol{V}_j) \cdot \boldsymbol{x}) = 0] \,.$$

So we have that

$$\begin{aligned} \underset{\boldsymbol{x} \sim \mathcal{N}_d}{\mathbf{Pr}}[\sigma_{1..r}(\boldsymbol{U}\boldsymbol{x}) \neq \sigma_{1..r}(\boldsymbol{V}\boldsymbol{x})] &= \sum_{j=1}^{k} \underset{\boldsymbol{x} \sim \mathcal{N}_d}{\mathbf{Pr}}[T_j(\boldsymbol{U}\boldsymbol{x}) = 1, T_j(\boldsymbol{V}\boldsymbol{x}) = 0] \\ &\leq \sum_{j=1}^{k} \underset{\boldsymbol{x} \sim \mathcal{N}_d}{\mathbf{Pr}}[T_j(\boldsymbol{U}\boldsymbol{x}) \neq T_j(\boldsymbol{V}\boldsymbol{x})] \,. \end{aligned}$$

In order to show the desired bound, it suffices to prove the following two lemmas.

Lemma D.25 (Disagreement Region)**.** *Consider a positive integer $r \leq k$. Fix $j \in [k]$ and let $\theta = \max_{i \in [k]} \theta(\boldsymbol{U}_i - \boldsymbol{U}_j, \boldsymbol{V}_i - \boldsymbol{V}_j)$. Then it holds that*

$$\lim_{\theta \to 0} \frac{\mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{N}_d}[T_j(\boldsymbol{U}\boldsymbol{x}) \neq T_j(\boldsymbol{V}\boldsymbol{x})]}{\theta} \leq c \cdot \sum_{i \in [k]} \Gamma(F_i^j) \sqrt{\log\left(\frac{1}{\Gamma(F_i^j)} + 1\right)} \,,$$

where $c > 0$ is some constant and F_i^j is the surface $\{\boldsymbol{x} : j \in \sigma_{1..r}(\boldsymbol{V}\boldsymbol{x})\} \cap \{\boldsymbol{x} : \boldsymbol{V}_i \cdot \boldsymbol{x} = \boldsymbol{V}_j \cdot \boldsymbol{x}\}$ for the matrix $\boldsymbol{V} \in \mathbb{R}^{k \times d}$.

and,

Lemma D.26. *Let F_i^j, r, k as in the previous lemma. It holds that*

$$\sum_{i\in[k]}\sum_{j\in[k]} \Gamma(F_i^j) \le 2kr\,.$$

Applying these two lemmas with $\theta = \max_{i\neq j} \theta(\boldsymbol{U}_i - \boldsymbol{U}_j, \boldsymbol{V}_i - \boldsymbol{V}_j)$, we get that

$$Z := \lim_{\theta\to 0} \frac{\sum_{j\in[k]} \mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}_d}[T_j(\boldsymbol{Ux}) \neq T_j(\boldsymbol{Vx})]}{\theta} \le c \cdot \sum_{j\in[k]}\sum_{i\in[k]} \Gamma(F_i^j)\sqrt{\log\left(\frac{1}{\Gamma(F_i^j)}+1\right)}\,.$$

Let us set $\Gamma'(F_i^j) = \Gamma(F_i^j)/(2kr)$. Then we have that

$$Z \le 2ckr \cdot \sum_{j\in[k]}\sum_{i\in[k]} \Gamma'(F_i^j)\sqrt{\log\left(\frac{1}{2kr\cdot\Gamma'(F_i^j)}+1\right)}\,.$$

It suffices to bound the quantity

$$\sum_{j\in[k]}\sum_{i\in[k]} \Gamma'(F_i^j)\sqrt{\log\left(\frac{1}{\Gamma'(F_i^j)}+1\right)} = O\left(kr\sqrt{\log(kr)}\right)\,,$$

where we used a similar “entropy-like” inequality as we did in the top-1 case. This yields (by recalling that it is sufficient to consider only the case of arbitrarily small angles, as in the top-1 case) that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[\sigma_{1..r}(\boldsymbol{Ux}) \neq \sigma_{1..r}(\boldsymbol{Vx})] \le c\, rk\,\sqrt{\log(kr)}\cdot \max_{i\neq j} \theta(\boldsymbol{U}_i - \boldsymbol{U}_j, \boldsymbol{V}_i - \boldsymbol{V}_j)\,,$$

for some universal constant c. □

The proof of Lemma D.25

We proceed with the proof of the key lemma concerning the disagreement region. We first show the following claim where we only change a single vector. Recall that

$$T(\boldsymbol{V}\boldsymbol{x}) = \sum_{S:|S|\le r-1} \prod_{i\in S} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \ge 0\} \prod_{i\notin S} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \le 0\}\,.$$

We will be interested in the surface $F_1 := F_1(\boldsymbol{V}\boldsymbol{x}) = T(\boldsymbol{V}\boldsymbol{x})\mathbb{1}\{\boldsymbol{v}_1 \cdot \boldsymbol{x} = 0\}$.

Claim D.27. *Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k \in \mathbb{R}^d$ and $\boldsymbol{r} \in \mathbb{R}^d$ with $\theta(\boldsymbol{v}_1, \boldsymbol{r}) \le \theta$ for some sufficiently small $\theta \in (0, \pi/2)$. It holds that*

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{N}_d}[T(\boldsymbol{v}_1\cdot\boldsymbol{x}, \dots, \boldsymbol{v}_k\cdot\boldsymbol{x}) \ne T(\boldsymbol{r}\cdot\boldsymbol{x}, \boldsymbol{v}_2\cdot\boldsymbol{x}, \dots, \boldsymbol{v}_k\cdot\boldsymbol{x})] \le c\cdot\theta\cdot\Gamma(F_1)\sqrt{\log\left(\frac{1}{\Gamma(F_1)}+1\right)}\,,$$

where F_1 is the surface $T(\boldsymbol{V}\boldsymbol{x}) \cap \{\boldsymbol{x} : \boldsymbol{v}_1 \cdot \boldsymbol{x} = 0\}$ and c is some universal constant.

Proof. We first decompose the sum of $T(\boldsymbol{V}\boldsymbol{x})$ depending on whether $1 \in S$ or not. Hence, we have that $T(\boldsymbol{v}_1\cdot\boldsymbol{x}, \dots, \boldsymbol{v}_k\cdot\boldsymbol{x}) = T^+(\boldsymbol{v}_1\cdot\boldsymbol{x}, \dots, \boldsymbol{v}_k\cdot\boldsymbol{x}) + T^-(\boldsymbol{v}_1\cdot\boldsymbol{x}, \dots, \boldsymbol{v}_k\cdot\boldsymbol{x})$ where

$$\begin{aligned}
&T^+(\boldsymbol{v}_1\cdot\boldsymbol{x}, \dots, \boldsymbol{v}_k\cdot\boldsymbol{x}) \\
&= \sum_{S\subseteq[k]:|S|\le r-1, 1\in S} \prod_{i\in S} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \ge 0\} \prod_{i\notin S} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \le 0\} \\
&= \sum_{S\subseteq[k]:|S|\le r-1, 1\in S} \mathbb{1}\{\boldsymbol{v}_1 \cdot \boldsymbol{x} \ge 0\} \cdot \prod_{i\in S\setminus\{1\}} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \ge 0\} \prod_{i\notin S} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \le 0\} \\
&= \mathbb{1}\{\boldsymbol{v}_1 \cdot \boldsymbol{x} \ge 0\} \cdot \sum_{S\subseteq[k]:|S|\le r-1, 1\in S} \prod_{i\in S\setminus\{1\}} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \ge 0\} \prod_{i\notin S} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \le 0\} \\
&=: \mathbb{1}\{\boldsymbol{v}_1 \cdot \boldsymbol{x} \ge 0\} \cdot G^+(\boldsymbol{v}_2\cdot\boldsymbol{x}, \dots, \boldsymbol{v}_k\cdot\boldsymbol{x})\,,
\end{aligned}$$

and similarly

$$\begin{aligned}
&T^-(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \\
&= \mathbb{1}\{\boldsymbol{v}_1 \cdot \boldsymbol{x} \le 0\} \cdot \sum_{S \subseteq [k] : |S| \le r-1, 1 \notin S} \prod_{i \in S} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \ge 0\} \prod_{i \notin S \setminus \{1\}} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \le 0\} \\
&=: \mathbb{1}\{\boldsymbol{v}_1 \cdot \boldsymbol{x} \le 0\} \cdot G^-(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \,.
\end{aligned}$$

Notice that the indicator G^s does not depend on the alternative 1 for $s \in \{-, +\}$. Since $T : \mathbb{R}^k \to \{0, 1\}$, we have that

$$\begin{aligned}
&\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d} [T(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \ne T(\boldsymbol{r} \cdot \boldsymbol{x}, \boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x})] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d} [|T(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) - T(\boldsymbol{r} \cdot \boldsymbol{x}, \boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x})|] \\
&\le \sum_{s \in \{-, +\}} \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d} [|T^s(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) - T^s(\boldsymbol{r} \cdot \boldsymbol{x}, \boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x})|] \\
&= \sum_{s \in \{-, +\}} \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_d} [G^s(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \dots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \cdot |\mathbb{1}\{s \cdot \boldsymbol{v}_1 \cdot \boldsymbol{x} \ge 0\} - \mathbb{1}\{s \cdot \boldsymbol{r} \cdot \boldsymbol{x} \ge 0\}|] \,.
\end{aligned}$$

Let us focus on the case $s = +$. The difference between the two indicators in the last line of the above equation corresponds to the event that the halfspaces $\mathbb{1}\{\boldsymbol{v}_1 \cdot \boldsymbol{x} \ge 0\}$ and $\mathbb{1}\{\boldsymbol{r} \cdot \boldsymbol{x} \ge 0\}$ disagree. Hence, we have that $|\mathbb{1}\{\boldsymbol{v}_1 \cdot \boldsymbol{x} \ge 0\} - \mathbb{1}\{\boldsymbol{r} \cdot \boldsymbol{x} \ge 0\}| = \mathbb{1}\{(\boldsymbol{v}_1 \cdot \boldsymbol{x})(\boldsymbol{r} \cdot \boldsymbol{x}) < 0\}$. Note that the above indicator depends on both $\boldsymbol{v}_1$ and $\boldsymbol{r}$. We would like to work only with one of these two vectors. To this end, let us introduce $\boldsymbol{q}$, the normalized projection of $\boldsymbol{r}$ onto the orthogonal complement of $\boldsymbol{v}_1$, i.e., $\boldsymbol{q} = \mathrm{proj}_{\boldsymbol{v}_1^\perp} \boldsymbol{r} / \|\mathrm{proj}_{\boldsymbol{v}_1^\perp} \boldsymbol{r}\|_2$. We have that $\boldsymbol{v}_1$ and $\boldsymbol{q}$ is an orthonormal basis of the subspace spanned by the vectors $\boldsymbol{v}_1$ and $\boldsymbol{r}$. Notice that $\boldsymbol{r} = \cos\theta(\boldsymbol{v}_1, \boldsymbol{r})\boldsymbol{v}_1 + \sin\theta(\boldsymbol{v}_1, \boldsymbol{r})\boldsymbol{q}$, by the construction of $\boldsymbol{q}$. Our goal is to understand the structure of the region $(\boldsymbol{v}_1 \cdot \boldsymbol{x})(\boldsymbol{r} \cdot \boldsymbol{x}) < 0$. This set is equal to

$$\{0 < \boldsymbol{v}_1 \cdot \boldsymbol{x} < -(\boldsymbol{q} \cdot \boldsymbol{x}) \tan\theta(\boldsymbol{v}_1, \boldsymbol{r})\} \cup \{-(\boldsymbol{q} \cdot \boldsymbol{x}) \tan\theta(\boldsymbol{v}_1, \boldsymbol{r}) < \boldsymbol{v}_1 \cdot \boldsymbol{x} < 0\} \,.$$

To see this, we have that $(\boldsymbol{v}_1 \cdot \boldsymbol{x})(\boldsymbol{r} \cdot \boldsymbol{x}) = (\boldsymbol{v}_1 \cdot \boldsymbol{x})(\cos\theta(\boldsymbol{v}_1, \boldsymbol{r})\boldsymbol{v}_1 \cdot \boldsymbol{x} + \sin\theta(\boldsymbol{v}_1, \boldsymbol{r})\boldsymbol{q} \cdot \boldsymbol{x})$.

This quantity must be negative. The left-hand set considers the case where $\boldsymbol{v}_1 \cdot \boldsymbol{x} > 0$ and so $\tan\theta(\boldsymbol{v}_1, \boldsymbol{r})(\boldsymbol{q} \cdot \boldsymbol{x}) < -\boldsymbol{v}_1 \cdot \boldsymbol{x}$. We obtain the right-hand set in a similar way. Thus, we have that the disagreement region $(\boldsymbol{v}_1 \cdot \boldsymbol{x})(\boldsymbol{r} \cdot \boldsymbol{x}) < 0$ is a subset of the region $\{|\boldsymbol{v}_1 \cdot \boldsymbol{x}| \leq |\boldsymbol{q} \cdot \boldsymbol{x}| \tan\theta(\boldsymbol{v}_1, \boldsymbol{r})\}$. Since $\tan\theta(\boldsymbol{v}_1, \boldsymbol{r}) \leq \theta$ and we have that θ is sufficiently small we can also replace the above region by the larger region: $\{|\boldsymbol{v}_1 \cdot \boldsymbol{x}| \leq 2\theta|\boldsymbol{q} \cdot \boldsymbol{x}|\}$. Therefore, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_d}\left[G^+(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x})\, \mathbb{1}\{(\boldsymbol{v}_1 \cdot \boldsymbol{x})(\boldsymbol{r} \cdot \boldsymbol{x}) < 0\}\right] \\
&\qquad \leq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_d}\left[G^+(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x})\, \mathbb{1}\{|\boldsymbol{v}_1 \cdot \boldsymbol{x}| \leq 2\theta|\boldsymbol{q} \cdot \boldsymbol{x}|\}\right] .
\end{aligned}$$

From this point, the proof goes as in the top-1 case. In total, we will get that

$$\begin{aligned}
&\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[T(\boldsymbol{v}_1 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}) \neq T(\boldsymbol{r} \cdot \boldsymbol{x}, \boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x})] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_d}\left[(G^+(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}) + G^-(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}))\, |\boldsymbol{q} \cdot \boldsymbol{x}|\, \delta(|\boldsymbol{v}_1 \cdot \boldsymbol{x}|)\right] \\
&\leq 2 \int_{\boldsymbol{x}\in F_1} \phi_d(\boldsymbol{x}) |\boldsymbol{q} \cdot \boldsymbol{x}| d\mu(\boldsymbol{x}) \\
&\leq 2 \int_{\boldsymbol{x}\in F_1} \phi_d(\boldsymbol{x}) |\boldsymbol{q} \cdot \boldsymbol{x}| \mathbb{1}\{|\boldsymbol{q} \cdot \boldsymbol{x}| \leq \xi\} d\mu(\boldsymbol{x}) + 2 \int_{\boldsymbol{x}\in F_1} \phi_d(\boldsymbol{x}) |\boldsymbol{q} \cdot \boldsymbol{x}| \mathbb{1}\{|\boldsymbol{q} \cdot \boldsymbol{x}| \geq \xi\} d\mu(\boldsymbol{x}) \\
&\leq 2\xi \int_{\boldsymbol{x}\in F_1} \phi_d(\boldsymbol{x}) d\mu(\boldsymbol{x}) + 2 \int_{\boldsymbol{x}\in F_1} \phi_d(\boldsymbol{x}) |\boldsymbol{q} \cdot \boldsymbol{x}| \mathbb{1}\{|\boldsymbol{q} \cdot \boldsymbol{x}| \geq \xi\} d\mu(\boldsymbol{x})\,,
\end{aligned}$$

where $d\mu(\boldsymbol{x})$ is the standard surface measure in $\mathbb{R}^d$. Let us explain the first inequality above. Note that the space induced by $G^-(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x})$ contains the space induced by $G^+(\boldsymbol{v}_2 \cdot \boldsymbol{x}, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x})$. Hence, in the integration, we can integrate over the surface $F_1 = T(\boldsymbol{V}\boldsymbol{x}) \cap \mathbb{1}\{\boldsymbol{x} : \boldsymbol{v}_1 \cdot \boldsymbol{x} = 0\}$ twice. Essentially, this surface corresponds to $\mathbb{1}\{\boldsymbol{v}_1 \cdot \boldsymbol{x} = 0\} \cdot \sum_{S \subseteq [k]\setminus\{1\}:|S|\leq r-1} \prod_{i\in S} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \geq 0\} \prod_{i\notin S} \mathbb{1}\{\boldsymbol{v}_i \cdot \boldsymbol{x} \leq 0\}$. Applying the steps of the top-1 case, we can obtain the desired bound in terms of the Gaussian surface area of F_1. □

Next, for fixed $j \in [k]$, we can apply the above claim sequentially (as we did in the end of the top-1 case) to get

$$\lim_{\theta \to 0} \frac{\mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{N}_d}[T_j(\boldsymbol{U}\boldsymbol{x}) \neq T_j(\boldsymbol{V}\boldsymbol{x})]}{\theta} \leq c \cdot \sum_{i \in [k]} \Gamma(F_i^j) \sqrt{\log\left(\frac{1}{\Gamma(F_i^j)} + 1\right)},$$

for some small constant $c > 0$.

The proof of Lemma D.26

Using the above result, we get that it suffices to control the value $\Gamma(F_i^j)$, where F_i^j is the surface of $T_j(\boldsymbol{V}\boldsymbol{x}) \cap \{\boldsymbol{x} : \boldsymbol{V}_i \cdot \boldsymbol{x} = \boldsymbol{V}_j \cdot \boldsymbol{x}\}$ for the matrix $\boldsymbol{V}$ and $i, j \in [k]$. We next have to control the Gaussian surface area of the induced shape, i.e., the quantity

$$\Gamma(\{\boldsymbol{x} : j \in \sigma_{1..r}(\boldsymbol{V}\boldsymbol{x})\} \cap \{\boldsymbol{x} : \boldsymbol{V}_i \cdot \boldsymbol{x} = \boldsymbol{V}_j \cdot \boldsymbol{x}\}).$$

To this end, we give the next lemma.

Lemma D.28. *Let $r \leq k$ with $r, k \in \mathbb{N}$. For any matrix $\boldsymbol{V} \in \mathbb{R}^{k \times d}$ and $i, j \in [k]$, there exists a matrix $\boldsymbol{Q} = \boldsymbol{Q}^{(i)} \in \mathbb{R}^{k \times d}$ which depends only on i such that*

$$\Gamma(F_i^j) := \Gamma(\{\boldsymbol{x} : j \in \sigma_{1..r}(\boldsymbol{V}\boldsymbol{x})\} \cap \{\boldsymbol{x} : \boldsymbol{V}_i \cdot \boldsymbol{x} = \boldsymbol{V}_j \cdot \boldsymbol{x}\}) \leq 2 \cdot \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d}[j \in \sigma_{1..r}(\boldsymbol{Q}\boldsymbol{x})].$$

Before proving this result, let us see how to apply it in order to get Lemma D.26.

We will have that

$$\begin{aligned}\sum_{i\in[k]}\sum_{j\in[k]}\Gamma(F_i^j) &= \sum_{i\in[k]}\sum_{j\in[k]}\Gamma(\{\boldsymbol{x} : j \in \sigma_{1..r}(\boldsymbol{V}\boldsymbol{x})\} \cap \{\boldsymbol{x} : \boldsymbol{V}_i \cdot \boldsymbol{x} = \boldsymbol{V}_j \cdot \boldsymbol{x}\}) \\ &\leq 2\sum_{i\in[k]}\sum_{j\in[k]}\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[j \in \sigma_{1..r}(\boldsymbol{Q}^{(i)}\boldsymbol{x})] \\ &= 2\sum_{i\in[k]}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_d}[|\sigma_{1..r}(\boldsymbol{Q}^{(i)}\boldsymbol{x})|] \\ &= 2\sum_{i\in[k]} r \\ &= 2kr\,.\end{aligned}$$

Proof of Lemma D.28. For this proof, we fix $i, j \in [k]$. The first step is to design the matrix $\boldsymbol{Q}$. As a first observation, we can subtract the vector $\boldsymbol{V}_i$ from each weight vector and do not affect the resulting orderings. Second, we can assume that the weight vectors that correspond to indices which j beats are unit. Let us be more specific Assume that initially we have that

$$(\boldsymbol{V}_j - \boldsymbol{V}_\ell) \cdot \boldsymbol{x} \geq 0\,.$$

The first observation gives that

$$(\boldsymbol{V}_j - \boldsymbol{V}_i) \cdot \boldsymbol{x} \geq (\boldsymbol{V}_\ell - \boldsymbol{V}_i) \cdot \boldsymbol{x}\,.$$

Let us set $\widetilde{\boldsymbol{Q}}$ the intermediate matrix with rows $\boldsymbol{V}_j - \boldsymbol{V}_i$. The second observation states that the inequalities where j beats some index ℓ are not affected by normalization. Note that $\widetilde{\boldsymbol{Q}}_j \cdot \boldsymbol{x} = 0$ and hence $\widetilde{\boldsymbol{Q}}_\ell \cdot \boldsymbol{x} \leq 0$. Hence, dividing with non-negative numbers will not affect the order of these two values, i.e.,

$$\frac{\widetilde{\boldsymbol{Q}}_j \cdot \boldsymbol{x}}{\|\widetilde{\boldsymbol{Q}}_j\|_2} \geq \frac{\widetilde{\boldsymbol{Q}}_\ell \cdot \boldsymbol{x}}{\|\widetilde{\boldsymbol{Q}}_\ell\|_2}\,.$$

Note that the above ordering is $\boldsymbol{x}$-dependent, since the indices that j beats depend on $\boldsymbol{x}$. However, we can normalize any row of $\widetilde{\boldsymbol{Q}}$ without affecting the fact that the element j is top-r (since the sign of the inner products is not affected by normalization). This transformation yields a matrix $\boldsymbol{Q} = \boldsymbol{Q}^{(i)}$ and depends only on i (crucially, it is independent of j). For simplicity, we will omit the index i in what follows. For this matrix, we have that

$$\{\boldsymbol{x} : j \in \sigma_{1..r}(\boldsymbol{Q}\boldsymbol{x}), \boldsymbol{Q}_j \cdot \boldsymbol{x} = 0\} = \{\boldsymbol{x} : j \in \sigma_{1..r}(\boldsymbol{V}\boldsymbol{x}), \boldsymbol{V}_i \cdot \boldsymbol{x} = \boldsymbol{V}_j \cdot \boldsymbol{x}\} \,.$$

We will now prove that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathcal{N}_d} [j \in \sigma_{1..r}(\boldsymbol{Q}\boldsymbol{x})] \geq \frac{\Gamma(F_i^j)}{2} \,.$$

Let us fix some $\boldsymbol{x}$ and set $\boldsymbol{x}^{\|} = \mathrm{proj}_{\boldsymbol{Q}_j} \boldsymbol{x}$ and $x^{\perp} = \mathrm{proj}_{\boldsymbol{Q}_j^{\perp}} \boldsymbol{x}$. We assume that $\boldsymbol{x}$ lies in the set $\{\boldsymbol{x} : j \in \sigma_{1..r}(\boldsymbol{Q}\boldsymbol{x})\}$. This implies that there exist an index set I of size at least $k - r$ so that if $\ell \in I$ then

$$\boldsymbol{Q}_j \cdot \boldsymbol{x}^{\|} + \boldsymbol{Q}_j \cdot \boldsymbol{x}^{\perp} \geq \boldsymbol{Q}_\ell \cdot \boldsymbol{x}^{\|} + \boldsymbol{Q}_\ell \cdot \boldsymbol{x}^{\perp} \,.$$

Let us condition on the event

$$\boldsymbol{Q}_j \cdot \boldsymbol{x}^{\perp} \geq \boldsymbol{Q}_\ell \cdot \boldsymbol{x}^{\perp} \,.$$

We hence get that

$$\boldsymbol{Q}_j \cdot \boldsymbol{x}^{\|} = (\boldsymbol{Q}_j \cdot \boldsymbol{Q}_j) \cdot (\boldsymbol{Q}_j \cdot \boldsymbol{x}) \geq \boldsymbol{Q}_\ell \cdot \boldsymbol{x}^{\|} = (\boldsymbol{Q}_\ell \cdot \boldsymbol{Q}_j) \cdot (\boldsymbol{Q}_j \cdot \boldsymbol{x})$$

Using that $\boldsymbol{Q}_j$ is unit, that the inner product between $\boldsymbol{Q}_\ell$ and $\boldsymbol{Q}_j$ is at most one and that $\boldsymbol{Q}_j \cdot \boldsymbol{x}$ is a univariate Gaussian, we get that

$$\mathop{\mathbf{Pr}}_{z \sim \mathcal{N}(0,1)} [z \cdot (1 - \boldsymbol{Q}_\ell \cdot \boldsymbol{Q}_j) \geq 0] = 1/2 \,.$$

The above discussion implies that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[j \in \sigma_{1..r}(\boldsymbol{Q}\boldsymbol{x})] = \mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[(\forall \ell \in I)\ \boldsymbol{Q}_j \cdot \boldsymbol{x}^{\|} + \boldsymbol{Q}_j \cdot \boldsymbol{x}^{\perp} \geq \boldsymbol{Q}_\ell \cdot \boldsymbol{x}^{\|} + \boldsymbol{Q}_\ell \cdot \boldsymbol{x}^{\perp}]$$

and so $\mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{N}_d}[j \in \sigma_{1..r}(\boldsymbol{Q}\boldsymbol{x})]$ equals to

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[(\forall \ell \in I)\ \boldsymbol{Q}_j \cdot \boldsymbol{x}^{\|} \geq \boldsymbol{Q}_j \cdot \boldsymbol{x}^{\|} \mid (\forall \ell \in I)\ \boldsymbol{Q}_j \cdot \boldsymbol{x}^{\perp} \geq \boldsymbol{Q}_\ell \cdot \boldsymbol{x}^{\perp}] \cdot$$
$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[(\forall \ell \in I)\ \boldsymbol{Q}_j \cdot \boldsymbol{x}^{\perp} \geq \boldsymbol{Q}_\ell \cdot \boldsymbol{x}^{\perp}] .$$

However, in the above product, we have that the first term is $1/2$ and the second term is the probability that $j \in \sigma_{1..r}(\boldsymbol{Q}\boldsymbol{x}^{\perp})$, i.e.,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim\mathcal{N}_d}[j \in \sigma_{1..r}(\boldsymbol{Q}\boldsymbol{x})] \geq \frac{\mathbf{Pr}[j \in \sigma_{1..r}(\boldsymbol{Q}\boldsymbol{x}^{\perp})]}{2} = \Gamma(F_i^j)/2 ,$$

since the space in the RHS is low-dimensional and corresponds to the desired surface. □

D.4 Distribution-Free Lower Bounds for Top-1 Disagreement Error

We begin with some definitions concerning the PAC Label Ranking setting. Let $\mathcal{X}$ be an instance space and $\mathcal{Y} = \mathbb{S}_k$ be the space of labels, which are rankings over k elements. A sorting function or hypothesis is a mapping $h : \mathcal{X} \to \mathbb{S}_k$. We denote by $h_1(x)$ the top-1 element of the ranking $h(x)$. A hypothesis class is a set of classifiers $\mathcal{H} \subset \mathbb{S}_k^{\mathcal{X}}$.

Top-1 Disagreement Error. The top-1 disagreement error with respect to a joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathbb{S}_k$ equals to the probability $\mathbf{Pr}_{(x,\sigma)\sim\mathcal{D}}[h_1(x) \neq \sigma^{-1}(1)]$. We mainly consider learning in the **realizable** case, which means that there is $h^\star \in \mathcal{H}$ which has (almost surely) zero error. Therefore, we can focus on the marginal distribution $\mathcal{D}_x$ over $\mathcal{X}$ and denote the top-1 disagreement error of a sorting function

h with respect to the true hypothesis $h^\star$ by $\mathrm{Err}_{\mathcal{D}_x,h^\star}(h) := \mathbf{Pr}_{x\sim\mathcal{D}_x}\left[h_1(x) \neq h_1^\star(x)\right]$.

A learning algorithm is a function $\mathcal{A}$ that receives a training set of m instances, $S \in \mathcal{X}^m$, together with their labels according to $h^\star$. We denote the restriction of $h^\star$ to the instances in S by $h^\star|_S$. The output of the algorithm $\mathcal{A}$, denoted $\mathcal{A}(S, h^\star|_S)$ is a sorting function. A learning algorithm is proper if it always outputs a hypothesis from $\mathcal{H}$.

The top-1 PAC Label Ranking sample complexity of a learning algorithm $\mathcal{A}$ is the function $m^{(1)}_{\mathcal{A},\mathcal{H}}$ defined as follows: for every $\epsilon, \delta > 0$, $m^{(1)}_{\mathcal{A},\mathcal{H}}(\epsilon,\delta)$ is the minimal integer such that for every $m \geq m^{(1)}_{\mathcal{A},\mathcal{H}}(\epsilon,\delta)$, every distribution $\mathcal{D}_x$ on $\mathcal{X}$, and every target hypothesis $h^\star \in \mathcal{H}$, $\mathbf{Pr}_{S\sim\mathcal{D}_x^m}\left[\mathrm{Err}_{\mathcal{D}_x,h^\star}(\mathcal{A}(S,h^\star|_S)) > \epsilon\right] \leq \delta$. In this case, we say that the learning algorithm (ϵ,δ)-learns the class of sorting functions $\mathcal{H}$ with respect to the top-1 disagreement error. If no integer satisfies the inequality above, define $m^{(1)}_{\mathcal{A}}(\epsilon,\delta) = \infty$. $\mathcal{H}$ is learnable with $\mathcal{A}$ if for all ϵ and δ the sample complexity is finite. The **top-1 PAC Label Ranking sample complexity** of a class $\mathcal{H}$ is $m^{(1)}_{\mathrm{PAC},\mathcal{H}}(\epsilon,\delta) = \inf_{\mathcal{A}} m^{(1)}_{\mathcal{A},\mathcal{H}}(\epsilon,\delta)$, where the infimum is taken over all learning algorithms. Clearly, the above top-1 definition can be extended to the top-r setting.

In this section, we show the next result. We denote by $\mathcal{L}_{d,k}$ the class of Linear Sorting functions in d dimensions with k labels.

Theorem D.29. *In the realizable PAC Label Ranking setting, any algorithm that (ϵ,δ)-learns the class $\mathcal{L}_{d,k}$ with respect to the top-1 disagreement error requires at least $\Omega((dk + \log(1/\delta))/\epsilon)$ samples.*

Top-1 Ranking Natarajan Dimension

In order to establish the above result, we introduce a variant of the standard Natarajan dimension Natarajan (1989); Ben-David et al. (1992); Daniely et al. (2011); Daniely and Shalev-Shwartz (2014). For a ranking π, we will also let $L_1(\pi)$ its top-1 element and $L_{3..k}(\pi)$ the ranking after deleting its top-2 part.

Definition D.30 (Top-1 Ranking Natarajan Dimension). *Let $\mathcal{H} \subseteq \mathbb{S}_k^{\mathcal{X}}$ be a hypothesis class of sorting functions and let $S \subseteq \mathcal{X}$. We say that $\mathcal{H}$ N-shatters S if there exist* *two mappings $f_1, f_2 : S \to \mathbb{S}_k$ such that for every $y \in S$, $L_1(f_1(y)) \neq L_1(f_2(y))$ and $L_{3..k}(f_1(y)) = L_{3..k}(f_2(y))$ and for every $T \subseteq S$, there exists a sorting function $g \in \mathcal{H}$ such that*

$$(i)\ \forall x \in T,\ \ g(x) = f_1(x), \text{ and } (ii)\ \forall x \in S \setminus T,\ \ g(x) = f_2(x)\,.$$

*The **top-1 Ranking Natarajan dimension** of $\mathcal{H}$, denoted $d_N^{(1)}(\mathcal{H})$ is the maximal cardinality of a set that is N-shattered by $\mathcal{H}$.*

First, we connect PAC Label Ranking learnability to the top-1 disagreement error with the notion of top-1 Ranking Natarajan dimension.

Theorem D.31 (Top-1-Natarajan Lower Bounds Sample Complexity)**.** *In the realizable PAC Label Ranking setting, we have for every hypothesis class $\mathcal{H} \subseteq \mathbb{S}_k^{\mathcal{X}}$*

$$m^{(1)}_{\mathrm{PAC},\mathcal{H}}(\epsilon, \delta) = \Omega\left(\frac{d_N^{(1)}(\mathcal{H}) + \ln(1/\delta)}{\epsilon}\right)\,.$$

Proof. Let $\mathcal{H} \subseteq \mathbb{S}_k^{\mathcal{X}}$ be a class of sorting functions of top-1-Natarajan dimension $d_N^{(1)} = d_N$. Consider the binary hypothesis class $\mathcal{H}_{\mathrm{bin}} = \{0,1\}^{[d_N]}$ which contains all the classifiers from $[d_N] = \{1, ..., d_N\}$ to $\{0,1\}$. It suffices to show the following.

Claim D.32. *It holds that $m^{(1)}_{\mathrm{PAC},\mathcal{H}}(\epsilon, \delta) \geq m_{\mathrm{PAC},\mathcal{H}_{\mathrm{bin}}}(\epsilon, \delta)$.*

This is sufficient since we have that $m_{\mathrm{PAC},\mathcal{H}_{\mathrm{bin}}}(\epsilon, \delta) = \Omega\left(\frac{\mathrm{VC}(\mathcal{H}_{\mathrm{bin}}) + \ln(1/\delta)}{\epsilon}\right)$ and $\mathrm{VC}(\mathcal{H}_{\mathrm{bin}}) = d_N$. Let us now prove the claim.

We assume that the instance space is the set $\mathcal{X}$. Assume that A is a learning algorithm for the hypothesis class $\mathcal{H} \subseteq \mathbb{S}_k^{\mathcal{X}}$ and A_{bin} is a learning algorithm for the associated binary class $\mathcal{H}_{\mathrm{bin}}$. It suffices to show that A requires at least as many samples as A_{bin}. In fact, we will show that whenever A_{bin} errs, so does A. Let $S = \{s_1, ..., s_{d_N}\}, f_0, f_1$ be the set and the two functions that witness that the top-1-Natarajan dimension of $\mathcal{H}$ is d_N. Given a training set $(x_i, y_i)_{i \in [m]} \in ([d_N] \times \{0,1\})^m$, we set $g : \mathcal{X} \to \mathbb{S}_k$ be equal to the output of the algorithm A with input $(s_{x_i}, f_{y_i}(x_i))_{i \in [m]} \in (S \times \mathbb{S}_k)^m$. We also set f be the output of the algorithm

A_{bin} with input $(x_i, y_i)_{i \in [m]}$ by setting $f(i) = 1$ if and only if $L_1(g(s_i)) = L_1(f_1(s_i))$. We will show that whenever A_{bin} errs, so does A. Fix $(x_i, y_i) \in S \times \{0, 1\}$. Assume that $A_{\text{bin}}(x_i) \neq y_i$ and say $y_i = 0$. Then $f(i) = 1$ and so $L_1(g(s_i)) = L_1(f_1(s_i)) \neq L_1(f_0(s_i))$. This implies that A errs. The case $y_i = 1$ is similar. □

Lower Bound for top-1 disagreement error for LSFs

Theorem D.33 (Top-1 Natarajan Dimension of LSFs)**.** *Consider the hypothesis class* $\mathcal{L}_{d,k} = \{\sigma_{\boldsymbol{W}} : \mathbb{R}^d \to \mathbb{S}_k : \sigma_{\boldsymbol{W}}(\boldsymbol{x}) = \text{argsort}(\boldsymbol{W}\boldsymbol{x}), \boldsymbol{W} \in \mathbb{R}^{k \times d}\}$*. Then,* $d_N^{(1)}(\mathcal{L}_{d,k}) = \Omega(dk)$.

Proof. Fix $k \in \mathbb{N}$. Let us consider the case $d = 2$ that will correspond as the building block for the general case $d > 2$. Let us first choose the set of points: Set P be the collection of pairs $P = \{(2i-1, 2i)\}_{i \in [b]}$ for any $i \in [b]$ with $b = \lfloor k/2 \rfloor$ and $S = \{\boldsymbol{x}_m\}_{m \in P}$ where these points correspond to $|P|$ equidistributed points on the unit sphere in $\mathbb{R}^2$. This set of points has size $|P| = \Theta(k)$ and we are going to N-shatter it using $\mathcal{L}_{2,k}$.

Consider the matrix $\boldsymbol{W} \in \mathbb{R}^{k \times 2}$ so that $\{\boldsymbol{W}_i\}_{i \in [k]}$ correspond to the rows of $\boldsymbol{W}$. The structure of the problem relies on the hyperplanes with normal vectors $(\boldsymbol{W}_i - \boldsymbol{W}_j)_{i \neq j}$ and our choice of $\boldsymbol{W}$ will rely on these hyperplanes. For any $m = (2i-1, 2i)$, we set $\boldsymbol{W}_{2i-1}, \boldsymbol{W}_{2i}$ on the unit sphere so that $\boldsymbol{W}_{2i-1} \cdot \boldsymbol{W}_{2i} = 1 - \phi$ with $\phi \in (0, 1)$ sufficiently small (set $\arccos(1-\phi) = 2\pi/(100k)$) and let C_m be the cone generated by these two vectors with axis I_m. We place $\boldsymbol{W}_{2i-1}$ so that the distance between $\boldsymbol{x}_m$ and the hyperplane I_m is sufficiently small (say that the angle between $\boldsymbol{x}_m$ and I_m is $\arccos(1-\phi)/100$). Note that the normal vector of I_m is $\boldsymbol{W}_{2i-1} - \boldsymbol{W}_{2i}$ and we place $\boldsymbol{x}_m$ so that it has positive correlation with this vector. This uniquely identifies the location of $\boldsymbol{W}_{2i}$. Crucially, each vector $\boldsymbol{x}_m$ has the following properties: (i) $\boldsymbol{x}_m$ is very close to the boundary of the hyperplane with normal vector $(\boldsymbol{W}_{2i-1} - \boldsymbol{W}_{2i})$, (ii) $\boldsymbol{W}_{2i-1} \cdot \boldsymbol{x}_m > \boldsymbol{W}_{2i} \cdot \boldsymbol{x} > \boldsymbol{W}_j \cdot \boldsymbol{x}_m$ for any $j \notin m$ and (iii) $\boldsymbol{x}_m$ is far from any boundary induced by hyperplanes with normal vectors $\boldsymbol{W}_j - \boldsymbol{W}_{j'}$ for any $(j, j') \neq m$.

Since the points are well-separated on the unit sphere, for any $m = (2i-1, 2i) \in P$, we have $\boldsymbol{W}_{2i-1} \cdot \boldsymbol{W}_{2i} = 1 - \phi \approx 1$ and for any other pair of indices $(i, j) \notin P$, there exists $c = c(k) \in (0, 1)$, $|\langle \boldsymbol{W}_i, \boldsymbol{W}_j \rangle| \leq c$.

For any $m = (2i-1, 2i) \in P$, we set $\boldsymbol{W}'_{2i-1} - \boldsymbol{W}'_{2i} = \boldsymbol{R}_\theta(\boldsymbol{W}_{2i-1} - \boldsymbol{W}_{2i})$ for some θ to be chosen, where $\boldsymbol{R}_\theta$ is the 2×2 rotation matrix. We choose θ so that each point $\boldsymbol{x}_m$ for $m = (2i-1, 2i) \in P$ with $(\boldsymbol{W}_{2i-1} - \boldsymbol{W}_{2i}) \cdot \boldsymbol{x}_m > 0$ satisfies $(\boldsymbol{W}'_{2i-1} - \boldsymbol{W}'_{2i}) \cdot \boldsymbol{x}_m < 0$. The main idea is that since $\boldsymbol{x}_m$ has the properties (i)-(iii) described above, the rankings induced by the vectors $\boldsymbol{W}\boldsymbol{x}_m$ and $\boldsymbol{W}'\boldsymbol{x}_m$ will be different in the first two positions but the same in the rest.

Given the training set $\{\boldsymbol{x}_m\}_{m \in P}$, we have to construct f_0, f_1 and verify that they satisfy the top-1 Ranking Natarajan conditions. For $m = (2i-1, 2i)$, we have that $f_0(\boldsymbol{x}_m) = (2i-1, 2i, \pi)$ and $f_1(\boldsymbol{x}_m) = (2i, 2i-1, \pi)$ for some ranking π of size $k-2$ that depends on m. Specifically, we will set $f_0(\boldsymbol{x}) = \sigma(\boldsymbol{W}\boldsymbol{x})$ and $f_1(\boldsymbol{x}) = \sigma(\boldsymbol{W}'\boldsymbol{x})$, where σ gives the decreasing ordering of the elements of the input vector. By the choice of the set S and $\boldsymbol{W}, \boldsymbol{W}'$, it remains to show that the $k-2$ last elements of the rankings $f_0(\boldsymbol{x}_m)$ (say π_0) and of $f_1(\boldsymbol{x}_m)$ (say π_1) are in the same order, i.e., $L_{3..k}(f_0(\boldsymbol{x}_m)) = L_{3..k}(f_1(\boldsymbol{x}_m))$. Assume that $u \succ v$ in π_0. It suffices to show that $(\boldsymbol{W}'_u - \boldsymbol{W}'_v) \cdot \boldsymbol{x}_m \geq 0$, i.e., the order of u and v is preserved when transforming $\boldsymbol{W}$ to $\boldsymbol{W}'$. We have that $(\boldsymbol{W}_u - \boldsymbol{W}_v) \cdot \boldsymbol{x}_m > c_1$ for some constant $c_1 > 0$ (c_1 is the minimum over $(u, v) \neq m = (2i-1, 2i)$). Hence, we can pick θ small enough so that $(\boldsymbol{W}'_u - \boldsymbol{W}'_v) \cdot \boldsymbol{x}_m > c_2$ and this can be done for any pair u, v that does not correspond to m. This implies that $\pi_0 = \pi_1 = \pi$. In particular, we have that

$$(\boldsymbol{W}'_u - \boldsymbol{W}'_v) \cdot \boldsymbol{x}_m = \cos(\theta) \cdot (\boldsymbol{W}_u - \boldsymbol{W}_v) \cdot \boldsymbol{x}_m + \sin(\theta) \cdot (W_{uv}^{(1)} x_m^{(2)} - W_{uv}^{(2)} x_m^{(1)}) > c_2 > 0$$

for some θ sufficiently small, where $W_{uv}^{(t)}$ is the t-th entry of the vector $\boldsymbol{W}_u - \boldsymbol{W}_v$ for $t \in \{1, 2\}$ and $\boldsymbol{x}_m, \boldsymbol{W}_u, \boldsymbol{W}_v$ are unit vectors.

For any subset T of S, it remains to choose a linear classifier in $\mathcal{L}_{2,k}$ (which is allowed to depend on T). For any $T \subseteq S = \{\boldsymbol{x}_m\}_{m \in P}$, we consider the matrix $\overline{\boldsymbol{W}} \in \mathbb{R}^{k \times 2}$ so that for the i-th row $\overline{W}_i = \boldsymbol{W}_i 1\{i \in m \in T\} + \boldsymbol{W}'_i 1\{i \in m \in S \setminus T\}$ for any $i \in [k]$. This is valid since the pairs $m \in P$ partition $[k]$. We have to show the following two properties: (i) $\sigma(\overline{\boldsymbol{W}}\boldsymbol{x}) = f_0(\boldsymbol{x})$ for $x \in T$ and (ii) $\sigma(\overline{\boldsymbol{W}}\boldsymbol{x}) = f_1(\boldsymbol{x})$ for $x \in S \setminus T$.

Assume that $m = (2i-1, 2i)$ and $\boldsymbol{x}_m \in T$. We have that $f_0(\boldsymbol{x}_m) = (2i-1, 2i, \pi)$ and $\overline{\boldsymbol{W}}_{2i-1} - \overline{\boldsymbol{W}}_{2i} = \boldsymbol{W}_{2i-1} - \boldsymbol{W}_{2i}$ and so $2i-1 \succ 2i$ in the ranking $\sigma(\overline{\boldsymbol{W}}\boldsymbol{x}_m)$. It remains to show that the remaining $\binom{k}{2} - 1$ pairwise comparisons are the same in the two rankings. Let us consider a pair of points $u \neq v$ so that $u \succ v$ in $f_0(\boldsymbol{x}_m)$. It suffices to show that $u \succ v$ in $\sigma(\overline{\boldsymbol{W}}\boldsymbol{x}_m)$.

1. If u, v are so that $\overline{\boldsymbol{W}}_u - \overline{\boldsymbol{W}}_v = \boldsymbol{W}_u - \boldsymbol{W}_v$, the result holds.
2. If u, v are so that $\overline{\boldsymbol{W}}_u - \overline{\boldsymbol{W}}_v = \boldsymbol{W}_u - \boldsymbol{W}'_v$: In this case, u and v lie in a different pair of P and this implies that the correct direction is preserved if θ is appropriately chosen. For θ as above, it holds that $(\boldsymbol{W}_u - \boldsymbol{R}_\theta \boldsymbol{W}_v) \cdot \boldsymbol{x}_m$ has the same sign as $(\boldsymbol{W}_u - \boldsymbol{W}_v) \cdot \boldsymbol{x}_m$. In particular,

   $$\begin{aligned} &\boldsymbol{W}_u \cdot \boldsymbol{x}_m - \boldsymbol{R}_\theta \boldsymbol{W}_v \cdot \boldsymbol{x}_m \\ &= \boldsymbol{W}_u \cdot \boldsymbol{x}_m - (\cos(\theta) W_v^{(1)} - \sin(\theta) W_v^{(2)}) x_m^{(1)} - (\sin(\theta) W_v^{(1)} + \cos(\theta) W_v^{(2)}) x_m^{(2)}\,, \end{aligned}$$

   and so

   $$(\boldsymbol{W}_u - \boldsymbol{W}'_v) \cdot \boldsymbol{x}_m = \cos(\theta) \cdot (\boldsymbol{W}_u - \boldsymbol{W}_v) \cdot \boldsymbol{x}_m + \sin(\theta)(W_v^{(2)} x_m^{(1)} - W_v^{(1)} x_m^{(2)}) > 0\,.$$

3. If u, v are so that $\overline{\boldsymbol{W}}_u - \overline{\boldsymbol{W}}_v = \boldsymbol{W}'_u - \boldsymbol{W}'_v$, the analysis for the inner product with $\boldsymbol{x}_m$ will be similar.

We now have to extend this proof for $d > 2$. We will "tensorize" the above construction as follows. Let $S = \{\boldsymbol{y}_{mj}\}_{m \in [b], j \in [d/2]}$ with $|S| = \lfloor k/2 \rfloor \cdot \lfloor d/2 \rfloor$. We first define the points of S: For $s \in [d]$, set $y_{mj}[s] = x_m[1]1\{s = 2j-1\} + x_m[2]1\{s = 2j\}$ with $\boldsymbol{y}_{mj} \in \mathbb{R}^d$, i.e., $\boldsymbol{y}_{mj}$ has the values of $\boldsymbol{x}_m$ at the consecutive entries indicated by $m = (2i-1, 2i) \in P$ and zeros at the other positions.

We have to show that the set S is N-shattered. Given $T \subseteq S$, we are going to create the matrix $\overline{\boldsymbol{W}} \in \mathbb{R}^{k \times d}$. For illustration, think of each row of the matrix as having $d/2$ blocks of size two. If $\boldsymbol{y}_{mj} \in T$ with $m = (2i-1, 2i)$, set the two associated rows (indicated by m) of $\overline{\boldsymbol{W}}$ with $\boldsymbol{W}_{2i-1}, \boldsymbol{W}_{2i}$ at the j-th block and with $\boldsymbol{W}'_{2i-1}, \boldsymbol{W}'_{2i}$ otherwise. We will have that $\sigma(\overline{\boldsymbol{W}}\boldsymbol{y}) = f_0(\boldsymbol{y})$ if $y \in T$ and $\sigma(\overline{\boldsymbol{W}}\boldsymbol{y}) = f_1(\boldsymbol{y})$ otherwise and the analysis is the same as the $d = 2$ case. □

D.5 Examples of Noisy Ranking Distributions

Definition D.34 (Mallows model Mallows (1957)). *Consider k alternatives and let $\pi \in \mathbb{S}_k, \phi \in [0,1]$. The Mallows distribution $\mathcal{M}_{\mathrm{Mal}}(\pi, \phi)$ with central ranking π and spread parameter ϕ is a probability measure over $\mathbb{S}_k$ with density $\mathbf{Pr}_{\sigma \sim \mathcal{M}_{\mathrm{Mal}}(\pi,\phi)}[\sigma]$ that is proportional to $\phi^{d(\sigma,\pi)}$, where d is a ranking distance.*

We focus on Mallows models accociated with the Kendall's Tau distance $d = d_{KT}$ (the standard distance, not the normalized one), which measures the number of discordant pairs.

Fact D.35. *When $\phi < 1$, the Mallows model $\mathcal{M}_{\mathrm{Mal}}(\pi, \phi)$ is a ranking distribution with bounded noise at most $\frac{1+\phi}{4} < 1/2$.*

Proof. The following property holds Mallows (1957)

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\sigma \sim \mathcal{M}_{\mathrm{Mal}}(\pi,\phi)}[\sigma(i) < \sigma(j) | \pi(i) < \pi(j)] &= \frac{\pi(j) - \pi(i) + 1}{1 - \phi^{\pi(j)-\pi(i)+1}} - \frac{\pi(j) - \pi(i)}{1 - \phi^{\pi(j)-\pi(i)}} \\ &\geq \frac{1}{2} + \frac{1-\phi}{4}\,.\end{aligned}$$

□

The Bradley-Terry-Luce model Bradley and Terry (1952); Luce (2012) is the most studied pairwise comparisons model. In his seminal paper, Mallows Mallows (1957) also studied the following natural ranking distribution:

Definition D.36 (Bradley-Terry-Mallows Mallows (1957)). *Consider a score vector $\boldsymbol{w} \in \mathbb{R}_+^k$ with k distinct entries and let π be the ranking induced by the values of $\boldsymbol{w}$ in decreasing order. The Bradley-Terry-Mallows distribution $\mathcal{M}_{\mathrm{BTM}}(\boldsymbol{w})$ with central ranking*

π is a probability measure over $\mathbb{S}_k$ with density $\mathbf{Pr}_{\sigma \sim \mathcal{M}_{\mathrm{BTM}}(\boldsymbol{w})}[\sigma]$ that is proportional to $\prod_{i \succ_\sigma j} \frac{w_i}{w_i + w_j}$.

Lemma D.37. *There exists a real number $0 < \eta < 1/2$ so that the Bradley-Terry-Mallows distribution $\mathcal{M}_{\mathrm{BTM}}(\boldsymbol{w})$ is a ranking distribution with bounded noise at most η.*

Proof. In the standard Bradley-Terry-Luce model, the pairwise comparison between the alternatives i, j is a Bernoulli random variable with $\mathbf{Pr}[i \succ j] = w_i/(w_i + w_j)$. The Bradley-Terry-Mallows distribution can be considered as the Bradley-Terry-Luce model conditioned on the event that all the pairwise comparisons are consistent to a ranking. Hence, we have that

$$\mathop{\mathbf{Pr}}_{\sigma \sim \mathcal{M}_{\mathrm{BTM}}(\boldsymbol{w})}[\sigma] \qquad = \frac{1}{Z(k, \boldsymbol{w})} \prod_{i \succ_\sigma j} \frac{w_i}{w_i + w_j} \,.$$

Let us set $\mathcal{A}_{i \succ j} = \{\sigma \in \mathbb{S}_k : \sigma(i) < \sigma(j)\}$. We are interested in the following probability

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\sigma \sim \mathcal{M}_{\mathrm{BTM}}(\boldsymbol{w})}[i \succ_\sigma j | w_i > w_j] &= \mathop{\mathbf{Pr}}_{\sigma \sim \mathcal{M}_{\mathrm{BTM}}(\boldsymbol{w})}[\sigma(i) < \sigma(j) | w_i > w_j] \\ &= \frac{1}{Z(k, \boldsymbol{w})} \sum_{\sigma \in \mathcal{A}_{i \succ j}} \prod_{p \succ_\sigma q} \frac{w_p}{w_p + w_q} \,.\end{aligned}$$

Note that in order to show the desired property, it suffices to show that

$$\sum_{\sigma \in \mathcal{A}_{i \succ j}} \prod_{p \succ_\sigma q} \frac{w_p}{w_p + w_q} > \sum_{\sigma \in \mathcal{A}_{i \prec j}} \prod_{p \succ_\sigma q} \frac{w_p}{w_p + w_q} \,.$$

First, observe that there exists a correspondence mapping $\sigma \in \mathcal{A}_{i \succ j}$ to $\mathcal{A}_{i \prec j}$, where one flips the elements i and j. Hence, it suffices to show that the mass of the ranking $(u_a)i(u_b)j(u_c)$ is larger than the one of the ranking $(u_a)j(u_b)i(u_c)$, where u_a, u_b, u_c are permutations of length between 0 and $k-2$ with elements in $[k] \setminus \{i, j\}$. For the two above rankings, the only terms of the product that are not identical are the following

$$\frac{w_i}{w_i + w_j} \prod_{x \in u_b} \frac{w_i}{w_i + w_x} \frac{w_x}{w_x + w_j} > \frac{w_j}{w_i + w_j} \prod_{x \in u_b} \frac{w_j}{w_j + w_x} \frac{w_x}{w_x + w_i},$$

since $w_i > w_j$ and so the result follows. □

E Appendix to Chapter 6

E.1 Additional Preliminaries and Notation

We first state the following simple lemma that connects the total variation distance of two Normal distributions with their parameter distance. For a proof see e.g. Corollaries 2.13 and 2.14 of Diakonikolas et al. (2016b).

Lemma E.1. *Let* $N_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{r}\Sigma_1)$, $N_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{r}\Sigma_2)$ *be two Normal distributions. Then*

$$d_{\mathrm{TV}}(N_1, N_2) \leq \frac{1}{2} \left\| \boldsymbol{r}\Sigma_1^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \right\|_2 + \sqrt{2} \left\| \boldsymbol{r}I - \boldsymbol{r}\Sigma_1^{-1/2} \boldsymbol{r}\Sigma_2 \boldsymbol{r}\Sigma_1^{-1/2} \right\|_F$$

We readily use the following two lemmas from Daskalakis et al. (2018). The first suggests that we can accurately estimate the parameters (μ_S, Σ_S).

Lemma E.2. *Let* $(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S)$ *be the mean and covariance of the truncated Gaussian* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ *for a set* S *such that* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) \geq \alpha$. *Using* $\tilde{O}(\frac{d}{\epsilon^2} \log(1/\alpha) \log^2(1/\delta))$ *samples, we can compute estimates* $\widetilde{\boldsymbol{\mu}}_S$ *and* $\widetilde{\boldsymbol{\Sigma}}_S$ *such that, with probability at least* $1 - \delta$,

$$\|\boldsymbol{\Sigma}^{-1/2}(\widetilde{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_S)\|_2 \leq \epsilon \quad \textit{and} \quad (1 - \epsilon)\boldsymbol{\Sigma}_S \preceq \widetilde{\boldsymbol{\Sigma}}_S \preceq (1 + \epsilon)\boldsymbol{\Sigma}_S$$

The second lemma suggests that the empirical estimates are close to the true parameters of underlying truncated Gaussian.

Lemma E.3. *The empirical mean and covariance* $\widetilde{\boldsymbol{\mu}}_S$ *and* $\widetilde{\boldsymbol{\Sigma}}_S$ *computed using* $\tilde{O}(d^2 \log^2(1/\alpha\delta))$ *samples from a truncated Normal* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ *with* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) \geq \alpha$ *satisfies with probability* $1 - \delta$ *that:*

$$\|\boldsymbol{\Sigma}^{-1/2}(\widetilde{\boldsymbol{\mu}}_S - \boldsymbol{\mu})\|_2^2 \leq O(\log \frac{1}{\alpha}), \quad \widetilde{\boldsymbol{\Sigma}}_S \succeq \Omega(\alpha^2)\boldsymbol{\Sigma}, \quad \left\| \boldsymbol{\Sigma}^{-1/2} \widetilde{\boldsymbol{\Sigma}}_S \boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I} \right\|_F^2 \leq O(\log \frac{1}{\alpha}).$$

Moreover, $\Omega(\alpha^2) \leq \left\| \widetilde{\boldsymbol{\Sigma}}_S^{-1/2} \boldsymbol{r}\Sigma \widetilde{\boldsymbol{\Sigma}}_S^{-1/2} \right\|_2 \leq O(1/\alpha^2)$.

In particular, the mean and covariance $\widetilde{\boldsymbol{\mu}}_S$ and $\widetilde{\boldsymbol{\Sigma}}_S$ that satisfy the conditions of Lemma E.3, are in $(O(\log(1/\alpha)), 1 - O(\alpha^2))$-near isotropic position.

We will use the following very useful anti-concentration result about the Gaussian mass of sets defined by polynomials.

Theorem E.4 (Theorem 8 of Carbery and Wright (2001))**.** *Let $q, \gamma \in \mathbb{R}_+$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{r}\Sigma \in \mathbb{R}^{d\times d}$ such that $\boldsymbol{\Sigma}$ is symmetric positive semidefinite and $p : \mathbb{R}^d \to \mathbb{R}$ be a multivariate polynomial of degree at most ℓ, we define*

$$\bar{Q} = \left\{ \boldsymbol{x} \in \mathbb{R}^d \mid |p(\boldsymbol{x})| \leq \gamma \right\},$$

then there exists an absolute constant C such that

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma; \bar{Q}) \leq \frac{Cq\gamma^{1/\ell}}{\left(\mathbf{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{r}\Sigma)}\left[|p(\boldsymbol{z})|^{q/\ell}\right]\right)^{1/q}}.$$

Hermite Polynomials, Ornstein-Uhlenbeck Operator, and Gaussian Surface Area.

We denote by $L^2(\mathbb{R}^d, \mathcal{N}_0)$ the vector space of all functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[f^2(x)] < \infty$. The usual inner product for this space is $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[f(\boldsymbol{x})g(\boldsymbol{x})]$. While, usually one considers the probabilists's or physicists' Hermite polynomials, in this work we define the *normalized* Hermite polynomial of degree i to be $H_0(x) = 1, H_1(x) = x, H_2(x) = \frac{x^2-1}{\sqrt{2}}, \ldots, H_i(x) = \frac{He_i(x)}{\sqrt{i!}}, \ldots$ where by $He_i(x)$ we denote the probabilists' Hermite polynomial of degree i. These normalized Hermite polynomials form a complete orthonormal basis for the single dimensional version of the inner product space defined above. To get an orthonormal basis for $L^2(\mathbb{R}^d, \mathcal{N}_0)$, we use a multi-index $V \in \mathbb{N}^d$ to define the d-variate normalized Hermite polynomial as $H_V(\boldsymbol{x}) = \prod_{i=1}^d H_{v_i}(x_i)$. The total degree of H_V is $|V| = \sum v_i \in V v_i$. Given a function $f \in L^2$ we compute its Hermite coefficients as $\hat{f}(V) = \mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[f(\boldsymbol{x})H_V(\boldsymbol{x})]$ and express it uniquely as $\sum_{V\in\mathbb{N}^d} \hat{f}(V)H_V(\boldsymbol{x})$. We denote by $S_k f(x)$ the degree k partial sum of the Hermite expansion of f,

$S_k f(\boldsymbol{x}) = \sum_{|V| \le k} \hat{f}(V) H_V(\boldsymbol{x})$. Then, since the basis of Hermite polynomials is complete, we have $\lim_{k \to \infty} \mathbf{E}_{x \sim \mathcal{N}_0}[(f(\boldsymbol{x}) - S_k f(\boldsymbol{x}))^2] = 0$. We would like to quantify the convergence rate of $S_k f$ to f. Parseval's identity states that

$$\mathop{\mathbf{E}}_{x \sim \mathcal{N}_0}[(f(\boldsymbol{x}) - S_k f(\boldsymbol{x}))^2] = \sum_{|V|=k}^{\infty} \hat{f}(V)^2.$$

Definition E.5 (Hermite Concentration). *Let $\gamma(\epsilon, d)$ be a function $\gamma : (0, 1/2) \times \mathbb{N} \mapsto \mathbb{N}$. We say that a class of functions $\mathcal{F}$ over $\mathbb{R}^d$ has a Hermite concentration bound of $\gamma(\epsilon, d)$, if for all $d \ge 1$, all $\epsilon \in (0, 1/2)$, and $f \in \mathcal{F}$ it holds $\sum_{|V| \ge \gamma(\epsilon,d)} \hat{f}(V)^2 \le \epsilon$.*

We now define the Gaussian Noise Operator as in O'Donnell (2014). Using a different parametrization, which is not convenient for our purposes, these operators are also known as the Ornstein-Uhlenbeck semigroup, or the Mehler transform.

Definition E.6. *The Gaussian Noise operator T_ρ is the linear operator defined on the space of functions $L^1(\mathbb{R}^d, \mathcal{N}_0)$ by*

$$T_\rho f(\boldsymbol{x}) = \mathop{\mathbf{E}}_{\boldsymbol{y} \sim \mathcal{N}_0}\left[f(\rho \boldsymbol{x} + \sqrt{1 - \rho^2} \boldsymbol{y}) \right].$$

A nice property of operator $T_{1-\rho}$ that we will use is that it has a simple Hermite expansion

$$S_k(T_\rho f)(\boldsymbol{x}) = \sum_{V : |V| \le k} \rho^{|V|} \widehat{f}(V) H_V(\boldsymbol{x}) \tag{E.1}$$

We also define the noise sensitivity of a function f.

Definition E.7 (Noise Sensitivity). *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a function in $L^2(\mathbb{R}^d, \mathcal{N}_0)$. The noise sensitivity of f at $\rho \in [0, 1]$ is defined to be*

$$\mathbf{NS}_\rho[f] = 2 \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_0}[f(\boldsymbol{x})^2 - f(\boldsymbol{x}) T_{1-\rho} f(\boldsymbol{x})]$$

Since, the vectors $\boldsymbol{x}$ and $\boldsymbol{z} = (1 - \rho)\boldsymbol{x} + \sqrt{1 - \rho^2} \boldsymbol{y}$ are jointly distributed according to

$$D_\rho = \mathcal{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{I} & (1-\rho)\boldsymbol{I}\\ (1-\rho)\boldsymbol{I} & \boldsymbol{I}\end{pmatrix}\right). \tag{E.2}$$

we can write

$$\mathbf{NS}_\rho[f] = \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}\left[f(\boldsymbol{x})^2\right] + \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}\left[f(\boldsymbol{z})^2 - 2f(\boldsymbol{x})f(\boldsymbol{z})\right] = \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}[(f(\boldsymbol{x}) - f(\boldsymbol{z}))^2]. \tag{E.3}$$

When f is an indicator of a set, the noise sensitivity is

$$\mathbf{NS}_\rho[\mathbb{1}_S] = 2\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})}[\mathbb{1}_S(\boldsymbol{x})(1-\mathbb{1}_S(\boldsymbol{z}))] = 2\mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{z})}[\mathbb{1}_S(\boldsymbol{x})\mathbb{1}_{S^c}(\boldsymbol{z})], \tag{E.4}$$

which is equal to the probability of the correlated points $\boldsymbol{x}, \boldsymbol{z}$ landing at "opposite" sides of S.

Ledoux Ledoux (1994) and Pisier Pisier (1986) showed that the noise sensitivity of a set can be bounded by its Gaussian surface area.

Definition E.8 (Gaussian Surface Area)**.** *For a Borel set $A \subseteq \mathbb{R}^d$, its Gaussian surface area is* $\Gamma(A) = \liminf_{\delta\to 0} \frac{\mathcal{N}_0(A_\delta \setminus A)}{\delta}$*, where* $A_\delta = \{x : \boldsymbol{h}rmdist(x, A) \le \delta\}$*.*

We will use the following lemma given in Klivans et al. (2008).

Lemma E.9 (Corollary 14 of Klivans et al. (2008))**.** *For a Borel set $S \subseteq \mathbb{R}^d$ and $\rho \ge 0$,* $\mathbf{NS}_\rho[\mathbb{1}_S(\boldsymbol{x})] \le \sqrt{\pi}\sqrt{\rho}\,\Gamma(S)$.

For more details on the Gaussian space and Hermite Analysis (especially from the theoretical computer science perspective), we refer the reader to O'Donnell (2014). Most of the facts about Hermite polynomials that we shall use in this work are well known properties and can be found, for example, in Szegö (1967).

E.2 Missing proofs of Section 6.2

We will use a standard tournament based approach for selecting a good hypotheses. We will use a version of the tournament from Daskalakis and Kamath (2014). See also Devroye and Lugosi (2012).

Lemma E.10 (Tournament Daskalakis and Kamath (2014))**.** *There is an algorithm, which is given sample access to some distribution X and a collection of distributions $\mathcal{H} = \{H_1, \dots, H_N\}$ over some set, access to a PDF comparator for every pair of distributions H_i, $H_j \in \mathcal{H}$, an accuracy parameter $\epsilon > 0$, and a confidence parameter $\delta > 0$. The algorithm makes $O(\log(1/\delta)\epsilon^2) \log N)$ draws from each of $X, H_1, \dots, H_N$ and returns some $H \in \mathcal{H}$ or declares "failure" If there is some $H \in \mathcal{H}$ such that $d_{\mathrm{TV}}(H, X) \leq \epsilon$ then with probability at least $1 - \delta$ the returned distribution H satisfies $d_{\mathrm{TV}}(H, X) \leq 512\epsilon$. The total number of operations of the algorithm is $O(\log(1/\delta)(1/\epsilon^2)(N \log N + \log 1/\delta))$.*

We first argue that if the class of sets $\mathcal{S}$ has VC-dimension $\mathrm{VC}(\mathcal{S})$ then we can learn the truncated model in ϵ total variation by drawing roughly $\mathrm{VC}(\mathcal{S})/\epsilon$ samples. We will use the following standard fact whose proof may be found for example in page 398 of Shalev-Shwartz and Ben-David (2014c). For convenience we restate the result using our notation.

Lemma E.11 (Shalev-Shwartz and Ben-David (2014c))**.** *Let D be a distribution on $\mathbb{R}^d$. Let $\mathcal{S}$ be a family of subsets of $\mathbb{R}^d$. Fix $\epsilon \in (0, 1), \delta \in (0, 1/4)$ and let $N = O(\mathrm{VC}(\mathcal{S}) \log(1/\epsilon)/\epsilon + \log(1/\delta))$ Then, with probability at least $1 - \delta$ over a choice of a sample $X \sim D^N$ we have that if $D(S) \geq \epsilon$ then $|S \cap X| \neq \emptyset$.*

The proof of Lemma 6.8 We define the class of sets $\mathcal{A} = \{S^* \setminus S : S \in \mathcal{S}\}$. We first argue that for any $A \subset \mathbb{R}^d$ we have $\mathrm{VC}(\mathcal{A}) \leq \mathrm{VC}(\mathcal{S})$. Let $X \subset \mathbb{R}^d$ be a set of points. The set of different labelings of X using sets of $\mathcal{S}$ resp. $\mathcal{A}$ is $L_{\mathcal{S}} = \{X \cap S : S \in \mathcal{S}\}$ resp. $L_{\mathcal{A}} = \{X \cap S : S \in \mathcal{A}\} = \{X \cap (A \setminus S) : S \in \mathcal{S}\}$. We define the function $g : L_{\mathcal{A}} \to L_{\mathcal{S}}$ by $g(X \cap (A \setminus S)) = X \cap S$. We that observe for $S_1, S_2 \in \mathcal{S}$ we have that $X \cap S_1 = X \cap S_2$ implies that $X \cap (A \setminus S_1) = X \cap (A \setminus S_2)$. Therefore, g is one-to-one and we obtain that $|L_{\mathcal{A}}| \leq |L_{\mathcal{S}}|$. We draw N samples $X = \{x_i, i \in N\}$. Applying Lemma E.11 for the family $\mathcal{A}$, we have that with N samples, with probability at least $1 - \delta$ it holds that if $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma; S^* \setminus S) \geq \epsilon$ for some set $S \in \mathcal{S}$ then $|(S^* \setminus S) \cap X| > 0$. Therefore, every set that is consistent with the samples, i.e. every S that that contains the samples, satisfies the property

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma; S^* \setminus S) \le \epsilon$. Moreover, since $d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma)) \le \epsilon$ we obtain that $\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, S^* \setminus S) \le 2\epsilon$ for any set S consistent with the data.

Next, we use the fact that $\widetilde{S}$ is chosen so that $\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, S^*) \ge \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$. This means that for all $x \in S^* \cap \widetilde{S}$ it holds $\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, S^*; \boldsymbol{x}) \le \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S}; \boldsymbol{x})$. To simplify notation we set $\widetilde{\mathcal{N}_{\widetilde{S}}} = \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$, $\widetilde{\mathcal{N}_{S^*}} = \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, S^*)$, and $\mathcal{N}_{S^*} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma, S^*)$. We have

$$\begin{aligned} 2d_{\mathrm{TV}}(\widetilde{\mathcal{N}_{\widetilde{S}}}, \widetilde{\mathcal{N}_{S^*}}) &= \int_{\widetilde{\mathcal{N}_{S^*}}(\boldsymbol{x}) \ge \widetilde{\mathcal{N}_{\widetilde{S}(\boldsymbol{x})}}} \left(\widetilde{\mathcal{N}_{S^*}}(\boldsymbol{x}) - \widetilde{\mathcal{N}_{\widetilde{S}}}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \\ &\le \int_{S^* \setminus \widetilde{S}} \widetilde{\mathcal{N}_{S^*}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\le \frac{\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}; S^* \setminus \widetilde{S})}{\alpha} \le \frac{\epsilon}{\alpha}. \end{aligned}$$

Moreover,

$$d_{\mathrm{TV}}(\widetilde{\mathcal{N}_{S^*}}, \mathcal{N}_{S^*}) \le \frac{d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma))}{\alpha} \le \frac{\epsilon}{\alpha}$$

Using the triangle inequality we obtain that $d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S}), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma, S^*)) \le 3\epsilon/(2\alpha)$.

The proof of Lemma 6.9 Using Lemma E.3 we know that we can draw $\widetilde{O}(d^2 \log^2(1/\alpha\delta))$ samples and obtain estimates of the conditional mean and covariance $\widetilde{\boldsymbol{\mu}}_C$, $\widetilde{\boldsymbol{r}\Sigma}_C$. Transforming the space so that $\widetilde{\boldsymbol{\mu}}_C = 0$ and $\widetilde{\boldsymbol{r}\Sigma}_C = \boldsymbol{r}I$. For simplicity, we will keep denoting the parameters of the unknown Gaussian $\boldsymbol{\mu}, \boldsymbol{r}\Sigma$ after transforming the space. From Lemma E.3 we have that $\left\|\boldsymbol{r}\Sigma^{-1/2}\boldsymbol{\mu}\right\|_2 \le O(\log(1/\alpha)^{1/2}/\alpha)$, $\Omega(\alpha^2) \le \left\|\Sigma^{-1/2}\right\|_2 \le O(1/\alpha^2)$ and $\|I - \Sigma\|_F \le O(\log(1/\alpha)/\alpha^2)$. Therefore, the cube of $\mathbb{R}^{d+d^2}$ where all the parameters μ_i, Σ_{ij} of the mean and the covariance lie has side length at most $O(1/\mathrm{poly}(a))$. We can partition this cube into smaller cubes of side length $O(\epsilon\mathrm{poly}(a)/d)$ and obtain that there exists a point of the grid $(\boldsymbol{u}, \boldsymbol{r}B)$ such that $\left\|\Sigma^{-1/2}(\boldsymbol{u} - \boldsymbol{\mu})\right\|_2 \le \epsilon$, $\left\|\boldsymbol{r}I - \boldsymbol{r}\Sigma^{-1/2}\boldsymbol{r}B\boldsymbol{r}\Sigma^{-1/2}\right\|_F \le \epsilon$, which implies that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{u}, \boldsymbol{r}B), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma)) \le \epsilon$. Assume now that for each guess $(\boldsymbol{u}, \boldsymbol{r}B)$ of our grid we solve the optimization problem as defined in Lemma 6.8 and find a candidate set $S_{\boldsymbol{u}, \boldsymbol{r}B}$. Notice that the set of our hypotheses $\boldsymbol{u}, \boldsymbol{r}B, S_{\boldsymbol{u}, \boldsymbol{r}B}$ is $O((d^2/\epsilon)^{d^2+d})$. Moreover, using Lemma 6.8 and the fact that there exists a point $\boldsymbol{u}, \boldsymbol{r}B)$ in the grid so that

$d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{u}, r B), \mathcal{N}(\boldsymbol{\mu}, r\Sigma)) \leq \epsilon$, we obtain that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{u}, r B, S_{\boldsymbol{u}, r B}), \mathcal{N}(\boldsymbol{\mu}, r\Sigma, S)) \leq \epsilon$. Now we can use Lemma E.10 we can select a hypotheses $\mathcal{N}(\boldsymbol{u}, r B, \widetilde{S})$ within $O(\epsilon)$ total variation distance of $\mathcal{N}(\boldsymbol{\mu}, r\Sigma, S)$, and the number samples required to run the tournament is as claimed.

E.3 Missing Proofs of Section 6.3

To prove Theorem 6.12 we shall use the inequalities of Lemma E.12.

Lemma E.12. *Let $k \in \mathbb{N}$. Then for all $0 < x < \frac{2k+1}{2k}$ it holds,*

$$-k \log x - \frac{1}{2} \log(1 - 2k(x-1) \leq 2k^2 (x-1)^2 \left(\frac{1}{x} + \frac{1}{1 - 2k(x-1)} \right)$$

Moreover, for all $x > \frac{2k-1}{2k}$

$$k \log x - \frac{1}{2} \log(1 - 2k(1-x)) \leq k^2 (1-x)^2 \left(1 + \frac{1}{1 - 2k(1-x)} \right).$$

Proof. We start with the first inequality. Let $f(x) = -k \log x - \frac{1}{2} \log(1 - 2k(x-1)$. We first assume that $1 \leq x \frac{2k+1}{2k}$. We have

$$\begin{aligned} f(x) &= \int_1^x \left(\frac{k}{1 - 2k(t-1)} - \frac{k}{t} \right) \mathrm{d}t \\ &= k(1+2k) \int_1^x \frac{t-1}{t(1 - 2k(t-1))} \mathrm{d}t \\ &\leq \frac{k(1+2k)}{1 - 2k(x-1)} \int_1^x (t-1) \mathrm{d}t \\ &\leq 2k^2 \frac{(x-1)^2}{1 - 2k(x-1)} \end{aligned}$$

If $0 < x \leq 1$ we have

$$f(x) \leq \frac{k(1+2k)}{x} \int_1^x (t-1) \mathrm{d}t \leq 2k^2 \frac{(x-1)^2}{x}$$

Adding these two bounds gives an upper bound for all $0 < x < \frac{2k+1}{2k}$. Similarly, we now show the second inequality. Let $g(x) = k \log x - \frac{1}{2}\log(1-2k(1-x))$. We first assume that $1 \leq x$ and write

$$\begin{aligned} g(x) &= \int_1^x \left(\frac{k}{t} - \frac{k}{1-2k(1-t)}\right) \mathrm{d}t \\ &= k\int_1^x \frac{(t-1)(2k-1)}{t(1+2k(t-1))}\mathrm{d}t \\ &\leq k(2k-1)\int_1^x \frac{t-1}{\mathrm{d}}t \\ &\leq k^2(x-1)^2. \end{aligned}$$

Similarly, if $\frac{2k-1}{2k} < x \leq 1$ we have

$$g(x) \leq k^2 \frac{(1-x)^2}{1-2k(1-x)}.$$

We add the two bounds together to get the desired upper bound.

□

The proof of Lemma 6.13 For simplicity we denote $\mathcal{N}_i = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_i)$. We start by proving the upper bound. Using Schwarz's inequality we write

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}\left[\left(\frac{\mathcal{N}_1(\boldsymbol{x})}{\mathcal{N}_0(\boldsymbol{x})}\right)^k \left(\frac{\mathcal{N}_0(\boldsymbol{x})}{\mathcal{N}_2(\boldsymbol{x})}\right)^k\right] \leq \left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}\left(\frac{\mathcal{N}_1(\boldsymbol{x})}{\mathcal{N}_0(\boldsymbol{x})}\right)^{2k}\right)^{1/2} \left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}\left(\frac{\mathcal{N}_0(\boldsymbol{x})}{\mathcal{N}_2(\boldsymbol{x})}\right)^{2k}\right)^{1/2}.$$

We can now bound each term independently. We start by the ratio of $\mathcal{N}_1/\mathcal{N}_0$. Without loss of generality we may assume that $\boldsymbol{\Sigma}^1$ is diagonal, $\boldsymbol{\Sigma}_1 = \boldsymbol{hrmdiag}(\lambda_1, \ldots, \lambda_d)$.

We also let $\boldsymbol{\mu}_1 = (\mu_1, \ldots, \mu_d)$. We write

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}\left[\left(\frac{\mathcal{N}_1(\boldsymbol{x})}{\mathcal{N}_0(\boldsymbol{x})}\right)^{2k}\right] &= \frac{1}{|\boldsymbol{\Sigma}_1|^k}\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}\left[\exp\left(-k(\boldsymbol{x}-\boldsymbol{\mu}_1)^T{\boldsymbol{\Sigma}_1}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + k\boldsymbol{x}^T\boldsymbol{x}\right)\right] \\
&= \frac{\exp(-k{\boldsymbol{\mu}_1}^T{\boldsymbol{\Sigma}_1}^{-1}\boldsymbol{\mu}_1)}{|\boldsymbol{\Sigma}_1|^k}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\exp\left(k\boldsymbol{x}^T(\boldsymbol{I}-\boldsymbol{\Sigma}_1^{-1})\boldsymbol{x} + 2k\boldsymbol{\mu}_1^T{\boldsymbol{\Sigma}_1}^{-1}\boldsymbol{x}\right)\right] \\
&\le \frac{1}{|\boldsymbol{\Sigma}_1|^k}\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}\left[\exp\left(\sum_{i=1}^d\left(k(1-1/\lambda_i)x_i^2 + 2k\frac{\mu_i}{\lambda_i}x_i\right)\right)\right] \\
&= \underbrace{\prod_{i=1}^d\frac{1}{\lambda_i^k}\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}\left[\exp\left(k(1-1/\lambda_i)x^2 + 2k\frac{\mu_i}{\lambda_i}x\right)\right]}_{A}
\end{aligned}$$

We now use the fact that for all $a < 1/2$.

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}[\exp(ax^2+bx)] = \frac{1}{\sqrt{1-2a}}\exp\left(\frac{b^2}{2-4a}\right)$$

At this point notice that since for all i it holds $\lambda_i < 2k/(2k-1)$ we have that term A is bounded. We get that

$$A = \underbrace{\exp\left(\sum_{i=1}^d\left(k\log\frac{1}{\lambda_i} - \frac{1}{2}\log\left(1-2k\left(1-\frac{1}{\lambda_i}\right)\right)\right)\right)}_{A_1}$$
$$\underbrace{\exp\left(\sum_{i=1}^d\frac{2k^2\mu_i^2}{\lambda_i^2(1-2k(1-1/\lambda_i))}\right)}_{A_2}$$

To bound the term A_1 we use the second inequality of Lemma E.12 to get

$$A_1 \le \exp\left(\sum_{i=1}^d k^2(1-1/\lambda_i)^2\left(1+\frac{1}{1-2k(1-1/\lambda_i)}\right)\right) \le \exp\left(\frac{2k^2B}{\delta}\right)$$

Bounding A_2 is easier

$$A_2 \leq \exp\left(\frac{2k^2 \left\|\boldsymbol{\mu}_1\right\|_2^2}{\lambda_{\min}^2 \delta}\right)$$

Combining the bounds for A_1 and A_2 we obtain

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\left(\frac{\mathcal{N}_1(\boldsymbol{x})}{\mathcal{N}_0(\boldsymbol{x})}\right)^{2k}\right] \leq \exp\left(\frac{10k^2}{\delta}B\right)$$

We now work similarly to bound the ratio $\mathcal{N}_0/\mathcal{N}_2$. We will again assume that $\boldsymbol{\Sigma}_2 = \boldsymbol{h}rmdiag(\lambda_1, \ldots, \lambda_d)$ and $\mu_2 = (\mu_1, \ldots, \mu_d)$. We have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\left(\frac{\mathcal{N}_0(\boldsymbol{x})}{\mathcal{N}_2(\boldsymbol{x})}\right)^{2k}\right] \\
&= \exp(k\boldsymbol{\mu}_2^T \Sigma_2^{-1} \boldsymbol{\mu}_2) \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[|\boldsymbol{\Sigma}_2|^k \exp\left(k\boldsymbol{x}^T(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{I})\boldsymbol{x} - 2k\boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1}\boldsymbol{x}\right)\right] \\
&\leq \exp((k+1)B) \prod_{i=1}^{d} \mathop{\mathbf{E}}_{x\sim\mathcal{N}_0}\left[\exp\left(k(1/\lambda_i - 1)x^2 - k\log(1/\lambda_i) - 2k(\mu_i/\lambda_i)x\right)\right] \\
&= \exp\left(\left(\frac{8k^2}{\delta} + k + 1\right) B\right) \exp\left(\sum_{i=1}^{d}\left(-k\log(1/\lambda_i) - \frac{1}{2}\log\left(1 - 2k(1/\lambda_i - 1)\right)\right)\right) \\
&\leq \exp\left(\left(\frac{10k^2}{\delta} + 4k^2 + k + 1\right) B\right),
\end{aligned}$$

where to obtain the last inequality we used the first inequality of Lemma E.12 and the bounds for the maximum and minimum eigenvalues of $\boldsymbol{\Sigma}_2$. Finally, plugging in the bounds for the two ratios we get for $i = 1, 2$

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\left(\frac{\mathcal{N}_{3-i}(\boldsymbol{x})}{\mathcal{N}_i(\boldsymbol{x})}\right)^{k}\right] \leq \exp\left(\frac{13k^2}{\delta}B\right).$$

Having the upper bound it is now easy to prove the lower bound using the convexity of $x \mapsto x^{-1}$ and Jensen's inequality.

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\left(\frac{\mathcal{N}_1(\boldsymbol{x})}{\mathcal{N}_2(\boldsymbol{x})}\right)^k\right] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\left(\frac{\mathcal{N}_2(\boldsymbol{x})}{\mathcal{N}_1(\boldsymbol{x})}\right)^{-k}\right] &\geq \left(\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[\left(\frac{\mathcal{N}_2(\boldsymbol{x})}{\mathcal{N}_1(\boldsymbol{x})}\right)^k\right]\right)^{-1} \\ &\geq \exp\left(-\frac{13k^2}{\delta}B\right).\end{aligned}$$

The proof of Lemma 6.14 For any $\rho \in (0,1)$, using identity E.1, we write

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[f(\boldsymbol{x})T_{1-\rho}(\boldsymbol{x})] = \sum_{V\in\mathbb{N}^d}(1-\rho)^{|V|}\widehat{f}(V)^2$$

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[f(\boldsymbol{x})^2 - f(\boldsymbol{x})T_{1-\rho}f(\boldsymbol{x})\right] &= \sum_{V\in\mathbb{N}^d}\widehat{f}(V)^2 - \sum_{V\in\mathbb{N}^d}(1-\rho)^{|V|}\widehat{f}(V)^2 \\ &= \sum_{V\in\mathbb{N}^d}\left(1-(1-\rho)^{|V|}\right)\widehat{f}(V)^2 \\ &\geq \sum_{|V|\geq 1/\rho}\left(1-(1-\rho)^{|V|}\right)\widehat{f}(V)^2 \\ &\geq \sum_{|V|\geq 1/\rho}\left(1-(1-\rho)^{1/\rho}\right)\widehat{f}(V)^2 \\ &\geq (1-1/\mathrm{e})\sum_{|V|\geq 1/\rho}\widehat{f}(V)^2\end{aligned}$$

The proof of Lemma 6.16 We first write

$$\begin{aligned}\frac{1}{2}\mathop{\mathbf{E}}_{(x,z)\sim D_\rho}[(r(\boldsymbol{x})-r(\boldsymbol{z}))^2] &= \frac{1}{2}\mathop{\mathbf{E}}_{(x,z)\sim D_\rho}\left[\frac{r(\boldsymbol{x})^2}{2}+\frac{r(\boldsymbol{z})^2}{2} - r(\boldsymbol{x})r(\boldsymbol{z})\right] \\ &= \mathop{\mathbf{E}}_{(x,z)\sim D_\rho}[r(\boldsymbol{x})^2 - r(\boldsymbol{z})r(\boldsymbol{x})].\end{aligned}$$

Let

$$\sum_{V\in\mathbb{N}^d}\widehat{r}(V)H_V(\boldsymbol{x})$$

be the Hermite expansion of $r(\boldsymbol{x})$. From Parseval's identity and the Hermite expansion of Ornstein–Uhlenbeck operator, (E.1) we have

$$\begin{aligned}\mathop{\mathbf{E}}_{(x,z)\sim D_\rho}[r(\boldsymbol{x})^2 - r(\boldsymbol{x})r(\boldsymbol{z})] &= \sum_{V\in\mathbb{N}^d} \widehat{r}(V)^2 - \sum_{V\in\mathbb{N}^d} (1-\rho)^{|V|}\widehat{r}(V)^2 \\ &\leq \rho \sum_{V\in\mathbb{N}^d} |V|\widehat{r}(V)^2,\end{aligned}$$

where the last inequality follows from Bernoulli's inequality $1-\rho|V| \leq (1-\rho)^{|V|}$. We know that (see for example Szegö (1967))

$$\frac{\partial}{\partial x_i} H_V(\boldsymbol{x}) = \frac{\partial}{\partial x_i} \prod_{v_i\in V} H_{v_i}(x_i) = \prod_{v_j\in V\setminus v_i} H_{v_j}(x_j)\sqrt{v_i}H_{v_i-1}(x_i)$$

Therefore,

$$\frac{\partial r(\boldsymbol{x})}{\partial x_i} = \sum_{V\in\mathbb{N}^d} \widehat{r}(V)\sqrt{v_i}H_{v_i-1}(x_i) \prod_{v_j\in V\setminus v_i} H_{v_j}(x_j)$$

From Parseval's identity we have

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left(\frac{\partial r(\boldsymbol{x})}{\partial x_i}\right)^2\right] = \sum_{V\in\mathbb{N}^d} \widehat{r}(V)^2 v_i.$$

Therefore,

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\|\nabla r(\boldsymbol{x})\|_2^2\right] = \sum_{V\in\mathbb{N}^d} |V|\widehat{r}(V)^2.$$

The lemma follows.

Learning the Hermite Expansion

In this section we present a way to bound the variance of the empirical estimation of Hermite coefficients. To bound the variance of estimating Hermite polynomials we shall need a bound for the expected value of the fourth power of a Hermite polynomial.

Lemma E.13. *For any $V \in \mathbb{N}^d$ it holds $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[H_V^4(\boldsymbol{x})] \le 9^{|V|}$.*

Proof. We compute

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[H_V^4(\boldsymbol{x})] &= \prod_{v_i\in V} \mathop{\mathbf{E}}_{x\sim\mathcal{N}(0,1)}[H_{v_i}^2(x_i)H_{v_i}^2(x_i)] \\
&= \prod_{v_i\in V} \mathop{\mathbf{E}}_{x\sim\mathcal{N}(0,1)}\left[\left(\sum_{r=0}^{v_i}\binom{v_i}{r}\frac{\sqrt{2r!}}{r!}H_{2r}(x_i)\right)\left(\sum_{r=0}^{v_i}\binom{v_i}{r}\frac{\sqrt{2r!}}{r!}H_{2r}(x_i)\right)\right] \\
&= \prod_{v_i\in V}\sum_{r=0}^{v_i}\binom{v_i}{r}^2\frac{(2r)!}{(r!)^2}\mathop{\mathbf{E}}_{x\sim\mathcal{N}(0,1)}\left[H_{2r}(x_i)^2\right] = \prod_{v_i\in V}\sum_{r=0}^{v_i}\binom{v_i}{r}^2\frac{(2r)!}{(r!)^2} \\
&\le \prod_{v_i\in V}\sum_{r=0}^{v_i}\binom{v_i}{r}^2 2^{2r} \le \prod_{v_i\in V}\left(\sum_{r=0}^{v_i}\binom{v_i}{r}2^r\right)^2 \le \prod_{v_i\in V} 9^{v_i} = 9^{|V|}.
\end{aligned}$$

In the above computation we used the formula for the product of two (normalized) Hermite polynomials

$$H_i(x)H_i(x) = \sum_{r=0}^{v_i}\binom{v_i}{r}\frac{\sqrt{2r!}}{r!}H_{2r}(x_i),$$

see, for example, Szegö (1967). □

The proof of Lemma 6.17 We have

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}[(H_V(\boldsymbol{x})-c_V)^2] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}[H_V^2(\boldsymbol{x})] - c_V^2 \le \frac{1}{\alpha}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}[H_V^2(\boldsymbol{x})]$$

We have

$$\begin{aligned}\left|\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}[H_V^2(\boldsymbol{x})]-1\right| &= \left|\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}[H_V^2(\boldsymbol{x})]-\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[H_V^2(\boldsymbol{x})]\right| \\ &\leq \int H_V^2(\boldsymbol{x})|\mathcal{N}^*(\boldsymbol{x})-\mathcal{N}_0(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} \\ &= \int H_V^2(\boldsymbol{x})\sqrt{\mathcal{N}_0(\boldsymbol{x}))}\frac{|\mathcal{N}^*(\boldsymbol{x})-\mathcal{N}_0(\boldsymbol{x})|}{\sqrt{\mathcal{N}_0(\boldsymbol{x})}}\mathrm{d}\boldsymbol{x} \\ &\leq \Big(\underbrace{\int H_V^4(\boldsymbol{x})\mathcal{N}_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}_{A}\Big)^{1/2}\Big(\underbrace{\int\frac{(\mathcal{N}^*(\boldsymbol{x})-\mathcal{N}_0(\boldsymbol{x}))^2}{\mathcal{N}_0(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}}_{B}\Big)^{1/2}\end{aligned}$$

To bound term A we use Lemma E.13. Using Lemma 6.13 we obtain

$$B \leq \mathop{\mathbf{E}}_{x\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\left(\frac{\mathcal{N}^*(\boldsymbol{x})}{\mathcal{N}_0(\boldsymbol{x})}\right)^2\right] \leq \mathrm{poly}(1/\alpha).$$

The bound for the variance follows from the independence of the samples.

E.4 Missing Proofs of Section 6.3

The proof of Lemma 6.23 We have that

$$\left|M_{\psi_k}(\boldsymbol{u},\boldsymbol{B})-M'_{\psi}(\boldsymbol{u},\boldsymbol{B})\right| \leq \left|M_{\psi_k}(\boldsymbol{u},\boldsymbol{B})-M_{\psi}(\boldsymbol{u},\boldsymbol{B})\right| + \left|M_{\psi}(\boldsymbol{u},\boldsymbol{B})-M'_{\psi}(\boldsymbol{u},\boldsymbol{B})\right|.$$

For the first term we have that

$$\begin{aligned}\left|M_{\psi_k}(\boldsymbol{u},\boldsymbol{B})-M'_{\psi}(\boldsymbol{u},\boldsymbol{B})\right| &\leq C_{\boldsymbol{u},\boldsymbol{B}} \mathop{\mathbf{E}}_{x\sim\mathcal{N}_S^*}\left[\frac{\mathbb{N}_0(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}\left|\psi_k(\boldsymbol{x})-\psi(\boldsymbol{x})\right|\right]\\ &\leq C_{\boldsymbol{u},\boldsymbol{B}}\sqrt{\mathop{\mathbf{E}}_{x\sim\mathbb{N}_0}\left[\left(\frac{\mathcal{N}_S^*(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}\right)^2\right]\cdot\mathop{\mathbf{E}}_{x\sim\mathbb{N}_0}\left[(\psi_k(\boldsymbol{x})-\psi(\boldsymbol{x}))^2\right]}\\ &\leq \frac{C_{\boldsymbol{u},\boldsymbol{B}}}{\alpha^*}\sqrt{\mathop{\mathbf{E}}_{x\sim\mathbb{N}_0}\left[\left(\frac{\mathcal{N}^*(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}\right)^2\right]\cdot\mathop{\mathbf{E}}_{x\sim\mathbb{N}_0}\left[(\psi_k(\boldsymbol{x})-\psi(\boldsymbol{x}))^2\right]}\end{aligned}$$

now we can use Lemma 6.13, Lemma 6.19 and Theorem 6.18 to get

$$\left|M_{\psi_k}(\boldsymbol{u},\boldsymbol{B})-M'_{\psi}(\boldsymbol{u},\boldsymbol{B})\right| \leq \mathrm{poly}(1/\alpha^*)\sqrt{\epsilon}$$

For the second term we have that

$$\left|M_{\psi}(\boldsymbol{u},\boldsymbol{B})-M'_{\psi}(\boldsymbol{u},\boldsymbol{B})\right| \leq \left|1-\frac{C'_{\boldsymbol{u},\boldsymbol{B}}}{C_{\boldsymbol{u},\boldsymbol{B}}}\right| C_{\boldsymbol{u},\boldsymbol{B}} \mathop{\mathbf{E}}_{x\sim\mathcal{N}_S^*}\left[\frac{\mathcal{N}^*(\boldsymbol{x})}{\alpha^*\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}\right]$$

We need to bound

$$\begin{aligned}&\left|1-\frac{C'_{\boldsymbol{u},\boldsymbol{B}}}{C_{\boldsymbol{u},\boldsymbol{B}}}\right|\\ &=\left|1-e^{-\frac{1}{2}\left(\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\boldsymbol{\Sigma}_S+\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T-\widetilde{\boldsymbol{\Sigma}}_S)))-\boldsymbol{u}^T\boldsymbol{\mu}_S\right)}\right|\\ &\leq e^{\left|\frac{1}{2}\left(\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\boldsymbol{\Sigma}_S+\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T-\widetilde{\boldsymbol{\Sigma}}_S)))-\boldsymbol{u}^T\boldsymbol{\mu}_S\right)\right|}-1\\ &\leq e^{\frac{1}{2}\left(\|\boldsymbol{B}-\boldsymbol{I}\|_F\|\boldsymbol{\Sigma}_S+\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T-\widetilde{\boldsymbol{\Sigma}}_S\|_F+\|\boldsymbol{u}\|_2\|\boldsymbol{\mu}_S\|_2\right)}-1\\ &\leq \|\boldsymbol{B}-\boldsymbol{I}\|_F\|\boldsymbol{\Sigma}_S+\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T-\widetilde{\boldsymbol{\Sigma}}_S\|_F+\|\boldsymbol{u}\|_2\|\boldsymbol{\mu}_S\|_2\end{aligned}$$

where the last inequality holds when $\|\boldsymbol{B}-\boldsymbol{I}\|_F\|\boldsymbol{\Sigma}_S+\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T-\widetilde{\boldsymbol{\Sigma}}_S\|_F+\|\boldsymbol{u}\|_2\|\boldsymbol{\mu}_S\|_2 \leq 1$. But we know that $(\boldsymbol{u},\boldsymbol{B}) \in \mathcal{D}$ and hence $\|\boldsymbol{B}-\boldsymbol{I}\|_F \leq \mathrm{poly}(1/\alpha^*)$, $\|\boldsymbol{u}\|_2 \leq \mathrm{poly}(1/\alpha^*)$. Also from Section 6.1 we have that $\|\boldsymbol{\Sigma}_S+\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T-\widetilde{\boldsymbol{\Sigma}}_S\|_F \leq \epsilon$ and

$\|\boldsymbol{\mu}_S\|_2 \leq \epsilon$ and we can set ϵ to be any inverse polynomial in $1/\alpha^*$ times ϵ. Hence we get

$$\left|1 - \frac{C'_{\boldsymbol{u},\boldsymbol{B}}}{C_{\boldsymbol{u},\boldsymbol{B}}}\right| \leq \epsilon$$

Now we can also use Lemma 6.19 and Lemma 6.13 which imply that

$$C_{\boldsymbol{u},\boldsymbol{B}} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\frac{\mathcal{N}^*(\boldsymbol{x})}{\alpha^*\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}\right] \leq \mathrm{poly}(1/\alpha^*)$$

and therefore we have

$$\left|M_\psi(\boldsymbol{u},\boldsymbol{B}) - M'_\psi(\boldsymbol{u},\boldsymbol{B})\right| \leq \mathrm{poly}(1/\alpha^*)\epsilon.$$

Hence we can once again divide ϵ by any polynomial of $1/\alpha^*$ without increasing the complexity presented in Section 6.1 and the lemma follows.

The proof of Lemma 6.23 We apply successive Cauchy-Schwarz inequalities to separate the terms that appear in the expression for the squared norm of the gradient. We have that

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\|\boldsymbol{v}(\boldsymbol{u},\boldsymbol{B})\|_2^2\right] \\
&= C_{\boldsymbol{u},rB}^2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\left(\left\|\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right\|_F^2 + \|\tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}\|_2^2\right)\frac{\mathbb{N}_0^2(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},rB}^2(x)}\psi_k^2(\boldsymbol{x})\right] \\
&= C_{\boldsymbol{u},rB}^2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\left(\left\|\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right\|_F^2 + \|\tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}\|_2^2\right)\frac{\mathbb{N}_0(\boldsymbol{x})\mathcal{N}_S^*(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},rB}^2(x)}\psi_k^2(\boldsymbol{x})\right] \\
&\le C_{\boldsymbol{u},rB}^2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\left(\left\|\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right\|_F^2 + \|\tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}\|_2^2\right)\frac{\mathbb{N}_0(\boldsymbol{x})\mathcal{N}_S^*(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},rB}^2(x)}\right]^{1/2}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\psi_k^4(\boldsymbol{x})\right]^{1/2} \\
&\le C_{\boldsymbol{u},rB}^2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\left(\left\|\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right\|_F^2 + \|\tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}\|_2^2\right)^2\right]^{1/4} \\
&\qquad \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\frac{\mathbb{N}_0^2(\boldsymbol{x})\mathcal{N}^*{}_S^2(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},rB}^4(x)}\right]^{1/4}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\psi_k^4(\boldsymbol{x})\right]^{1/2} \\
&\le C_{\boldsymbol{u},rB}^2 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\left(\left\|\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right\|_F + \|\tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}\|_2\right)^4\right]^{1/4} \\
&\qquad \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\frac{\mathbb{N}_0^4(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},rB}^4(x)}\right]^{1/8}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\frac{\mathcal{N}^*{}_S^4(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},rB}^4(x)}\right]^{1/8}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\psi_k^4(\boldsymbol{x})\right]^{1/2}
\end{aligned}
$$

We now bound each term separately.

- By Lemma 6.19, $C_{\boldsymbol{u},rB}^2 \le \operatorname{poly}(1/\alpha)$.
- Given that $(\tilde{\boldsymbol{\mu}}_S, \tilde{\boldsymbol{\Sigma}}_S)$ are near-isotropic,

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\left(\left\|\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right\|_F + \|\tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}\|_2\right)^4\right]^{1/4} \\
&\quad \le \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\left(\left\|\boldsymbol{x}\boldsymbol{x}^T\right\|_F + \|\tilde{\boldsymbol{\Sigma}}_S\|_F + \left\|\tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right\|_F + \|\tilde{\boldsymbol{\mu}}_S\| + \|\boldsymbol{x}\|_2\right)^4\right]^{1/4} \\
&\quad \le d\operatorname{poly}(1/\alpha).
\end{aligned}
$$

- By Lemma 6.13,

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\frac{\mathbb{N}_0^4(\boldsymbol{x})}{\mathbb{N}^4_{\boldsymbol{u},rB}(x)}\right]^{1/8}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\frac{\mathcal{N}^{*4}_{S}(\boldsymbol{x})}{\mathbb{N}^4_{\boldsymbol{u},rB}(x)}\right]^{1/8}\leq \mathrm{poly}(1/\alpha).$$

- For the last term, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\psi_k^4(\boldsymbol{x})\right] &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\left(\sum_{0\leq|V|\leq k}\tilde{c}_V H_V(\boldsymbol{x})\right)^4\right]\\ &\leq 2^3\sum_{0\leq|V|\leq k}\tilde{c}_V^4\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[H_V^4(\boldsymbol{x})\right]\\ &\leq 8\left(\sum_{0\leq|V|\leq k}\tilde{c}_V^2\right)^2\cdot\left(\max_{0\leq|V|\leq k}\left\{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[H_V^4(\boldsymbol{x})\right]\right\}\right)\end{aligned}$$

From Lemma 6.17 and the conditioning on the event that the estimators of the Hermite coefficients are accurate we have that $(\tilde{c}_V - c_V)^2 \leq 1$ and hence we get the following.

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[\psi_k^4(\boldsymbol{x})\right]\leq 2^{10}d^{2k}\left(\sum_{0\leq|V|\leq\infty}c_V^2\right)^4\cdot\left(\max_{0\leq|V|\leq k}\left\{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[H_V^4(\boldsymbol{x})\right]\right\}\right)$$

To bound $\mathbf{E}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[H_V^4(\boldsymbol{x})\right]$ we use Lemma E.13. Moreover, from Parseval's identity we obtain that $\sum_{0\leq|V|\leq\infty}c_V^2 = \mathbf{E}_{\boldsymbol{x}\sim\mathbb{N}_0}\psi^2(\boldsymbol{x})$. From Lemma 6.13 we get

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathbb{N}_0}\psi^2(\boldsymbol{x})\leq\ \frac{1}{\alpha}\mathop{\mathbf{E}}_{x\sim N_0}\left(\frac{\mathcal{N}^*(\boldsymbol{x})}{\mathcal{N}_0(\boldsymbol{x})}\right)^2 = \mathrm{poly}(1/\alpha).$$

From Lemma 6.17 we obtain that $\max_{0\leq|V|\leq k}\left\{\mathbf{E}_{\boldsymbol{x}\sim\mathbb{N}_0}\left[H_V^4(\boldsymbol{x})\right]\right\}\leq 2^k$. The result follows from the above estimates.

The proof of Lemma 6.24 We will prove this lemma in two steps, first we will prove

$$\left| \boldsymbol{z}^T \mathcal{H}_{M_{\psi_k}}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} - \boldsymbol{z}^T \mathcal{H}_{M_{\psi}}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} \right| \leq \lambda \tag{E.5}$$

and then we will prove that

$$\boldsymbol{z}^T \mathcal{H}_{M_{\psi}}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} \geq 2\lambda \tag{E.6}$$

for some parameter $\lambda \geq \operatorname{poly}(\alpha^*)$. To prove (E.5) we define

$$p(\boldsymbol{z}; \boldsymbol{x}) = \left(\boldsymbol{z}^T \begin{pmatrix} \frac{1}{2} \left(\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S \tilde{\boldsymbol{\mu}}_S^T \right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x} \end{pmatrix} \right)^2$$

and we have that

$$\begin{aligned} &\left| \boldsymbol{z}^T \mathcal{H}_{M_{\psi_k}}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} - \boldsymbol{z}^T \mathcal{H}_{M_{\psi}}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} \right| \\ &\quad = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_S^*} \left[e^{h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x}) \cdot p(\boldsymbol{z}; \boldsymbol{x}) \cdot |\psi_k(\boldsymbol{x}) - \psi(\boldsymbol{x})| \right] \\ &\quad = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{N}_0} \left[e^{h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \cdot \mathbb{1}_S(\boldsymbol{x}) \cdot \mathcal{N}^*(\boldsymbol{x}) \cdot p(\boldsymbol{z}; \boldsymbol{x}) \cdot |\psi_k(\boldsymbol{x}) - \psi(\boldsymbol{x})| \right] \end{aligned}$$

we then separate the terms using the Cauchy Schwarz inequality

$$\begin{aligned} &\left| \boldsymbol{z}^T \mathcal{H}_{M_{\psi_k}}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} - \boldsymbol{z}^T \mathcal{H}_{M_{\psi}}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} \right| \\ &\quad \leq \sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{N}_0} \left[e^{2h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \cdot \mathbb{1}_S(\boldsymbol{x}) \cdot (\mathcal{N}^*(\boldsymbol{x}))^2 \cdot p^2(\boldsymbol{z}; \boldsymbol{x}) \right]} \cdot \sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{N}_0} \left[(\psi_k(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2 \right]} \end{aligned}$$

we apply now the Hermite concentration from Theorem 6.18 and we get

$$\begin{aligned} &\leq \sqrt{\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbb{N}_0} \left[e^{2h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \cdot \mathbb{1}_S(\boldsymbol{x}) \cdot (\mathcal{N}^*(\boldsymbol{x}))^2 \cdot p^2(\boldsymbol{z}; \boldsymbol{x}) \right]} \cdot \sqrt{\epsilon} \\ &\leq \sqrt[4]{\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}^*} \left[e^{4h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})} \cdot \mathbb{1}_S(\boldsymbol{x}) \cdot (\mathcal{N}^*(\boldsymbol{x}))^2 \left(\mathbb{N}_0(\boldsymbol{x}) \right)^2 \right]} \cdot \sqrt[4]{\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}^*} \left[p^4(\boldsymbol{z}; \boldsymbol{x}) \right]} \cdot \sqrt{\epsilon} \end{aligned}$$

we now use (6.8), Lemma 6.19 and the fact that $\mathbb{1}_S(\boldsymbol{x}) \leq 1$ to get

$$\leq \sqrt[4]{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[\mathrm{e}^{4h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\left(\mathcal{N}^*(\boldsymbol{x})\right)^2\left(\mathbb{N}_0(\boldsymbol{x})\right)^2\right]} \cdot \sqrt[4]{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p^4(\boldsymbol{z};\boldsymbol{x})\right]} \cdot \sqrt{\epsilon}$$

and finally we use Lemma 6.13 to prove the following

$$\left|\boldsymbol{z}^T\mathcal{H}_{M_{\psi_k}}(\boldsymbol{u},\boldsymbol{B})\boldsymbol{z} - \boldsymbol{z}^T\mathcal{H}_{M_\psi}(\boldsymbol{u},\boldsymbol{B})\boldsymbol{z}\right| \leq \sqrt[4]{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p^4(\boldsymbol{z};\boldsymbol{x})\right]} \cdot \mathrm{poly}(1/\alpha^*) \cdot \sqrt{\epsilon} \quad \text{(E.7)}$$

Next we prove (E.6). We have that

$$\begin{aligned}\boldsymbol{z}^T\mathcal{H}_{M_\psi}(\boldsymbol{u},\boldsymbol{B})\boldsymbol{z} &= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x}) \cdot p(\boldsymbol{z};\boldsymbol{x}) \cdot \psi(\boldsymbol{x})\right] \\ &= \frac{1}{\alpha^*}C_{\boldsymbol{u},\boldsymbol{B}} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\frac{\mathbb{N}^*(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}p(\boldsymbol{z};\boldsymbol{x})\right].\end{aligned}$$

Now we define the set $\bar{Q}_{\boldsymbol{z}} = \left\{\boldsymbol{x} \in \mathbb{R}^d \mid |p(\boldsymbol{z};\boldsymbol{x})| \leq \frac{1}{32C}\left(\alpha^*\right)^4 \sqrt[4]{\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p^4(\boldsymbol{z};\boldsymbol{x})\right]}\right\}$, where C is the universal constant guaranteed from Theorem E.4. Then using Theorem E.4 and the fact that $p(\boldsymbol{z};\boldsymbol{x})$ has degree 4 we get that $\mathbb{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\bar{Q}) \leq \frac{\alpha^*}{2}$. Hence we define the set $S' = S \cap \bar{Q}$ and we have that $\mathbb{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;S') \geq \alpha^*/2$.

$$\begin{aligned}\boldsymbol{z}^T\mathcal{H}_{M_\psi}(\boldsymbol{u},\boldsymbol{B})\boldsymbol{z} &\geq \frac{1}{\alpha^*}C_{\boldsymbol{u},\boldsymbol{B}} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_{S'}^*}\left[\frac{\mathbb{N}^*(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}p(\boldsymbol{z};\boldsymbol{x})\right] \\ &\geq \left(\min_{\boldsymbol{x}\in S'} p(\boldsymbol{z};\boldsymbol{x})\right)\frac{1}{\alpha^*}C_{\boldsymbol{u},\boldsymbol{B}} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_{S'}^*}\left[\frac{\mathbb{N}^*(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}\right]\end{aligned}$$

and from the definition of S' and Lemma 6.19 we have that

$$\boldsymbol{z}^T\mathcal{H}_{M_\psi}(\boldsymbol{u},\boldsymbol{B})\boldsymbol{z} \geq \mathrm{poly}(\alpha^*) \cdot \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_{S'}^*}\left[\frac{\mathbb{N}^*(\boldsymbol{x})}{\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}\right] \cdot \sqrt[4]{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p^4(\boldsymbol{z};\boldsymbol{x})\right]}$$

now we can apply Jensen's inequality on the convex function $x \mapsto 1/x$ and we get

$$\begin{aligned} \boldsymbol{z}^T \mathcal{H}_{M_\psi}(\boldsymbol{u}, \boldsymbol{B})\boldsymbol{z} &\geq \text{poly}(\alpha^*) \cdot \frac{1}{\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}^*_{S'}}\left[\frac{\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}{\mathbb{N}^*(\boldsymbol{x})}\right]} \cdot \sqrt[4]{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p^4(\boldsymbol{z};\boldsymbol{x})\right]} \\ &\geq \text{poly}(\alpha^*) \cdot \frac{1}{\sqrt{\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[\left(\frac{\mathbb{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})}{\mathcal{N}^*(\boldsymbol{x})}\right)^2\right]}} \cdot \sqrt[4]{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p^4(\boldsymbol{z};\boldsymbol{x})\right]} \end{aligned}$$

finally using Lemma 6.13 we get

$$\boldsymbol{z}^T \mathcal{H}_{M_\psi}(\boldsymbol{u}, \boldsymbol{B})\boldsymbol{z} \geq \text{poly}(\alpha^*) \sqrt[4]{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p^4(\boldsymbol{z};\boldsymbol{x})\right]} \tag{E.8}$$

Now using (E.7) and (E.8) we can see that it is possible to pick ϵ in the Hermite concentration to be the correct polynomial in α^* so that

$$\left|\boldsymbol{z}^T \mathcal{H}_{M_{\psi_k}}(\boldsymbol{u}, \boldsymbol{B})\boldsymbol{z} - \boldsymbol{z}^T \mathcal{H}_{M_\psi}(\boldsymbol{u}, \boldsymbol{B})\boldsymbol{z}\right| \leq \boldsymbol{z}^T \mathcal{H}_{M_\psi}(\boldsymbol{u}, \boldsymbol{B})\boldsymbol{z}$$

which implies from Jensen's inequality that

$$\begin{aligned} \boldsymbol{z}^T \mathcal{H}_{M_{\psi_k}}(\boldsymbol{u}, \boldsymbol{B})\boldsymbol{z} &\geq \text{poly}(\alpha^*) \sqrt[4]{\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p^4(\boldsymbol{z};\boldsymbol{x})\right]} \\ &\geq \text{poly}(\alpha^*) \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p(\boldsymbol{z};\boldsymbol{x})\right] \end{aligned}$$

So the last step is to prove a lower bound for $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p(\boldsymbol{z};\boldsymbol{x})\right]$. For this we can use the Lemma 3 of Daskalakis et al. (2018) from which we can directly get $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}^*}\left[p(\boldsymbol{z};\boldsymbol{x})\right] \geq \text{poly}(\alpha^*)$ and the lemma follows.

E.5 Details of Section 6.3

We present here the of the proof of Theorem 6.27. We already proved that given only positive examples from a truncated normal can obtain arbitrarily good estimations of the unconditional (true) parameters of the normal using Algorithm 11. Recall also that with positive samples we can obtain an approximation of the function $\psi(\boldsymbol{x})$ defined in 6.4. From Theorem 6.18 we know that with $d^{\mathrm{poly}(1/\alpha)\Gamma(S)^2/\epsilon^4}$ samples we can obtain a function $\psi_k(\boldsymbol{x})$ such that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[((\psi_k(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2] \leq \epsilon.$$

Now we can construct an almost indicator function using ψ_k and the learned parameters $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{I}}$. We denote $\widetilde{\mathcal{N}} = \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$.

$$\widetilde{f}(x) = \frac{\mathcal{N}_0(\boldsymbol{x})}{\widetilde{\mathcal{N}}(\boldsymbol{x})}\psi_k(\boldsymbol{x}). \tag{E.9}$$

This function should be a good enough approximation to the function

$$f(\boldsymbol{x}) = \frac{\mathcal{N}_0(\boldsymbol{x})}{\mathcal{N}^*(\boldsymbol{x})}\psi(\boldsymbol{x}) = \frac{\mathbb{1}_S(\boldsymbol{x})}{\alpha^*}. \tag{E.10}$$

Notice that even though we do not know the mass of the truncation set α^* we can still construct a threshold function that achieves low error with respect to the zero-one loss. We first prove a standard lemma that upper bounds the zero-one loss with the distance of f and $\widetilde{f}$. We prove it so that we have a version consistent with our notation.

Lemma E.14. *Let S be a subset of $\mathbb{R}^d$. Let D be a distribution on $\mathbb{R}^d$ and let $f : \mathbb{R}^d \to \{0, B\}$, where $B > 1$ such that $f(\boldsymbol{x}) = B\,\mathbb{1}_S(\boldsymbol{x})$. For any $g : \mathbb{R}^d \mapsto [0, +\infty)$ it holds $\mathbf{E}_{x\sim D}\left[\mathbf{1}\{g(x) > 1/2)\} \neq \mathbb{1}_S(\boldsymbol{x})\}\right] \leq \sqrt{2}\,\mathbf{E}_{x\sim D}\left[\sqrt{|g(x) - f(x)|}\right]$.*

Proof. It suffices to show that for all $x \in \mathbb{R}^d$ it holds

$$\mathbf{1}\{\mathrm{sign}(g(x) - 1/2) \neq S(\boldsymbol{x})\} \leq \sqrt{2}\sqrt{|g(x) - f(x)|}. \tag{E.11}$$

We only need to consider the case where $\mathrm{sign}(g(x) - 1/2) \neq S(\boldsymbol{x})$. Assume first that $g(x) > 1/2$ and $x \neq S$. Then the LHS of Equation (E.11) is 1 and the RHS of (E.11) is $\sqrt{2}\sqrt{|g(x) - f(x)|} \geq \sqrt{2}\sqrt{|1/2 - 0|} = 1$. Assume now that

$g(x) < 1/2$ and $S(x) = 1$. Then the RHS of (E.11) equals $\sqrt{2}\sqrt{|g(x) - f(x)|} \geq \sqrt{2}\sqrt{|B - 1/2|} \geq 1$. □

We now state the following lemma that upper bounds the distance of f and $\widetilde{f}$ in with the sum of the total variation distance of the true and learned distributions as well as the approximation error of ψ_k.

Lemma E.15. *Let α be the absolute constant of* (6.3). *Let $S \subseteq \mathbb{R}^d$ and let $\mathcal{N}^*, \widetilde{\mathcal{N}}$ be $(O(\log(1/\alpha)), 1/16)$-isotropic. Let ψ be as in* (6.4). *Moreover, let $\widetilde{f}, f$ be as in* (E.9), (E.10). *Then,*

$$\mathop{\mathbf{E}}_{x \sim \mathcal{N}^*}\left[\sqrt{|\widetilde{f}(x) - f(x)|}\right] \leq \mathrm{poly}(1/\alpha)\left(\left(\mathop{\mathbf{E}}_{x \sim \mathcal{N}_0}\left[(\psi_k(x) - \psi(x))^2\right]\right)^{1/4} + \left(d_{\mathrm{TV}}(\mathcal{N}^*, \widetilde{\mathcal{N}})\right)^{1/4}\right)$$

Proof. We compute

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x \sim \mathcal{N}^*}\left[\sqrt{|\widetilde{f}(x) - f(x)|}\right] \\
&\leq \mathop{\mathbf{E}}_{x \sim \mathcal{N}^*}\left[\left(\left|\psi_k(x)\frac{\mathcal{N}_0(x)}{\widetilde{\mathcal{N}}(x)} - \psi(x)\frac{\mathcal{N}_0(x)}{\mathcal{N}^*(x)}\right|\right)^{1/2}\right] \\
&= \mathop{\mathbf{E}}_{x \sim \mathcal{N}^*}\left[\left(\left|\psi_k(x)\frac{\mathcal{N}_0(x)}{\widetilde{\mathcal{N}}(x)} - \psi(x)\frac{\mathcal{N}_0(x)}{\widetilde{\mathcal{N}}(x)} + \psi(x)\frac{\mathcal{N}_0(x)}{\widetilde{\mathcal{N}}(x)} - \psi(x)\frac{\mathcal{N}_0(x)}{\mathcal{N}^*(x)}\right|\right)^{1/2}\right] \\
&\leq \mathop{\mathbf{E}}_{x \sim \mathcal{N}^*}\left[\left(|\psi_k(x) - \psi(x)|\frac{\mathcal{N}_0(x)}{\widetilde{\mathcal{N}}(x)}\right)^{1/2}\right] + \mathop{\mathbf{E}}_{x \sim \mathcal{N}^*}\left[\left(\psi(x)\left|\frac{\mathcal{N}_0(x)}{\widetilde{\mathcal{N}}(x)} - \frac{\mathcal{N}_0(x)}{\mathcal{N}^*(x)}\right|\right)^{1/2}\right] \\
&\leq \Bigg(\underbrace{\mathop{\mathbf{E}}_{x \sim \mathcal{N}^*}\left[|\psi_k(x) - \psi(x)|\right]}_{A}\Bigg)^{1/2}\Bigg(\underbrace{\mathop{\mathbf{E}}_{x \sim \mathcal{N}^*}\left[\frac{\mathcal{N}_0(x)}{\widetilde{\mathcal{N}}(x)}\right]}_{B}\Bigg)^{1/2} \\
&+ \Bigg(\underbrace{\mathop{\mathbf{E}}_{x \sim \mathcal{N}^*}\left[\psi(x)\left|\frac{\mathcal{N}_0(x)}{\widetilde{\mathcal{N}}(x)} - \frac{\mathcal{N}_0(x)}{\mathcal{N}^*(x)}\right|\right]}_{C}\Bigg)^{1/2}
\end{aligned}$$

where for term C we used Jensen's inequality. Using Lemma E.16 and Lemma 6.13 we have that

$$\begin{aligned} A &\leq \left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0} \left[(\psi_k(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2 \right] \right)^{1/2} \left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}^*} \left[\frac{\mathcal{N}_0(\boldsymbol{x})}{\mathcal{N}^*(\boldsymbol{x})} \right] \right)^{1/2} \\ &\leq \left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0} \left[(\psi_k(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2 \right] \right)^{1/2} \mathrm{poly}(1/\alpha) \end{aligned}$$

Since $\mathcal{N}_0, \widetilde{\mathcal{N}}$, and $\mathcal{N}^*$ are $(O(\log(1/\alpha), 1/16)$-isotropic, using Lemma 6.13 we obtain that

$$B = \mathop{\mathbf{E}}_{x\sim\mathcal{N}_0} \left[\frac{\mathcal{N}_0(\boldsymbol{x})}{\widetilde{\mathcal{N}}(\boldsymbol{x})} \frac{\mathcal{N}^*(\boldsymbol{x})}{\widetilde{\mathcal{N}}(\boldsymbol{x})} \right] \leq \left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0} \left[\frac{\mathcal{N}_0(\boldsymbol{x})}{\widetilde{\mathcal{N}}(\boldsymbol{x})} \right] \right)^{1/2} \left(\mathop{\mathbf{E}}_{x\sim\mathcal{N}_0} \left[\frac{\mathcal{N}^*(\boldsymbol{x})}{\widetilde{\mathcal{N}}(\boldsymbol{x})} \right] \right)^{1/2} \leq \mathrm{poly}(1/\alpha)$$

We now bound term C. We write

$$C = \mathop{\mathbf{E}}_{x\sim\mathcal{N}^*} \left[\psi(\boldsymbol{x}) \left| \frac{\mathcal{N}_0(\boldsymbol{x})}{\widetilde{\mathcal{N}}(\boldsymbol{x})} - \frac{\mathcal{N}_0(\boldsymbol{x})}{\mathcal{N}^*(\boldsymbol{x})} \right| \right] = \frac{1}{\alpha^*} \mathop{\mathbf{E}}_{x\sim\mathcal{N}^*} \left[\left| \frac{\mathcal{N}^*(\boldsymbol{x})}{\widetilde{\mathcal{N}}(\boldsymbol{x})} - 1 \right| \right] \tag{E.12}$$

To simplify notation, let $\ell(\boldsymbol{x}) = \left| \frac{\mathcal{N}^*(\boldsymbol{x})}{\widetilde{\mathcal{N}}(\boldsymbol{x})} - 1 \right|$. Moreover, notice that $\mathbf{E}_{x\sim\widetilde{\mathcal{N}}}[\ell(\boldsymbol{x})] = d_{\mathrm{TV}}(\mathcal{N}^*, \widetilde{\mathcal{N}})$. Using the second bound of Lemma E.16 and Lemma 6.13 we obtain

$$C \leq \frac{1}{\alpha} d_{\mathrm{TV}}(\mathcal{N}^*, \widetilde{\mathcal{N}}) + \mathrm{poly}(1/\alpha) \sqrt{d_{\mathrm{TV}}(\mathcal{N}^*, \widetilde{\mathcal{N}})} \leq \mathrm{poly}(1/\alpha) \sqrt{d_{\mathrm{TV}}(\mathcal{N}^*, \widetilde{\mathcal{N}})}.$$

Combining the bounds for A, B and C we obtain the result. □

Since we have the means two make both errors of Lemma E.15 small we can now recover the unknown truncation set S.

The proof of Theorem 6.27 We first run Algorithm 11 to find estimates $\widetilde{\boldsymbol{\mu}}$, $\widetilde{\boldsymbol{\Sigma}}$. From Theorem 6.4 we know that $N = d^{\mathrm{poly}(1/\alpha)\Gamma^2(\mathcal{S})/\epsilon^{32}}$ samples suffice to obtain parameters $\widetilde{\boldsymbol{\mu}}$, $\widetilde{\boldsymbol{\Sigma}}$ such that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})) \leq \mathrm{poly}(\alpha)\epsilon^4$. Notice, that from Theorem 6.12 we also know that N samples from the conditional distribution $\mathcal{N}_S^*$

suffice to learn a function ψ_k such that $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[(\psi_k(\boldsymbol{x})-\psi(\boldsymbol{x}))^2] \leq \mathrm{poly}(\alpha)\epsilon^4$. Now we can construct the approximation $\widetilde{f}(\boldsymbol{x}) = \psi_k(\boldsymbol{x})\mathcal{N}_0(\boldsymbol{x})/\widetilde{\mathcal{N}}(\boldsymbol{x})$. Let our indicator $\widetilde{S} = \mathbf{1}\{\widetilde{f}((x) > 1/2\}$ and from Lemma E.14 and Lemma E.15 we obtain the result.

Lemma E.16. *Let P, Q be two distributions on $\mathbb{R}^d$ such that $P(\boldsymbol{x}), Q(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$ and $\ell : \mathbb{R}^d \mapsto \mathbb{R}$ be a function. Then it holds*

$$\left|\mathop{\mathbf{E}}_{\boldsymbol{x}\sim P}[\ell(\boldsymbol{x})] - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim Q}[\ell(\boldsymbol{x})]\right| \leq \left(\mathop{\mathbf{E}}_{\boldsymbol{x}\sim P}[\ell^2(\boldsymbol{x})]\mathop{\mathbf{E}}_{\boldsymbol{x}\sim P}\right)^{1/2}\left(\left[\left(\frac{Q(\boldsymbol{x})}{P(\boldsymbol{x})}\right)^2\right]\right)^{1/2}$$

Moreover,

$$\left|\mathop{\mathbf{E}}_{\boldsymbol{x}\sim P}[\ell(\boldsymbol{x})] - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim Q}[\ell(\boldsymbol{x})]\right| \leq 2\left(\left(\mathop{\mathbf{E}}_{\boldsymbol{x}\sim P}[\ell^2(\boldsymbol{x})] + \mathop{\mathbf{E}}_{\boldsymbol{x}\sim Q}[\ell^2(\boldsymbol{x})]\right)\right)^{1/2}\sqrt{d_{\mathrm{TV}}(P,Q)}$$

Proof. Write

$$\begin{aligned}\left|\mathop{\mathbf{E}}_{\boldsymbol{x}\sim P}[\ell(\boldsymbol{x})] - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim Q}[\ell(\boldsymbol{x})]\right| &\leq \int \ell(\boldsymbol{x})\sqrt{P(\boldsymbol{x})}\frac{|P(\boldsymbol{x})-Q(\boldsymbol{x})|}{\sqrt{P(\boldsymbol{x})}}\mathrm{d}x \\ &= \left(\int \ell^2(\boldsymbol{x})P(\boldsymbol{x})\mathrm{d}x \int \frac{(P(\boldsymbol{x})-Q(\boldsymbol{x}))^2}{P(\boldsymbol{x})}\mathrm{d}x\right)^{1/2}\end{aligned}$$

For the second inequality we have

$$\begin{aligned}\left|\mathop{\mathbf{E}}_{\boldsymbol{x}\sim P}[\ell(\boldsymbol{x})] - \mathop{\mathbf{E}}_{\boldsymbol{x}\sim Q}[\ell(\boldsymbol{x})]\right| &\leq \int \ell(\boldsymbol{x})|P(\boldsymbol{x})-Q(\boldsymbol{x})|\mathrm{d}x \\ &\leq \int \ell(\boldsymbol{x})\sqrt{P(\boldsymbol{x})+Q(\boldsymbol{x})}\ \frac{|P(\boldsymbol{x})-Q(\boldsymbol{x})|}{\sqrt{P(\boldsymbol{x})+Q(\boldsymbol{x})}}\ \mathrm{d}x \\ &\leq \left(\mathop{\mathbf{E}}_{x\sim P}[\ell^2(x)] + \mathop{\mathbf{E}}_{x\sim Q}[\ell^2(x)]\right)^{1/2}\left(\int \frac{(P(\boldsymbol{x})-Q(\boldsymbol{x}))^2}{P(\boldsymbol{x})+Q(\boldsymbol{x})}\mathrm{d}x\right)^{1/2}\end{aligned}$$

Now observe that

$$\left(\int \frac{(P(\boldsymbol{x}) - Q(\boldsymbol{x}))^2}{P(\boldsymbol{x}) + Q(\boldsymbol{x})} \mathrm{d}x\right)^{1/2} \leq \left(2 \int \left(\sqrt{P(\boldsymbol{x})} - \sqrt{Q(\boldsymbol{x})}\right)^2 \mathrm{d}x\right)^{1/2}$$
$$= 2d_{\mathrm{H}}(P, Q) \leq 2\sqrt{d_{\mathrm{TV}}(P, Q)}$$

□

E.6 Missing Proofs of Section 6.5

In the following we use the polynomial norms. Let $p(\boldsymbol{x}) = \sum_{V:|V|\leq k} c_V x^V$ be a multivariate polynomial. We define the $\|p\|_\infty = \max_{V:|V|\leq k} |c_V|$, $\|p\|_1 = \sum_{V:|V|\leq k} |c_V|$.

The proof of Lemma 6.31 Let $W = S_1 \cap S_2 \cap \{f_1 > f_2\} \cup S_1 \setminus S_2$, that is the set of points where the first density is larger than the second. We now write the L_1 distance between f_1, f_2 as

$$\int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} = \int \mathbb{1}_W(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$$

Denote $p(\boldsymbol{x})$ the polynomial that will do the approximation of the L_1 distance. From Lemma 6.30 we know that there exists a Normal distribution within small chi-squared divergence of both $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{r}\Sigma_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{r}\Sigma_2)$. Call the density function of this distribution $g(\boldsymbol{x})$. We have

$$\begin{aligned}
&\Big| \int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} - \int p(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})) \Big| \qquad\qquad (\text{E.13}) \\
&\qquad = \left| \int (\mathbb{1}_W(\boldsymbol{x}) - p(\boldsymbol{x}))\ (f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \right| \\
&\qquad \leq \int |\mathbb{1}_W(\boldsymbol{x}) - p(\boldsymbol{x})|\ |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \\
&\qquad \leq \int |\mathbb{1}_W(\boldsymbol{x}) - p(\boldsymbol{x})| \sqrt{g(\boldsymbol{x})}\ \frac{|f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|}{\sqrt{g(\boldsymbol{x})}} \mathrm{d}x \\
&\qquad \leq \sqrt{\int (\mathbb{1}_W(\boldsymbol{x}) - p(\boldsymbol{x}))^2 g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}} \sqrt{\int \frac{(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))^2}{g(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}}, \qquad (\text{E.14})
\end{aligned}$$

where we use Schwarzs' inequality. From Lemma 6.30 we know that

$$\int \frac{f_1(\boldsymbol{x})^2}{g(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \leq \int \frac{\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{r}\Sigma_1; \boldsymbol{x})^2}{g(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} = \exp(\mathrm{poly}(1/\alpha)).$$

Similarly, $\int \frac{f_2(\boldsymbol{x})^2}{g(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} = \exp(\mathrm{poly}(1/\alpha))$. Therefore we have,

$$\begin{aligned}
&\Big| \int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} - \int p(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})) \Big| \\
&\leq \exp(\mathrm{poly}(1/\alpha)) \sqrt{\int (\mathbb{1}_W(\boldsymbol{x}) - p(\boldsymbol{x}))^2 g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}
\end{aligned}$$

Recall that $g(\boldsymbol{x})$ is the density function of a Gaussian distribution, and let $\boldsymbol{\mu}, \boldsymbol{r}\Sigma$ be the parameters of this Gaussian. Notice that it remains to show that there exists a good approximating polynomial $p(\boldsymbol{x})$ to the indicator function $\mathbb{1}_W$. We can now transform the space so that $g(\boldsymbol{x})$ becomes the standard normal. Notice that this is an affine transformation that also transforms the set W; call the transformed set W^t. We now argue that the Gaussian surface area of the transformed set W^t at most a constant multiple of the Gaussian surface area of the original set W. Let $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{r}\Sigma_i; S_i) = \alpha_i$ for $i = 1, 2$ and let $h_1(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{r}\Sigma_1; \boldsymbol{x})/\alpha_1$ resp. $h_2(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{r}\Sigma_2; \boldsymbol{x})/\alpha_2$ be the density of first resp. second Normal ignoring the truncation sets S_1, S_2. Notice that instead of f_1, f_2 we may use h_1, h_2 in the definition of W, that is

$$W = (S_1 \cap S_2 \cap \{h_1 \geq h_2\}) \cup S_1 \setminus S_2.$$

Now, since $\boldsymbol{r}\Sigma^{-1/2} > 0$ we have that the affine map $T(x) = \boldsymbol{r}\Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ is a bijection. Therefore $T(A \cap B) = T(A) \cap T(B)$ and $T(A \cup B) = T(A) \cup T(B)$. Similarly to $W^t = T(W)$, let S_1^t, S_2^t, $\{h_1 \geq h_2\}^t$ be the transformed sets. Therefore,

$$W^t = (S_1^t \cap S_2^t \cap \{h_1 \geq h_2\}^t) \cup S_1^t \setminus S_2^t.$$

We will use some elementary properties of Gaussian surface area (see for example Fact 17 of Klivans et al. (2008)). We have that for any sets S_1, S_2 $\Gamma(S_1 \cap S_2)$ and $\Gamma(S_1 \cup S_2)$ are upper bounded from $\Gamma(S_1) + \Gamma(S_2)$. Moreover, $\Gamma(S_1 \setminus S_2) \leq \Gamma(S_1) + \Gamma(S_2^c) = \Gamma(S_1) + \Gamma(S_2)$. From our assumptions, we know that the Gaussian surface area of the sets S_1^t, S_2^t is $O(\Gamma(\mathcal{S})$. Notice now that the set $\{h_1 \geq h_2\}^t$ is a degree 2 polynomial threshold function. Therefore, using the result of Kane (2011) (see also Table 1.1) we obtain that $\Gamma(\{h_1 \geq h_2\}^t) = O(1)$. Combining the above we obtain that $\Gamma(W^t) = O(\Gamma(\mathcal{S})$. To keep the notation simple we from now on we will by W the transformed set W^t. Now, assuming that a good approximating polynomial $p(\boldsymbol{x})$ of degree k exists with respect to $\mathcal{N}(\boldsymbol{0}, \boldsymbol{r}I)$ then $p(\boldsymbol{r}\Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}))$ is a polynomial of degree k that approximates $\mathbb{1}_W(\boldsymbol{x})$ with respect to $g(\boldsymbol{x})$. Since $\mathbb{1}_W \in L^2(\mathbb{R}^d, \mathbb{N}_0)$ we can approximate it using Hermite polynomials. For some $k \in \mathbb{N}$ we set $p(\boldsymbol{x}) = S_k \mathbb{1}_W(x)$, that is

$$p_k(\boldsymbol{x}) = \sum_{V:|V| \leq k} \widehat{\mathbb{1}_W} H_V(\boldsymbol{x}).$$

Combining Lemma 6.14 and Lemma E.9 we obtain

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{N}_0}[(\mathbb{1}_W(\boldsymbol{x}) - p_k(x))^2] = O\left(\frac{\Gamma(\mathcal{S})}{k^{1/2}}\right).$$

Therefore,

$$\Big| \int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} - \int p_k(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})) \Big| = \exp(\mathrm{poly}(1/\alpha)) \frac{\Gamma(\mathcal{S})^{1/2}}{k^{1/4}}$$

Therefore, ignoring the dependence on the absolute constant α, to achieve error $O(\epsilon)$ we need degree $k = O(\Gamma(\mathcal{S})^2/\epsilon^4)$.

To complete the proof, it remains to obtain a bound for the coefficients of the polynomial $q(\boldsymbol{x}) = p_k(\boldsymbol{r}\Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}))$. We use the standard notation of polynomial norms, e.g. $\|p\|_\infty$ is the maximum (in absolute value) coefficient, $\|p\|_1$ is the sum of the absolute values of all coefficients etc. From Parseval's identity we obtain that the sum of the squared weights is less than 1 so these coefficients are clearly not large. The large coefficients are those of the Hermite Polynomials. We consider first the 1 dimensional Hermite polynomial and take an even degree Hermite polynomial H_n. The explicit formula for the k-th degree coefficient is

$$\frac{2^{k/2-n/2}\sqrt{n!}}{(n/2 - k/2)!k!} \leq 2^n,$$

see, for example, Szegö (1967). Similarly, we show the same bound when the degree of the Hermite polynomial is odd. Therefore, we have that the maximum coefficient of $H_V(x) = \prod_{i=1}^d H_i(x_i)$ is at most $\prod_{i=1}^d 2^{v_i} = 2^{\sum_{i=1}^d v_i} = 2^{|V|}$. Using Lemma E.17 we obtain that

$$\begin{aligned} \left\| H_V(\boldsymbol{r}\Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})) \right\|_1 &\leq \binom{d + |V|}{|V|} 2^{|V|} \left(\sqrt{d} \left\| \boldsymbol{r}\Sigma^{-1/2} \right\|_2 + \left\| \boldsymbol{r}\Sigma^{-1/2}\boldsymbol{\mu} \right\|_2 \right)^{|V|} \\ &\leq \binom{d + |V|}{|V|} (4d)^{|V|/2} (O(1/\alpha^2))^{|V|} \end{aligned}$$

Now we have

$$\|q(\boldsymbol{x})\|_\infty \leq \sum_{V:|V| \leq k} |c_V| \left\| H_V(\boldsymbol{r}\Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})) \right\|_\infty \leq \binom{d+k}{k}^2 (4d)^{k/2} (O(1/\alpha^2))^k,$$

where we used the fact that since $\sum_V |c_v|^2 \leq 1$ it holds that $|c_V| \leq 1$ for all V. To conclude the proof we notice that we can pick the degree k so that

$$\left|\int q(\boldsymbol{x})(f_1(\boldsymbol{x})-f_2(\boldsymbol{x}))\right| = \left|\sum_{V:|V|\le k} \boldsymbol{x}^V (f_1(\boldsymbol{x})-f_2(\boldsymbol{x}))\right| \ge \epsilon/2.$$

Since the maximum coefficient of $q(\boldsymbol{x})$ is bounded by $d^{O(k)}$ we obtain the result.

The proofof Theorem 6.5 We first draw $O(d^2/\epsilon^2)$ and compute estimates of the conditional mean $\widetilde{\boldsymbol{\mu}}_C$ and covariance $\widetilde{\boldsymbol{\Sigma}}_C$ that satisfy the guarantees of Lemma E.2. We now transform the space so that $\widetilde{\boldsymbol{\mu}}_C = \boldsymbol{0}$ and $\boldsymbol{r}\Sigma_C = \boldsymbol{r}I$. For simplicity we still denote $\boldsymbol{\mu}$ and $\boldsymbol{r}\Sigma$ the parameters of the unknown Gaussian after the transformation. From Lemma E.3 we have that $\left\|\boldsymbol{r}\Sigma^{-1/2}\boldsymbol{\mu}\right\|_2 \le O(\log(1/\alpha)^{1/2}/\alpha)$, and $\Omega(\alpha^2) \le \left\|\boldsymbol{r}\Sigma^{1/2}\right\|_2 \le O(1/\alpha^2)$. Let $\widetilde{m}_V$ be the empirical moments of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma, S)$, $\widetilde{m}_V = \frac{\sum_{i=1}^N \boldsymbol{x}^V}{N}$. We first bound the variance of a moment $\boldsymbol{x}^V$.

$$\mathrm{Var}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{r}\Sigma,S)}[\boldsymbol{x}^V] \le \mathop{\mathbf{E}}_{x\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{r}\Sigma,S)}[\boldsymbol{x}^{2V}] \le \frac{1}{\alpha}\mathop{\mathbf{E}}_{x\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{r}\Sigma)}[\boldsymbol{x}^{2V}] = \frac{1}{\alpha}\mathop{\mathbf{E}}_{x\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{r}I)}[(\boldsymbol{r}\Sigma^{1/2}\boldsymbol{x}+\boldsymbol{\mu})^{2V}]$$

Following the proof of Lemma E.17 we get that $\left\|(\boldsymbol{r}\Sigma^{1/2}\boldsymbol{x}+\boldsymbol{\mu})^{2V}\right\|_\infty \le (\sqrt{d}\left\|\boldsymbol{r}\Sigma^{1/2}\right\|_2 + \left\|\boldsymbol{\mu}\right\|_2)^{|V|}$. Using Lemma 6.31 we know that if we set $k = \Gamma(\mathcal{S})/\epsilon^4$ then given any guess of the parameters $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S}$ we can check whether the corresponding truncated Gaussian $\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$ is in total variation distance ϵ from the true by checking that all moments $\mathbf{E}_{x\sim\mathcal{N}(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}},\widetilde{S})}[\boldsymbol{x}^V]$ of the guess are close to the (estimates) of the true moments. Using the above observations and ignoring the dependence on the constant α we get that $\left\|(\boldsymbol{r}\Sigma^{1/2}\boldsymbol{x}+\boldsymbol{\mu})^{2V}\right\|_\infty \le d^{O(k)}$. Chebyshev's inequality implies that with $d^{O(k)}/\epsilon^2$ samples we can get an estimate such that with probability at least $3/4$ it holds $|\widetilde{m}_V - m_V| \le \epsilon/d^{O(k)}$. Using the standard process of repeating and taking the median estimate we amplify the success probability to $1-\delta$. Since we want all the estimates of all the moments V with $|V| \le k$ to be accurate we choose $\delta = 1/d^{O(k)}$ and by the union bound we obtain that with constant probability $|\widetilde{m}_V - m_V| \le \epsilon/d^{O(k)}$ for all V with $|V| \le k$. Now, for any tuple of parameters

$(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$ we check whether the first $d^{O(k)}$ moments of the corresponding truncated Gaussian $\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$ are in distance $\epsilon/d^{O(k)}$ of the estimates $\widetilde{m}_V$. If this is true for all the moments, then Lemma 6.31 implies that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Sigma, S), \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})) \leq \epsilon$.

The proof of Lemma 6.30 Without loss of generality we may assume that $N_1 = \mathcal{N}(\mathbf{0}, \boldsymbol{r}I)$ and $N_2 = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Lambda)$, where $\boldsymbol{r}\Lambda$ is a diagonal matrix with elements $\lambda_i > 0$. We define the normal $N = \mathcal{N}(\mathbf{0}, \boldsymbol{r}R)$ with $r_i = \max(1, \lambda_i)$. We have

$$\begin{aligned}
D_{\chi^2}(N_2 \| N) + 1 &= \int \frac{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{r}\Lambda; \boldsymbol{x})^2}{\mathcal{N}(\mathbf{0}, \boldsymbol{r}R; \boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\
&= \frac{\sqrt{|\boldsymbol{r}R|}}{(2\pi)^{d/2}|\boldsymbol{r}\Lambda|} \exp(-\boldsymbol{\mu}^T \boldsymbol{r}\Lambda^{-1}\boldsymbol{\mu}) \\
&\quad \cdot \underbrace{\int \exp\left(\boldsymbol{x}^T\left(\frac{1}{2}\boldsymbol{r}R^{-1} - \boldsymbol{r}\Lambda^{-1}\right) + 2\boldsymbol{\mu}^T\boldsymbol{r}\Lambda^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}}_{I}
\end{aligned}$$

We have

$$I = \prod_{i=1}^{d} \int \exp\left(x_i^2\left(\frac{1}{2r_i} - \frac{1}{\lambda_i}\right) + 2\frac{\mu_i}{\lambda_i}x_i\right)\mathrm{d}x_i = (2\pi)^{d/2}\prod_{i=1}^{d}\frac{\exp\left(\frac{2r_i\mu_i^2}{2r_i\lambda_i - \lambda_i^2}\right)}{\sqrt{2/\lambda_i - 1/r_i}}$$

Therefore,

$$\begin{aligned}
D_{\chi^2}(N_2 \| N) + 1 &\leq \prod_{i=1}^{d}\sqrt{\frac{r_i}{2\lambda_i - \lambda_i^2/r_i}}\exp\left(\frac{2r_i\mu_i^2}{2r_i\lambda_i - \lambda_i^2}\right) \\
&= \exp\left(\sum_{i=1}^{d}\frac{1}{2}\log\left(\frac{r_i}{2\lambda_i - \lambda_i^2/r_i}\right) + \frac{2r_i\mu_i^2}{2r_i\lambda_i - \lambda_i^2}\right)
\end{aligned}$$

Using the fact that $r_i = \max(1, \lambda_i)$ we have

$$\sum_{i=1}^{d}\log\left(\frac{r_i}{2\lambda_i - \lambda_i^2/r_i}\right) = \sum_{i:\lambda_i<1}\log\left(\frac{1}{2\lambda_i - \lambda_i^2}\right) \leq \sum_{i:\lambda_i<1}\left(\frac{1}{\lambda_i} - 1\right)^2 \leq \left\|\boldsymbol{r}\Lambda^{-1} - \boldsymbol{r}I\right\|_F^2,$$

where we used the inequality $\log(1/(2x - x^2)) \leq (1/x - 1)^2$ which holds for all $x \in (0, 1)$. Moreover,

$$\sum_{i=1}^{d} \frac{2r_i\mu_i^2}{2r_i\lambda_i - \lambda_i^2} = \sum_{i:\lambda\leq 1} \frac{2\mu_i^2}{2\lambda_i - \lambda_i^2} + \sum_{i:\lambda>1} \frac{2\mu_i^2}{\lambda_i} \leq \sum_{i=1}^{d} \frac{2\mu_i^2}{\lambda_i} = 2\left\|\boldsymbol{r}\Lambda^{-1/2}\boldsymbol{\mu}\right\|_2^2,$$

where we used the inequality $1/(2x - x^2) \leq 1/x$ which holds for all $x \in (0, 1)$. Combining the above we obtain

$$D_{\chi^2}(N_2\|N) \leq \exp\left(\frac{1}{2}\left\|\boldsymbol{r}\Lambda^{-1/2}\boldsymbol{\mu}\right\|_2 + 2\left\|\boldsymbol{r}\Lambda^{-1} - \boldsymbol{r}I\right\|_F^2\right)$$

Similarly, we compute

$$\begin{aligned} D_{\chi^2}(N_1\|N) + 1 = \prod_{i=1}^{d} \sqrt{\frac{r_i}{2 - 1/r_i}} &= \exp\left(\frac{1}{2}\sum_{i:\lambda_i>1} \log\left(\frac{\lambda_i}{2 - 1/\lambda_i}\right)\right) \\ &\leq \exp\left(\frac{1}{2}\sum_{i:\lambda_i>1} \lambda_i\left(1 - \frac{1}{\lambda_i}\right)^2\right) \\ &\leq \exp\left(\frac{1}{2}\max(\|\boldsymbol{r}\Lambda\|_2, 1)\left\|\boldsymbol{r}\Lambda^{-1} - \boldsymbol{r}I\right\|_F^2\right) \end{aligned}$$

The following lemma gives a very rough bound on the maximum coefficient of multivariate polynomials of affine transformations.

Lemma E.17. *Let $p(\boldsymbol{x}) = \sum_{V:|V|\leq k} c_V x^V$ be a multivariate polynomial of degree k. Let $\boldsymbol{r}A \in \mathbb{R}^{d\times d}$, $\boldsymbol{b} \in \mathbb{R}^d$. Let $q(\boldsymbol{x}) = p(\boldsymbol{r}A\boldsymbol{x} + \boldsymbol{b})$. Then $\|q\|_\infty \leq \|p\|_\infty \binom{d+k}{k}\left(\sqrt{d}\,\|\boldsymbol{r}A\|_2 + \|\boldsymbol{b}\|_2\right)^k$.*

Proof. We have that

$$q(\boldsymbol{x}) = \sum_{V:|V|\leq k} c_V \prod_{i=1}^{d}\left(\sum_{j=1}^{d} A_{ij}x_j + b_i\right)^{v_i}$$

Therefore,

$$\begin{aligned}
\|q\|_1 &\le \sum_{V:|V|\le k} c_V \prod_{i=1}^{d} \left(\sum_{j=1}^{d} |A_{ij}| + |b_i| \right)^{v_i} \le \sum_{V:|V|\le k} c_V \prod_{i=1}^{d} \left(\|\boldsymbol{r}A\|_\infty + \|\boldsymbol{b}\|_\infty \right)^{v_i} \\
&= \sum_{V:|V|\le k} c_V \left(\|\boldsymbol{r}A\|_\infty + \|\boldsymbol{b}\|_\infty \right)^{|V|} \le \|p\|_\infty \binom{d+k}{k} \left(\|\boldsymbol{r}A\|_\infty + \|\boldsymbol{b}\|_\infty \right)^{k} \\
&\le \|p\|_\infty \binom{d+k}{k} \left(\sqrt{d}\, \|\boldsymbol{r}A\|_2 + \|\boldsymbol{b}\|_2 \right)^{k}
\end{aligned}$$

□

F Appendix to Chapter 7

F.1 Multidimensional Taylor's Theorem

In this section we present the Taylor's theorem for multiple dimensions and we prove Theorem 7.3. We remind the following notation from the preliminaries section $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_d^{\alpha_d}$.

Theorem F.1 (Multi-Dimensional Taylor's Theorem). *Let $S \subseteq \mathbb{R}^d$ and $f : S \to \mathbb{R}$ be a $(k+1)$-times differentiable function, then for any $\boldsymbol{x}, \boldsymbol{y} \in S$ it holds that*

$$f(\boldsymbol{y}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d, |\boldsymbol{\alpha}| \leq k} \frac{\mathbf{D}_{\boldsymbol{\alpha}} f(\boldsymbol{x})}{\boldsymbol{\alpha}!} (\boldsymbol{y} - \boldsymbol{x})^{\boldsymbol{\alpha}} + H_k(\boldsymbol{y}; \boldsymbol{x}), \quad \textit{with}$$

$$H_k(\boldsymbol{y}; \boldsymbol{x}) = \sum_{\boldsymbol{\beta} \in \mathbb{N}^d, |\boldsymbol{\beta}| = k+1} R_{\boldsymbol{\beta}}(\boldsymbol{y}; \boldsymbol{x}) (\boldsymbol{y} - \boldsymbol{x})^{\boldsymbol{\beta}}$$

and

$$R_{\boldsymbol{\beta}}(\boldsymbol{y}; \boldsymbol{x}) = \frac{|\boldsymbol{\beta}|}{\boldsymbol{\beta}!} \int_0^1 (1-t)^{|\boldsymbol{\beta}|-1} \mathbf{D}_{\boldsymbol{\beta}} f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) dt.$$

We now provide a proof of Theorem 7.3.

Proof of Theorem 7.3. We start by observing that

$$\left| f(\boldsymbol{y}) - \bar{f}_k(\boldsymbol{y}; \boldsymbol{x}) \right| \leq \left(\sum_{\boldsymbol{\beta} \in \mathbb{N}^d, |\boldsymbol{\beta}| = k+1} \frac{1}{\boldsymbol{\beta}!} \right) \cdot R^{k+1} \cdot W.$$

This inequality follows from multidimensional Taylor's Theorem by some simple calculations. Now to show wanted result it suffices to show that $\sum_{\boldsymbol{\beta} \in \mathbb{N}^d, |\boldsymbol{\beta}| = k+1} \frac{1}{\boldsymbol{\beta}!} \leq \left(\frac{15d}{k} \right)^{k+1}$. To prove the latter we first show that $\min_{\boldsymbol{\beta} \in \mathbb{N}^d, |\boldsymbol{\beta}| = k+1} \boldsymbol{\beta}! = (\ell!)^{d-r}((\ell + 1)!)^r$ where $\ell = \lfloor \frac{k+1}{d} \rfloor$ and $r = k+1 \pmod{d}$. We prove this via contradiction, if this is not true then the minimum $\min_{\boldsymbol{\beta} \in \mathbb{N}^d, |\boldsymbol{\beta}| = k+1} \boldsymbol{\beta}!$ is achieved in multi-index

$\boldsymbol{\beta}$ such that there exist $i, j \in [d]$ such that $\beta_i < \ell$ and $\beta_j > \ell + 1$. In this case we define $\boldsymbol{\beta}'$ to be equal to $\boldsymbol{\beta}$ except for $\beta'_i = \beta_i + 1$ and $\beta'_j = \beta_j - 1$. In this case we get $\boldsymbol{\beta}'! < \frac{\beta_j}{\beta_i+1}\boldsymbol{\beta}'! = \boldsymbol{\beta}!$, which contradicts the optimality of $\boldsymbol{\beta}$. Therefore we have that $\sum_{\boldsymbol{\beta}\in\mathbb{N}^d, |\boldsymbol{\beta}|=k+1} \frac{1}{\boldsymbol{\beta}!} \le \binom{d+k+1}{k+1}\frac{1}{((\ell+1)!)^d}$. Now via upper bounds from Stirling's approximation we get that $\binom{d+k+1}{k+1}\frac{1}{((\ell+1)!)^d} \le \frac{e^{k+1}\left(1+\frac{d}{k+1}\right)^{k+1}}{((k+1)/d)^{k+1}e^{-k-1}} \le \left(\frac{e^2 d}{k}\right)^{k+1}\left(1+\frac{k}{k+1}\right)^{k+1}$ and the Theorem follows from simple calculations on the last expression. □

F.2 Missing Proofs for Single Dimensional Densities

In this section we provide the proof of the theorems presented in Section 7.2.

Proof of Theorem 7.7

We are going to use the following result that bounds the error of Hermite polynomial interpolation, wherein besides matching the values of the target function the approximating polynomial also matches its derivatives. The following theorem can be seen as a generalization of Lagrange interpolation, where the interpolation nodes are distinct, and Taylor's remainder theorem where we find a polynomial that matches the first k derivatives at a single node.

Lemma F.2 (Hermite Interpolation Error). *Let $x_1, \ldots, x_s$ be distinct nodes in $[a, b]$ and let $m_1, \ldots, m_s \in \mathbb{N}$ such that $\sum_{i=1}^{s} m_i = k + 1$. Moreover, let f be a $(k+1)$ times continuously differentiable function over $[a, b]$ and p be a polynomial of degree at most k such that for each x_i*

$$p(x_i) = f(x_i) \quad p'(x_i) = f'(x_i) \quad \ldots \quad p^{(m_i-1)}(x_i) = f^{(m_i-1)}(x_i)\,.$$

Then for all $x \in [a, b]$, there exists $\xi \in (a, b)$ such that

$$f(x) - p(x) = \frac{f^{(k+1)}(\xi)}{(k+1)!}\prod_{i=1}^{s}(x - x_i)^{m_i}\,.$$

We are also going to use the following upper bound on Kullback-Leibler divergence. For a proof see Lemma 1 of Barron and Sheu (1991).

Lemma F.3. *Let $\mathcal{P}, \mathcal{Q}$ be distributions on $\mathbb{R}$ with corresponding density functions p, q. Then for any $c > 0$ it holds*

$$\mathrm{D}_{\mathrm{KL}}(\mathcal{P} \| \mathcal{Q}) \leq e^{\|\log(p(x)/q(x))-c\|_\infty} \int p(x) \left(\log \frac{p(x)}{q(x)} - c \right)^2 \mathrm{d}x \,.$$

Before, the proof of Theorem 7.7 we are going to show a useful lemma. Let f, g be two density functions such that $\mathrm{D}_{\mathrm{KL}}(f \| g) > 0$ and let r another function r that lies strictly between the two densities f, g. The following lemma states that after we normalize r to become a density function $\bar{r}$ we get that $\bar{r}$ is closer to f in Kullback-Leibler divergence than g.

Lemma F.4 (Kullback-Leibler Monotonicity)**.** *Let f, g be density functions over $\mathbb{R}$ such that the measure defined by f is absolutely continuous with respect to that defined by g, i.e. the support of f is a subset of that of g. Let also r be an integrable function such that $r(x) \geq 0$, for all $x \in \mathbb{R}$, and moreover, for all $x \in \mathbb{R} \setminus Z$*

$$f(x) \leq r(x) < g(x) \quad \text{or} \quad g(x) < r(x) \leq f(x)$$

where Z is a set that has measure 0 under both f and g. Then, if $\bar{r}(x) = r(x) / \int r(x) \mathrm{d}x$ is the density function corresponding to $r(\cdot)$, it holds that

$$\mathrm{D}_{\mathrm{KL}}(f \| \bar{r}) < \mathrm{D}_{\mathrm{KL}}(f \| g).$$

Proof. To simplify notation we are going to assume that the support of f is the entire $\mathbb{R}$, and we define the sets $A_< = \{x \in \mathbb{R} : r(x) < g(x)\}$, $A_> = \{x \in \mathbb{R} : r(x) > g(x)\}$. In the following proof we are going to ignore the measure zero set where the assumptions about g, r, f do not hold. Denote $C = \int r(x) \mathrm{d}x =$

$1 - \int (g(x) - r(x)) \mathrm{d}x$. We have

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}(f \| g) - \mathrm{D}_{\mathrm{KL}}(f \| \bar{r}) &= \int f(x) \log \frac{\bar{r}(x)}{g(x)} \, \mathrm{d}x = \int f(x) \log \frac{r(x)}{g(x)} \, \mathrm{d}x - \log C \\ &\geq \int r(x) \log \frac{r(x)}{g(x)} \, \mathrm{d}x - \log C \,, \end{aligned}$$

where for the last inequality we used the fact that $f(x) < r(x)$ for all $x \in A_<$ and $f(x) > r(x)$ for all $x \in A_>$. Using the inequality $\log(1+z) \leq z$ we obtain $-\log C \geq \int (g(x) - r(x)) \mathrm{d}x$. Using this fact we obtain

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}(f \| p) - \mathrm{D}_{\mathrm{KL}}(f \| \bar{r}) &\geq \int r(x) \log \frac{r(x)}{g(x)} \, \mathrm{d}x + \int (g(x) - r(x)) \, \mathrm{d}x \\ &\geq \int \left(r(x) \log \frac{r(x)}{g(x)} + g(x) - r(x) \right) \mathrm{d}x. \end{aligned}$$

To finish the proof we observe that $r(x) \log(r(x)/g(x)) + g(x) - r(x) > 0$ for every x. To see this we rewrite the inequality as $\log \frac{r(x)}{g(x)} - 1 + \frac{g(x)}{r(x)} > 0$. To prove this we use the inequality $\log(z) - 1 + 1/z > 0$ for all $z \neq 1$. □

We are now ready to prove the main result of this section which is Theorem 7.7.

Proof of Theorem 7.7

Recall that

$$p = \operatorname*{argmin}_{q \in \mathcal{Q}_k} \mathrm{D}_{\mathrm{KL}}(D(f,S) \| D(q,S)) \,.$$

To simplify notation, we define the functions $\phi_f(x) = f(x) - \psi(f,S)$ and $\phi_p(x) = p(x) - \psi(p,S)$. Notice that these are the log densities of the conditional distributions $D(f,S)$ and $D(p,S)$ on the set S, viewed as functions over the entire interval I (i.e. they are the conditional log densities without the indicator $\mathbb{1}_S(x)$). Notice that ϕ_p is a polynomial of degree at most k. Let $g(x) = \phi_f(x) - \phi_p(x)$. We first show the following claim.

Claim F.5. *The equation $g(x) = 0$ has at least $k+1$ roots in S, counting multiplicities.*

Proof of Claim F.5. To reach a contradiction, assume that it has k (or fewer) roots. Let $\xi_1, \ldots, \xi_s \in I$ be the distinct roots of $g(x) = 0$ ordered in increasing order and let $m_1, \ldots, m_s$ be their multiplicities. Denote by $I_0, \ldots, I_s$ the partition of I using the roots of $g(x)$, that is

$$I_0 = (-\infty, \xi_1] \cap I = [\xi_0, \xi_1],\ I_1 = [\xi_1, \xi_2],\ \ldots,\ I_s = [\xi_s, \xi_{s+1}] = [\xi_s, +\infty) \cap I\,.$$

Let q be the polynomial that has the same roots as $g(x)$ and also the same sign as $g(x)$ in every set I_j of the partition. We claim that there exists $\lambda_j > 0$ such that, for every j and $x \in \mathrm{int}(I_j)$, the expression $\phi_f(x) - (\phi_p(x) + \lambda q(x))$ has the same sign as $\phi_f(x) - \phi_p(x)$ and also $|\phi_f(x) - (\phi_p(x) + \lambda_j q(x))| < |\phi_f(x) - \phi_p(x)|$. Indeed, fix an interval I_j and without loss of generality assume that $g(x) > 0$ for all $x \in I_j \setminus \{\xi_j, \xi_{j+1}\}$. Then it suffices to show that there exists $\lambda_j > 0$ such that $0 < g(x) - \lambda_j q(x) < g(x)$. Since $q(x) > 0$ for every $x \in \mathrm{int}(I_j)$, we need to choose $\lambda_j < g(x)/q(x)$. Since ξ_j is a root of the same multiplicity of both $g(x)$ and $q(x)$ and $g(x), q(x) > 0$ for all $x \in \mathrm{int}(I_j)$ we have $\lim_{x \to \xi_j^+} \frac{g(x)}{q(x)} = a > 0$. Similarly, we have $\lim_{x \to \xi_{j+1}^-} \frac{g(x)}{q(x)} = b > 0$. We can now define the following function

$$h(x) = \begin{cases} a, & x = \xi_j \\ g(x)/q(x), & \xi_j < x < \xi_{j+1} \\ b, & x = \xi_{j+1} \end{cases} .$$

We showed that $h(x)$ is continuous in $I_j = [\xi_j, \xi_{j+1}]$ and therefore has a minimum value $r_j > 0$ in the closed interval I_j. We set $\lambda_j = r_j$. With the same argument as above we obtain that for each interval I_j we can pick $\lambda_j > 0$. Since the number of intervals in our partition is finite we may set $\lambda = \min_{j=0,\ldots,s} \lambda_j$ and still have $\lambda > 0$.

We have shown that the polynomial $r(x) = \phi_p(x) + \lambda q(x)$ is almost everywhere, that is apart from a measure zero set, *strictly* closer to $\phi_f(x)$. In particular, we have that $\phi_p(x)$ for every $x \in [0,1]$, $|\phi_f(x) - r(x)| \leq |\phi_f(x) - \phi_p(x)|$ and for every $x \in S \setminus \{\xi_0, \ldots, \xi_{s+1}\}$ it holds $|\phi_f(x) - r(x)| < |\phi_f(x) - \phi_p(x)|$. Moreover, by construction we have that $\phi_f(x) - r(x)$ and $\phi_f(x) - p(x)$ are always of the same sign. Finally, the degree of $r(x)$ is at most k. Using Lemma F.4 we obtain that $\mathrm{D_{KL}}(D(f,S) \| D(r,S)) < \mathrm{D_{KL}}(D(f,S) \| D(p,S))$, which is impossible since we know that p is the polynomial that minimizes the Kullback-Leibler divergence. □

We are now ready to finish the proof of our lemma. Using Lemma F.2 and Claim F.5 we have that ϕ_p and ϕ_f are close not only in S but in the whole interval I. In particular, for every $x \in I$ it holds

$$\left| f(x) - \log \int_S e^{f(x)} \mathrm{d}x - p(x) + \log \int_S e^{p(x)} \mathrm{d}x \right| \le \frac{M}{(k+1)!} R^{k+1} := W_k \,. \tag{F.1}$$

Using the above bound together with Lemma F.3, where we set $c = \log \frac{\psi(f,I)\psi(p,S)}{\psi(p,I)\psi(f,S)}$, we obtain

$$\mathrm{D_{KL}}(D(f,I) \| D(p,I)) \le e^{W_k} W_k^2.$$

Proof of Theorem 7.12

We first show that the minimizer of the Kullback-Leibler divergence belongs to the set D_k. Let $q^* = \mathrm{argmin}_{q \in \mathcal{Q}_k} \mathrm{D_{KL}}(D(f,S) \| D(q,S))$ be the minimizer over the set of degree k polynomials. From the assumption that $f \in L_\infty(I,B)$ we have that $e^{-B}\alpha \le \int_S e^{f(x)} \mathrm{d}x \le e^B$ and therefore $|\psi(f,S)| \le B + \log(1/\alpha)$. We know from Equation (F.1) that for all $x \in I$ it holds

$$|q(x) - \psi(q,S) - f(x) + \psi(f,S)| \le W_k = \frac{M^{k+1}}{(k+1)!}.$$

In particular, for $x = 0$ using the above inequality we have $|\psi(q^*,S)| \le W_k + |\psi(f,S)| + |f(0)| \le W_k + 2B + \log(1/\alpha)$. Therefore, $q^* \in D_k$. From the Pythagorean identity of the information projection we have that for any other $q \in D_k$ it holds

$$\mathrm{D_{KL}}(D(f,S) \| D(q,S)) = \mathrm{D_{KL}}(D(f,S) \| D(q^*,S)) + \mathrm{D_{KL}}(D(q^*,S) \| D(q,S)).$$

Therefore, from the definition of q as an approximate minimizer with optimality gap ϵ we have that $\mathrm{D}_{\mathrm{KL}}(D(q^*, S) \| D(q, S)) \leq \epsilon$ and from Pinsker's inequality we obtain $d_{\mathrm{TV}}(D(q^*, S), D(q, S)) \leq \sqrt{\epsilon}$. Using the triangle inequality, we obtain

$$d_{\mathrm{TV}}(D(q, I), D(f, I)) \leq d_{\mathrm{TV}}(D(q, I), D(q^*, I)) + d_{\mathrm{TV}}(D(q^*, I), D(f, I))$$

From Theorem 7.7 and Pinsker's inequality we obtain that $d_{\mathrm{TV}}(D(q^*, I), D(f, I)) \leq e^{W_k/2} W_k$. Moreover, from Lemma 7.18 we have that $d_{\mathrm{TV}}(D(q, I), D(q^*, I)) \leq 4e^{10B}(2C/\alpha)^{k+8}\sqrt{\epsilon}$, where C is the absolute constant of Theorem 7.17.

F.3 Missing Proofs for Multi-Dimensional Densities

In this section we provide the proofs of the lemmas and theorems presented in Section 7.3.

Proof of Lemma 7.16

We remind that $\mathcal{Q}_{d,k}$ is the space of polynomials of degree at most k with d variables and zero constant term, where we might drop d from the notation if it is clear from context.

The bound on the norm $\|\mathbb{1}_K p\|_\infty$ follows directly from Taylor's theorem. In particular, using Theorem 7.3 we obtain that there exists the Taylor polynomial $f_k(\cdot; \mathbf{0})$ of degree k around $\mathbf{0}$ satisfies $\|\mathbb{1}_K (f - f_k)\|_\infty \leq (15MRd/k)^{k+1}$.

The bound on the Kullback-Leibler now follows directly from the following simple inequality that bounds the Kullback-Leibler divergence in terms of the ℓ_∞

norm of the log-densities. We have

$$\begin{aligned}
&\mathrm{D_{KL}}(D(f,S)\|D(g,S)) && \text{(F.2)}\\
&= \int_S D(f,S;\boldsymbol{x})(f(\boldsymbol{x})-g(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} + \log\left(\int_S e^{g(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right) - \log\left(\int_S e^{f(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right)\\
&\le \|\mathbb{1}_S(f-g)\|_\infty + \log\left(\int_S e^{f(\boldsymbol{x})+\|\mathbb{1}_S(f-g)\|_\infty}\mathrm{d}\boldsymbol{x}\right) - \log\left(\int_S e^{f(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right)\\
&\le 2\,\|\mathbb{1}_S(f-g)\|_\infty\,. && \text{(F.3)}
\end{aligned}$$

The polynomial provided by Theorem 7.3 does not necessarily have zero constant term. We can simply subtract this constant and show that the L_∞ norm of the resulting polynomial does not grow by a lot. Using the triangle inequality we get

$$\begin{aligned}
\|\mathbb{1}_K(f_k - f_k(\mathbf{0}))\|_\infty &\le \|\mathbb{1}_K f\|_\infty + \|\mathbb{1}_K(f_k - f)\|_\infty + |f_k(\mathbf{0})|\\
&\le 2B + M\left(\frac{15Rd}{k}\right)^{k+1}\,.
\end{aligned}$$

Finally, we observe that the polynomials f_k and $f_k - f_k(\mathbf{0})$ correspond to the same distribution after the normalization, therefore it still holds $\mathrm{D_{KL}}(D(f,S)\|D(f_k - f_k(\mathbf{0}),S)) = \mathrm{D_{KL}}(D(f,S)\|D(f_k,S)) \le 2(15MRd/k)^{k+1}$. □

Proof of Distortion of Conditioning Lemma 7.18

The Distortion of Conditioning Lemma contains two inequalities; an upper bound and a lower bound on $d_{\mathrm{TV}}(D(p,K),D(q,K))/d_{\mathrm{TV}}(D(p,K),D(q,K))$. We begin our proof with the upper bound and then we move to the lower bound.

Upper Bound. We first observe that for every set $R \subseteq K$ it holds

$$e^{-\|\mathbb{1}_K p\|_\infty}\mathrm{vol}(R) \le \int_R e^{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \le e^{\|\mathbb{1}_K p\|_\infty}\mathrm{vol}(R) \tag{F.4}$$

which implies that $|\psi(p,R)| \le \|\mathbb{1}_K p\|_\infty + \log(1/\mathrm{vol}(R))$.

Now to prove the upper bound on the ratio $d_{\mathrm{TV}}(D(p,K),D(q,K))/d_{\mathrm{TV}}(D(p,K),D(q,K))$, we will prove an upper bound on $d_{\mathrm{TV}}(D(p,K),D(q,K))$ and a lower bound on $d_{\mathrm{TV}}(D(p,K),D(q,K))$. We begin with the lower bound on $d_{\mathrm{TV}}(D(p,K),D(q,K))$.

$$\begin{aligned} 2d_{\mathrm{TV}}(D(p,S),D(q,S)) &= \int_S \left| \frac{e^{p(\boldsymbol{x})}}{e^{\psi(p,S)}} - \frac{e^{q(\boldsymbol{x})}}{e^{\psi(q,S)}} \right| \mathrm{d}\boldsymbol{x} \\ &\geq \min_{\boldsymbol{x}\in S} \left(\frac{e^{p(\boldsymbol{x})}}{e^{\psi(p,S)}} \right) \int_S \left| 1 - \frac{e^{-p(\boldsymbol{x})}}{e^{-\psi(p,S)}} \cdot \frac{e^{q(\boldsymbol{x})}}{e^{\psi(q,S)}} \right| \mathrm{d}\boldsymbol{x} \\ &\geq \frac{e^{-2\|\mathbb{1}_K p\|_\infty}}{\mathrm{vol}(S)} \int_S \left| 1 - e^{r(\boldsymbol{x})} \right| \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{F.5}$$

where $r(\boldsymbol{x}) = q(\boldsymbol{x}) - \psi(q,S) - (p(\boldsymbol{x}) - \psi(p,S))$. For some $\gamma > 0$ we define the set $Q = K \cap \{\boldsymbol{z} : |r(\boldsymbol{z})| \leq \gamma\}$. Using Theorem 7.17 for the degree k polynomial $r(\boldsymbol{x})$ and setting $q = 2k$, $\gamma = \left(\frac{\mathrm{vol}(S)}{2C\min\{d,2k\}}\right)^k \sqrt{\int_K (r(x))^2 \mathrm{d}\boldsymbol{x}}$, we get that $\mathrm{vol}\,(Q) \leq \mathrm{vol}(S)/2$. Using these definitions we have

$$\int_S |1 - e^{r(\boldsymbol{x})}| \mathrm{d}\boldsymbol{x} \geq \int_{S\setminus Q} |1 - e^{r(\boldsymbol{x})}| \mathrm{d}\boldsymbol{x} \geq \frac{\mathrm{vol}(S)}{2} \min_{\boldsymbol{x}\in S\setminus Q} |1 - e^{r(\boldsymbol{x})}| .$$

Since $|r(\boldsymbol{x})| \geq \gamma$ for all $x \in S \setminus Q$ we have that if $\gamma \geq 1$ then from the inequality $|1 - \mathrm{e}^x| \geq 1/2$ for $|x| > 1$ and from equation (F.5) we obtain

$$2d_{\mathrm{TV}}(D(p,S),D(q,S)) \geq \frac{e^{-2\|\mathbb{1}_K p\|_\infty}}{4}$$

If $\gamma < 1$ then we can use the inequality $|1 - \mathrm{e}^x| \geq |x|/2$ for $|x| \leq 1$ together with (F.5) to get

$$2d_{\mathrm{TV}}(D(p,S),D(q,S)) \geq \frac{e^{-2\|\mathbb{1}_K p\|_\infty}}{4} \mathrm{vol}(S) \cdot \gamma.$$

and hence for every value of γ we have that

$$2d_{\mathrm{TV}}(D(p,S),D(q,S)) \geq \frac{e^{-2\|\mathbb{1}_K p\|_\infty}}{4} \min\{\mathrm{vol}(S) \cdot \gamma, 1\}. \tag{F.6}$$

Next we find an upper bound on $d_{\mathrm{TV}}(D(p,K), D(q,K))$. In particular, we are going to relate the total variation distance of $D(p)$ and $D(q)$ with the integral $\int_K (r(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x}$. Applying Lemma F.3 with $c = -(\psi(q,K) - \psi(p,K)) + (\psi(q,S) - \psi(p,S))$ we have that

$$\mathrm{D}_{\mathrm{KL}}(D(p,K) \| D(q,K)) \le e^{\|\mathbb{1}_K r\|_\infty} \int_K D(p,K;\boldsymbol{x})(r(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \le e^{\|\mathbb{1}_K r\|_\infty + 2\|1_K p\|_\infty} \int_K (r(\boldsymbol{x}))^2 \,\mathrm{d}\boldsymbol{x}.$$

From equation (F.4) we obtain that $\|\mathbb{1}_K r\|_\infty \le 2\,\|\mathbb{1}_K p\|_\infty + 2\,\|\mathbb{1}_K q\|_\infty + 2\log(1/\mathrm{vol}(S))$. Now, using Pinsker’s and the above inequality we obtain

$$\begin{aligned} d_{\mathrm{TV}}(D(p,K), D(q,K)) &\le \sqrt{\mathrm{D}_{\mathrm{KL}}(D(p,K) \| D(q,K))} \le \frac{e^{2\|\mathbb{1}_K p\|_\infty + \|\mathbb{1}_K q\|_\infty}}{\mathrm{vol}(S)} \sqrt{\int_K (r(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x}} \\ &\le e^{2\|\mathbb{1}_K p\|_\infty + \|\mathbb{1}_K q\|_\infty} \frac{(2C \min\{d, 2k\})^k}{\mathrm{vol}(S)^{k+1}}\, \gamma. \end{aligned}$$

which implies our desired upper bound on the ratio $\frac{d_{\mathrm{TV}}(D(p,K),D(q,K))}{d_{\mathrm{TV}}(D(p,S),D(q,S))}$, using (F.6) and $\mathrm{vol}(S) \le 1$.

Lower Bound. We now show the lower bound on $d_{\mathrm{TV}}(D(p,K), D(q,K))/d_{\mathrm{TV}}(D(p,S), D(q,S))$. We have that

$$\begin{aligned} 2 d_{\mathrm{TV}}(D(p,S), D(q,S)) &= \int 1_S(\boldsymbol{x}) \left| \frac{D(p,K;\boldsymbol{x})}{D(p,K;S)} - \frac{D(q,K;\boldsymbol{x})}{D(q,K;S)} \right| \mathrm{d}\boldsymbol{x} \\ &= \int 1_S(\boldsymbol{x}) \left| \frac{D(p,K;\boldsymbol{x})}{D(p,K;S)} - \frac{D(q,K;\boldsymbol{x})}{D(p,K;S)} + \frac{D(q,K;\boldsymbol{x})}{D(p,K;S)} - \frac{D(q,K;\boldsymbol{x})}{D(q,K;S)} \right| \mathrm{d}\boldsymbol{x} \\ &\le \frac{1}{D(p,K;S)} \int 1_S(\boldsymbol{x}) \left| D(p,K;\boldsymbol{x}) - D(q,K;\boldsymbol{x}) \right| \mathrm{d}\boldsymbol{x} \\ &+ \int \mathbb{1}_S(\boldsymbol{x}) \left| \frac{D(q,K;\boldsymbol{x})}{D(p,K;S)} - \frac{D(q,K;\boldsymbol{x})}{D(q,K;S)} \right| \mathrm{d}\boldsymbol{x} \\ &\le \frac{1}{D(p,K;S)} d_{\mathrm{TV}}(D(p,K), D(q,K)) + \left| \frac{D(q,K;S)}{D(p,K;S)} - 1 \right| \\ &\le \frac{2}{D(p,K;S)} d_{\mathrm{TV}}(D(p,K), D(q,K))\,, \end{aligned}$$

where for the last step we used the fact that $|D(p,K;S) - D(q,K;S)| \le d_{\mathrm{TV}}(D(p,K), D(q,K))$.

Using again equation (F.4) we obtain that

$$D(p,K;S) = e^{\psi(p,S)-\psi(p,K)} = \left(\int_S e^{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right)\cdot\left(\int_K e^{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right)^{-1}$$

$$\geq \frac{e^{-\|\mathbb{1}_K p\|_\infty}\mathrm{vol}(S)}{e^{\|\mathbb{1}_K p\|_\infty}\mathrm{vol}(K)} \geq e^{-2B}\frac{\mathrm{vol}(S)}{\mathrm{vol}(K)},$$

and since $\mathrm{vol}(K) = 1$ the wanted bound of the lemma follows. □

Proof of Theorem 7.8

From Lemma 7.16 we obtain that by choosing $\mathbf{0} \in K$, there exists $\boldsymbol{v}$ such that $\left\|\mathbb{1}_K \boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})\right\|_\infty \leq 2B + (15Md/k)^{k+1}$. Moreover, from the same lemma we have that $\min_{\boldsymbol{w}\in D} \mathrm{D_{KL}}(D(f,S)\|D(\boldsymbol{w},S)) \leq \mathrm{D_{KL}}(D(f,S)\|D(\boldsymbol{v},S)) \leq 2\,(15Md/k)^{k+1}$. To simplify notation set $r_k = 2\,(15Md/k)^{k+1}$. Now, let $q(\boldsymbol{x}) = \boldsymbol{u}^T\boldsymbol{m}_k(\boldsymbol{x})$ be any approximate minimizer in D of the KL-divergence between $D(q,S)$ and $D(f,S)$ that satisfies

$$\mathrm{D_{KL}}(D(f,S)\|D(\boldsymbol{u},S)) \leq \min_{\boldsymbol{w}\in D}\mathrm{D_{KL}}(D(f,S)\|D(\boldsymbol{w},S)) + \epsilon \leq r_k + \epsilon.$$

Using the triangle inequality, we have

$$d_{\mathrm{TV}}(D(f,K),D(\boldsymbol{u},K)) \leq d_{\mathrm{TV}}(D(f,K),D(\boldsymbol{v},K)) + d_{\mathrm{TV}}(D(\boldsymbol{v},K),D(\boldsymbol{u},K)).$$

Using Lemma 7.18 we obtain that $d_{\mathrm{TV}}(D(\boldsymbol{v},K),D(\boldsymbol{u},K)) \leq U d_{\mathrm{TV}}(D(\boldsymbol{v},S),D(\boldsymbol{u},S))$ where $U = 4e^{15B}(2Cd)^k/\alpha^{k+3}$. Using again the triangle inequality we obtain that

$$d_{\mathrm{TV}}(D(\boldsymbol{v},K),D(\boldsymbol{u},K)) \leq U\left(d_{\mathrm{TV}}(D(\boldsymbol{v},S),D(f,S)) + d_{\mathrm{TV}}(D(f,S),D(\boldsymbol{u},S))\right)$$

Overall, we have proved the following important inequality that shows that we can extend the conditional information to whole set K without increasing the error by a lot. In other words, the polynomial with parameters $\boldsymbol{u}$ that we found by (approximately) minimizing the Kullback-Leibler divergence to the *conditional* *distribution* is a good approximation on the whole convex set K.

$$\begin{aligned}&d_{\mathrm{TV}}(D(f,K),D(\boldsymbol{u},K))\\&\le d_{\mathrm{TV}}(D(\boldsymbol{v},K),D(f,K))+U\left(d_{\mathrm{TV}}(D(\boldsymbol{v},S),D(f,S))+d_{\mathrm{TV}}(D(f,S),D(\boldsymbol{u},S))\right)\\&\le(U+1)(2\sqrt{r_k}+\sqrt{\bar{\epsilon}})\,.\end{aligned}$$

where we set $\bar{\epsilon}=2^{-\widetilde{\Omega}\left(\frac{d^3M}{\alpha^2}+B\right)}(1/\epsilon)^{-\Omega(\log(d/\alpha))}$ is the optimality gap of the vector $\boldsymbol{u}$. For the last inequality we use Pinsker's inequality to get that $d_{\mathrm{TV}}(D(\boldsymbol{v},S),D(f,S))\le\sqrt{\mathrm{D}_{\mathrm{KL}}(D(\boldsymbol{v},S)\|D(f,S))}\le\sqrt{r_k}$ since the guarantee of Lemma 7.16 holds for every subset $R\subseteq K$. Using again Pinsker's inequality to upper bound the other two total variation distances we obtain the final inequality. Substituting the values of U,L we obtain that

$$\begin{aligned}(U+1)(2\sqrt{r_k})&=O\left(e^{15B}\frac{(2Cd)^k}{\alpha^{k+3}}\frac{(15Md)^{k/2+1/2}}{k^{k/2}}\right)\\&=O\left(\exp\left(15B+\log\left(\frac{\sqrt{Md}}{\alpha^3}\right)+k\log\left(C'\frac{\sqrt{d^3M}}{\alpha}\right)-k\log\sqrt{k}\right)\right),\end{aligned}$$

where C' is an absolute constant. Therefore, for $k=O(d^3M/\alpha^2+B)+2\log(1/\epsilon)$ it holds that $U(2\sqrt{r_k})\le\epsilon/2$. Moreover, we observe that $U\sqrt{\bar{\epsilon}}\le\epsilon/2$ and therefore for this value of k we have $d_{\mathrm{TV}}(D(f,K),D(\boldsymbol{u},K))\le\epsilon/2+\epsilon/2\le\epsilon$. □

Proof of Theorem 7.20

We start with two lemmas that we are going to use in our proof of Theorem 7.20.

Lemma F.6 (Theorem 46 of Ben-David et al. (2018))**.** *Let p be a polynomial with real coefficients on d variables with some degree k such that $p\in L_\infty([0,1]^d,B)$. Then, the magnitude of any coefficient of p is at most $B(2k)^{3k}$ and the sum of magnitudes of all coefficients of p is at most $B\min((2(d+k))^{3k},2^{O(dk)})$.*

We note that in Ben-David et al. (2018) the $(2(d+k))^{3k}$ upper bound is given, the other follows easily from the single dimensional bound $2^{O(k)}$.

Lemma F.7 (Renegar (1992a,b))**.** *Let $p_i : \mathbb{R}^d \mapsto \mathbb{R}, i \in [m]$ be m polynomials over the reals each of degree at most k. Let $K = \{\boldsymbol{x} \in \mathbb{R}^d : p_i(\boldsymbol{x}) \geq 0, \text{for all } i \in [m]\}$. If the coefficients of the p_i's are rational numbers with bit complexity at most L, there is an algorithm that runs in time* $\operatorname{poly}(L, (mk)^d)$ *and decides if K is empty or not. Furthermore, if K is non-empty,the algorithm runs in time* $\operatorname{poly}(L, (mk)^d, \log(1/\delta))$ *and outputs a point in K up to an L_2 error δ.*

Objective Function of MLE

Now we define our objective, which is the Kullback-Leibler divergence between f and the candidate distribution, or equivalently the maximum-likelihood objective.

$$\begin{aligned} L(\boldsymbol{v}) &= \mathrm{D}_{\mathrm{KL}}(D(f,S) \| D(\boldsymbol{v},S)) \qquad \text{(F.7)} \\ &= \int D(f,S;\boldsymbol{x}) \log D(f,S;\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int D(f,S;\boldsymbol{x}) \boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \log \int_S e^{\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \end{aligned}$$

The gradient of $L(\boldsymbol{v})$ with respect to $\boldsymbol{v}$ is

$$\begin{aligned} \nabla_{\boldsymbol{v}} L(\boldsymbol{v}) &= -\int D(f,S;\boldsymbol{x}) \boldsymbol{m}_k(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \frac{\int_S \boldsymbol{m}_k(\boldsymbol{x}) e^{\boldsymbol{a}^T \boldsymbol{m}_k(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}}{\int_S e^{\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}} \\ &= \mathop{\mathbf{E}}_{x \sim D(\boldsymbol{v},S)} [\boldsymbol{m}_k(\boldsymbol{x})] - \mathop{\mathbf{E}}_{x \sim D(f,S)} [\boldsymbol{m}_k(\boldsymbol{x})] \qquad \text{(F.8)} \end{aligned}$$

The Hessian of $L(\boldsymbol{a})$ with respect to $\boldsymbol{v}$ is

$$\begin{aligned} \nabla^2_{\boldsymbol{v}} L(\boldsymbol{v}) &= \frac{\int_S \boldsymbol{m}_k(\boldsymbol{x}) \boldsymbol{m}_k^T(\boldsymbol{x}) e^{\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}}{\int_S e^{\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})}} - \frac{\int_S \boldsymbol{m}_k(\boldsymbol{x}) \boldsymbol{m}_k^T(\boldsymbol{x}) e^{\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}}{\left(\int_S e^{\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})}\right)^2} \\ &= \mathop{\mathbf{E}}_{x \sim D(\boldsymbol{v},S)} [\boldsymbol{m}_k(\boldsymbol{x}) \boldsymbol{m}_k^T(\boldsymbol{x})] - \mathop{\mathbf{E}}_{x \sim D(\boldsymbol{v},S)} [\boldsymbol{m}_k(\boldsymbol{x})] \mathop{\mathbf{E}}_{x \sim D(\boldsymbol{v},S)} [\boldsymbol{m}_k^T(\boldsymbol{x})]. \qquad \text{(F.9)} \end{aligned}$$

We observe that the Hessian is positive semi-definite since it is the covariance of the vector $\boldsymbol{m}_k(\boldsymbol{x})$. Therefore, we verify that $L(\boldsymbol{v})$ is convex as a function of $\boldsymbol{v}$.

Convergence of PSGD

Now, we prove that using Algorithm 13 we can efficiently estimate the parameters of a polynomial whose density well approximates the unknown density $D(f, [0,1]^d)$ in the whole unit cube.

We want to optimize the function $L(\boldsymbol{v})$ of Equation F.7 constrained in the convex set

$$D = \left\{ \boldsymbol{v} \in \mathbb{R}^m \;:\; \left\| \mathbb{1}_K \boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x}) \right\|_\infty \le C \right\}.$$

To be able to perform SGD we need to have unbiased estimates of the gradients. In particular, from the expression of the gradient (see Equation F.8) we have that in order to have unbiased estimates we need to generate a sample from the distribution $D(\boldsymbol{v}, S)$. We first observe that the initialization $\boldsymbol{v}^{(0)} \in D$. Using rejection sampling we can generate with probability at least $1 - \delta$ a sample uniformly distributed on S after $\log(1/\delta)/\alpha$ draws from the uniform distribution on K. Using the samples distributed uniformly over S we can use again rejection sampling to create a sample from $D(\boldsymbol{v}, S)$ using as base density the uniform over S. Since $e^{\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})} \le e^C$, the acceptance probability is e^{-C}. We need to generate $e^C \log(1/\delta)$ samples from the uniform on S in order to generate one sample from $D(\boldsymbol{v}, S)$. Overall, the total samples from the uniform on K in order to generate a sample from $D(\boldsymbol{v}, S)$ with probability $1 - \delta$ is $O(e^C C \log(1/\delta))$. To generate an unbiased estimate of the gradient we can simply draw samples $\boldsymbol{x}_t \sim D(\boldsymbol{v}, S), \boldsymbol{y}_t \sim D(f, S)$ and then take their difference, i.e. $\boldsymbol{g}^{(t)} = \boldsymbol{m}_k(\boldsymbol{x}^{(t)}) - \boldsymbol{m}_k(\boldsymbol{y}^{(t)})$. We have $\left\| \boldsymbol{g}^{(t)} \right\|_2^2 \le 2\binom{d+k}{k}$ for any $\boldsymbol{x} \in K$. Moreover, we need a bound on the L_2 diameter of D. From Lemma F.6 we have that since $\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x}) \in L_\infty(K, C)$ we get that $\|\boldsymbol{v}\|_2 \le \|\boldsymbol{v}\|_1 \le C(2(d+k))^k$. Now, we have all the ingredients to use Lemma 7.21, and obtain that after

$$T = \frac{C^2 2^{O(dk)} \cdot \binom{d+k}{k}}{\epsilon^2} = \frac{C^2 2^{O(dk)}}{\epsilon^2}$$

rounds, we have a vector $\boldsymbol{v}^{(T)}$ with optimality gap ϵ.

We next describe an efficient way to project to the convex set D. The projection to D is defined as $\operatorname{argmin}_{\boldsymbol{u} \in D} \|\boldsymbol{u} - \boldsymbol{v}\|_2^2$. We can use the Ellipsoid algorithm (see for example Martin Grötschel (1993)) to optimize the above convex objective as long as we can implement a separation oracle for the set D. The set D has an infinite number of linear constraints (one constraint for each $\boldsymbol{x} \in K$ but we can still use Renegar's algorithm to find a violated constraint for a point $\boldsymbol{v} \notin D$. Specifically, given a guess $\boldsymbol{v}$ we set up the following system of polynomial inequalities,

$$\begin{aligned}
&\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x}) \geq C \\
&0 \leq x_i \leq 1 \text{ for all } i \in [d] \,,
\end{aligned}$$

where $\boldsymbol{x}$ is the variable. Using Lemma F.7 we can decide if the above system is infeasible or find $\boldsymbol{x}$ that satisfies $\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x}) \geq C$ in time $\operatorname{poly}(((d+1)k)^d)$, where we suppress the dependence on the accuracy and bit complexity parameters.[1] If Renegar's algorithm returns such an $\boldsymbol{x}$ we have a violated constraint of D. Since D bounds the absolute value of $\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})$ we need to run Renegar's algorithm also for the system $\{\boldsymbol{x} : \boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x}) \leq -C, \boldsymbol{x} \in K\}$. The overall runtime of our separation oracle is $\operatorname{poly}(((d+1)k)^d)$ and thus the runtime of Ellipsoid to implement the projection step is also of the same order. Combining the runtime for sampling, the projection, and the total number of rounds we obtain that the total runtime of our algorithm is $2^{O(dk+C)}/(\alpha\epsilon^2)$. □

[1] Since the dependence of Renegar's algorithm is polynomial in the bit size of the coefficients and the accuracy of the solution it is straightforward to do the analysis of our algorithm assuming finite precision rational numbers instead of reals.

Algorithm 13 Projected Stochastic Gradient Descent. Given access to samples from $D(f, S)$.

```
procedure SGD($T, \eta, C$)    ▷ $T$: number of steps, $\eta$: step size, $C$: projection parameter.
    $\boldsymbol{v}^{(0)} \leftarrow \mathbf{0}$
    Let $D = \{\boldsymbol{v} : \max_{\boldsymbol{x} \in K} |\boldsymbol{v}^T \boldsymbol{m}_k(\boldsymbol{x})| \leq C\}$
    for $t = 1, \ldots, T$ do
        Draw sample $\boldsymbol{x}^{(t)} \sim D(\boldsymbol{v}^{(t-1)}, S)$ and $\boldsymbol{y}^{(t)} \sim D(f, S)$
        $\boldsymbol{g}^{(t)} \leftarrow \boldsymbol{m}_k(\boldsymbol{x}^{(t)}) - \boldsymbol{m}_k(\boldsymbol{y}^{(t)})$
        $\boldsymbol{v}^{(t)} \leftarrow \text{proj}_D(\boldsymbol{v}^{(t-1)} - \eta \boldsymbol{g}^{(t)})$
    return
```

G APPENDIX TO CHAPTER 8

G.1 Training Models from Coarse Data

Consider a parameterized family of functions $\boldsymbol{x} \to f(\boldsymbol{x}; \boldsymbol{w})$, where the parameters $\boldsymbol{w}$ lie in some parameter space $\mathcal{W} \subseteq \mathbb{R}^p$. For instance, the family may correspond to a feed-forward neural network with L layers. Given a finely labeled training sample $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N) \in \mathcal{X} \times \mathcal{Y}$, the parameters $\boldsymbol{w}$ are chosen using a gradient method in order to minimize the empirical risk,

$$\mathcal{L}_N(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} \ell(f(\boldsymbol{x}_i; \boldsymbol{w}), y_i)\,,$$

for some loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and the goal of this optimization task is to minimize the population risk function $\mathcal{L}(\boldsymbol{w}) = \mathbf{E}_{(\boldsymbol{x},y) \sim D(\boldsymbol{w}^\star)}[\ell(f(\boldsymbol{x}; \boldsymbol{w}), y)]$ (where the distribution $D(\boldsymbol{w}^\star)$ is unknown). For simplicity, let us focus on differentiable loss functions. Performing the SGD algorithm, we can circumvent the lack of knowledge of the population risk function $\mathcal{L}$. Specifically, instead of computing the gradient of $\mathcal{L}(\boldsymbol{w})$, the algorithm steps towards a random direction $\boldsymbol{v}$ with the constraint that the expected value of $\boldsymbol{v}$ is equal to the negative of the true gradient, i.e., it is an unbiased estimate of $-\nabla \mathcal{L}(\boldsymbol{w})$. Such a random vector $\boldsymbol{v}$ can be computed without knowing $D(\boldsymbol{w}^\star)$ using the interchangeability between the expectation and the gradient operators. Assume that the algorithm is at iteration $t \geq 1$. Let $(\boldsymbol{x}, y) \sim D(\boldsymbol{w}^\star)$ be a fresh sample and define $\boldsymbol{v}_t$ be the gradient of the loss function with respect to $\boldsymbol{w}$, at the point $\boldsymbol{w}_t$, i.e.,

$$\mathbf{E}[\boldsymbol{v}_t | \boldsymbol{w}_t] = \mathop{\mathbf{E}}_{(\boldsymbol{x},y) \sim D(\boldsymbol{w}^\star)} [\nabla \ell(f(\boldsymbol{x}; \boldsymbol{w}_t), y)] = \nabla \mathop{\mathbf{E}}_{(\boldsymbol{x},y) \sim D(\boldsymbol{w}^\star)} [\ell(f(\boldsymbol{x}; \boldsymbol{w}_t), y)] = \nabla \mathcal{L}(\boldsymbol{w}_t)\,.$$

Hence, an algorithm that has query access to a SQ oracle can implement a noisy version of the above iterative process (with inexact gradients, see e.g., d'Aspremont (2008); Devolder et al. (2014); Feldman et al. (2015a)) using the query functions

$q_i(\boldsymbol{x}, y) = (\nabla \ell(f(\boldsymbol{x}; \boldsymbol{w}_t), y))_i$ for any $i \in [p]$. Note that the algorithm knows the loss function ℓ, the parameterized functions' family $\{f(\cdot\,; \boldsymbol{w}) : \boldsymbol{w} \in \mathcal{W}\}$ and the current guess $\boldsymbol{w}_t$. Specifically, the algorithm performs p queries (one for each coordinate of the parameter vector) and the oracle returns to the algorithm a noisy gradient vector $\boldsymbol{r}_t$ that satisfies $\|\boldsymbol{r}_t - \nabla \mathcal{L}(\boldsymbol{w}_t)\|_\infty \leq \tau$.

In our setting, we do not have access to the SQ oracle with finely labeled examples. Our main result of this section (Theorem 8.4) is a mechanism that enables us to obtain access to such an oracle using a few coarsely labeled examples (with high probability). Hence, we can still perform the noisy gradient descent of the previous paragraph with an additional overhead on the sample complexity, due to the reduction.

G.2 Multiclass Logistic Regression with Coarsely Labeled Data

A first application for the above generic reduction from coarse data to statistical queries is the case of coarse multiclass logistic regression. In the standard (finely labeled) multiclass logistic regression problem, there are k fine labels (that correspond to classes), each one associated with a weight vector $\boldsymbol{w}_z \in \mathbb{R}^n$ with $z \in [k]$. We can consider the weight matrix $\boldsymbol{W} \in \mathbb{R}^{k \times n}$. Given an example $\boldsymbol{x} \in \mathbb{R}^n$, the vector $\boldsymbol{x}$ is filtered via the softmax function $\sigma(\boldsymbol{W}, \boldsymbol{x})$, which is a probability distribution over Δ^k with $\sigma(\boldsymbol{W}, \boldsymbol{x}; z) = \exp(\boldsymbol{w}_z^T \boldsymbol{x}) / \sum_{y \in [k]} \exp(\boldsymbol{w}_y^T \boldsymbol{x}), z \in [k]$ and the output is the finely labeled example $(\boldsymbol{x}, z) \in \mathbb{R}^n \times [k]$. The goal is to estimate the weight matrix $\boldsymbol{W}$, given finely labeled examples. Let us denote by $D(\boldsymbol{W})$ the joint distribution over the finely labeled examples for simplicity. When we have access to finely labeled examples $(\boldsymbol{x}, z) \sim D(\boldsymbol{W}^\star)$, the population log-likelihood objective $\mathcal{L}$ of the multiclass logistic regression problem

$$\mathcal{L}(\boldsymbol{W}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},z) \sim D(\boldsymbol{W}^\star)} \left[\boldsymbol{w}_z^T \boldsymbol{x} - \log \Big(\sum_{j \in \mathcal{Z}} \exp(\boldsymbol{w}_j^T \boldsymbol{x}) \Big) \right],$$

is concave (see Friedman et al. (2001)) with respect to the weight matrix $\boldsymbol{W} \in \mathbb{R}^{k \times n}$ and is solved using gradient methods. On the other hand, if we have sample access only to coarsely labeled examples $(\boldsymbol{x}, S) \sim D_\pi(\boldsymbol{W}^\star)$, the population log-likelihood objective $\mathcal{L}_\pi$ of the coarse multiclass logistic regression problem

$$\mathcal{L}_\pi(\boldsymbol{W}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},S)\sim D_\pi(\boldsymbol{W}^\star)} \left[\log \Big(\sum_{z \in S} \exp(\boldsymbol{w}_z^T \boldsymbol{x}) \Big) - \log \Big(\sum_{j \in \mathcal{Z}} \exp(\boldsymbol{w}_j^T \boldsymbol{x}) \Big) \right],$$

which is no more concave. However, as an application of our main result (Theorem 8.4), we can still solve it. In fact, since we can implement statistical queries using the sample access to the coarse data generative process $D_\pi(\boldsymbol{W}^\star)$, we can compute the gradients of the log-likelihood objective that corresponds to the *finely labeled examples*. Hence, the total sample complexity of optimizing this non-convex objective is equal to the sample complexity of solving the convex problem with an additional overhead at each iteration of computing the gradients, that is given by Theorem 8.4.

G.3 Geometric Information Preservation

In this section, we aim to provide some intuition behind the notion of information preserving partitions. The following result provides a geometric property for the partition distribution π. We show that if the partition distribution satisfies this particular geometric property, then it is also information preserving. We underline that the geometric property is quite important for our better understanding and it has the advantage that it is easy to verify. Hence, while the notion of information preserving distributions may be less intuitive, we believe that the geometric preservation property that we state in Lemma G.1 can fulfill this lack of intuition. The property informally states that, for any hyperplane, the sets in the partition that are not cut by this hyperplane have non trivial probability mass with respect to the true Gaussian. In the case of mixtures of convex partitions, we would like the same property to hold in expectation.

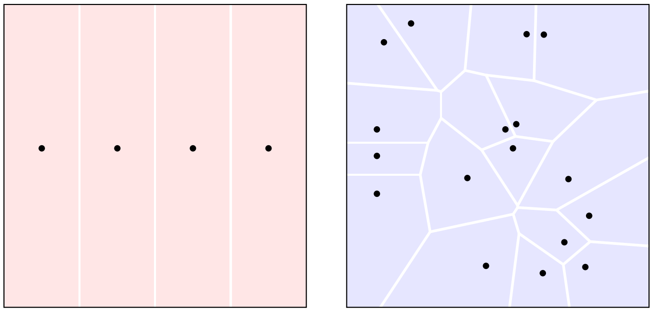


Figure G.1: (a) is a very rough partition that makes learning the mean impossible: Gaussians $\mathcal{N}((0,z))$ centered along the same vertical line $(0,z)$ assign exactly the same probability to all cells of the partitions and therefore, $d_{\mathrm{TV}}(\mathcal{N}_\pi((0,z_1)),\mathcal{N}_\pi((0,z_2)))=0$: it is impossible to learn the second coordinate of the mean. (b) is a convex partition of $\mathbb{R}^2$, that makes recovering the Gaussian possible.

Before stating Lemma G.1, let us return to Figure G.1. Observe that, in the first example with the four halfspaces, the geometric property does not hold, since there exists a line (i.e., a hyperplane) that intersects with all the sets. On the other hand, if we consider the second example with the Voronoi partition and assume that the true mean lies in the middle of the picture, we can see that any hyperplane does not intersect with a sufficient number of sets and, hence, the union of the uncut sets has non trivial probability mass for any hyperplane.

For a hyperplane $\mathcal{H}_{\boldsymbol{w},c}=\{\boldsymbol{x}\in\mathbb{R}^d:\boldsymbol{w}^T\boldsymbol{x}=c\}$ with normal vector $\boldsymbol{w}\in\mathbb{R}^d$ and threshold $c\in\mathbb{R}$, we denote the two associated halfspaces by $\mathcal{H}^+_{\boldsymbol{w},c}=\{\boldsymbol{x}\in\mathbb{R}^d:\boldsymbol{w}^T\boldsymbol{x}>c\}$ and $\mathcal{H}^-_{\boldsymbol{w},c}=\{\boldsymbol{x}\in\mathbb{R}^d:\boldsymbol{w}^T\boldsymbol{x}<c\}$. Before stating the next Lemma, we shortly describe what means for a hyperplane to cut a set with respect to a Gaussian $\mathcal{N}$. The set S is not cut by the hyperplane $\mathcal{H}$, if it totally lies in a halfspace induced by the hyperplane, say $\mathcal{H}^+$, i.e., it holds that $\mathcal{N}(S)=\mathcal{N}(S\cap\mathcal{H}^+)$.

Lemma G.1 (Geometric Information Preservation)**.** *Consider the generative process of coarse d-dimensional Gaussian data $\mathcal{N}_\pi(\boldsymbol{\mu}^\star)$, (see Definition 8.5). Consider an arbitrary hyperplane $\mathcal{H}_{\boldsymbol{w},c}$ with normal vector $\boldsymbol{w}\in\mathbb{R}^d$ and threshold $c\in\mathbb{R}$. For a partition $\boldsymbol{\Sigma}\in\operatorname{supp}(\pi)$ of $\mathbb{R}^d$, consider the collection that contains all the sets that are not cut by* *the hyperplane $\mathcal{H}_{\boldsymbol{w},c}$, i.e.,*

$$U_{\boldsymbol{w},c,\boldsymbol{\Sigma}} = \bigcup \left\{ S \in \boldsymbol{\Sigma} : \mathcal{N}^\star(S \cap \mathcal{H}^+_{\boldsymbol{w},c}) = \mathcal{N}^\star(S) \vee \mathcal{N}^\star(S \cap \mathcal{H}^-_{\boldsymbol{w},c}) = \mathcal{N}^\star(S) \right\}.$$

Assume that π satisfies

$$\mathop{\mathbf{E}}_{\boldsymbol{\Sigma} \sim \pi} \left[\mathcal{N}(\boldsymbol{\mu}^\star; U_{\boldsymbol{w},c,\boldsymbol{\Sigma}}) \right] \geq \alpha \,, \tag{G.1}$$

for some $\alpha \in (0,1]$. Then, for any Gaussian distribution $\mathcal{N}(\boldsymbol{\mu})$, it holds that

$$d_{\mathrm{TV}}(\mathcal{N}_\pi(\boldsymbol{\mu}), \mathcal{N}_\pi(\boldsymbol{\mu}^\star)) \geq C_\alpha \cdot d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}), \mathcal{N}(\boldsymbol{\mu}^\star)) \,,$$

for some C_α that depends only on α and satisfies $C_\alpha = \mathrm{poly}(\alpha)$, i.e., the partition distribution is C_α-information preserving.

Hence, the above geometric property is sufficient for information preservation. If we assume that the total variation distance between the true Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^\star)$ and a possible guess $\mathcal{N}(\boldsymbol{\mu})$ is at least ϵ and the partition distribution satisfies the geometric property of Equation (G.1), we get that the coarse generative process preserves a sufficiently large gap, in the sense that $d_{TV}(\mathcal{N}_\pi(\boldsymbol{\mu}^\star), \mathcal{N}_\pi(\boldsymbol{\mu})) \geq \mathrm{poly}(\alpha)\epsilon$. The proof of the above lemma, which relies on high-dimensional anti-concentration results on Gaussian distributions, follows.

Proof of Lemma G.1. Let us denote the true distribution by $\mathcal{N}^\star = \mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{I})$ for short. Consider an arbitrary hyperplane $\mathcal{H}_{\boldsymbol{w},c}$ with normal vector $\boldsymbol{w} \in \mathbb{R}^d$ and threshold $c \in \mathbb{R}$. Since the partition distribution (supported on a family of partitions $\mathcal{B}$) satisfies Equation (G.1), we have that, for the random variable $\mathcal{N}^\star(U_{\boldsymbol{w},c,\boldsymbol{\Sigma}})$, that takes values in $[0,1]$, there exists α such that

$$\mathop{\mathbf{E}}_{\boldsymbol{\Sigma} \sim \pi} \left[\mathcal{N}^\star(U_{\boldsymbol{w},c,\boldsymbol{\Sigma}}) \right] = \alpha \,.$$

We will use the following simple Markov-type inequality for bounded random variables.

Fact G.2 (Lemma B.1 from Shalev-Shwartz and Ben-David (2014a))**.** *Let Z be a random variable that takes values in $[0,1]$. Then, for any $\alpha \in (0,1)$, it holds that*

$$\mathbf{Pr}[Z > \alpha] \geq \frac{\mathbf{E}[Z] - \alpha}{1 - \alpha} \geq \mathbf{E}[Z] - \alpha\,.$$

By the Fact G.2, it holds that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\Sigma} \sim \pi} \left[\mathcal{N}^\star(U_{\boldsymbol{w},c,\boldsymbol{\Sigma}}) \geq \alpha/2 \right] \geq \alpha/2\,.$$

Hence, the mass of the "good" partitions is at least $\alpha/2$. Fix such a partition $\boldsymbol{\Sigma} \in \mathcal{B}$ (in the support of the partition distribution) and consider the true $\mathcal{N}^\star = \mathcal{N}(\boldsymbol{\mu}^\star)$ and the guess $\mathcal{N} = \mathcal{N}(\boldsymbol{\mu})$ distributions. For this pair of distributions, consider the set

$$\mathcal{H} = \left\{ \boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x}^T(\boldsymbol{\mu} - \boldsymbol{\mu}^\star) = \left(\|\boldsymbol{\mu}\|_2^2 - \|\boldsymbol{\mu}^\star\|_2^2 \right)/2 \right\}.$$

Observe that this set is a hyperplane with normal vector $\boldsymbol{\mu}^\star - \boldsymbol{\mu}$, that contains the midpoint $\frac{1}{2}(\boldsymbol{\mu} + \boldsymbol{\mu}^\star)$ (see Figure G.2).

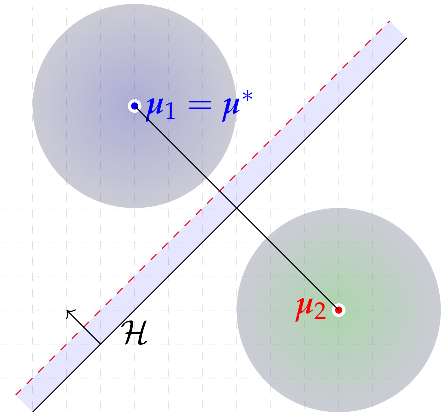


Figure G.2: Illustration of the worst-case set in testing the hypotheses $h_1 = \{\boldsymbol{\mu}_1 = \boldsymbol{\mu}^\star\}$ and $h_2 = \{\boldsymbol{\mu}_2 = \boldsymbol{\mu}^\star\}$.

Our main focus is to lower bound the total variation distance of the coarse distributions $\mathcal{N}^\star_\pi$ and $\mathcal{N}_\pi$. We claim that this lower bound can be described as a fractional knapsack problem and, hence, it is attained by a worst-case set, that (intuitively) places points as close as possible to the hyperplane $\mathcal{H}$, until its mass with respect to the true Gaussian $\mathcal{N}^\star$ is at least $\alpha/2$. Recall that the total variation distance between the two coarse distributions is

$$d_{\mathrm{TV}}(\mathcal{N}_\pi, \mathcal{N}_\pi^\star) = \sum_{\mathbf{\Sigma} \in \mathcal{B}} \pi(\mathbf{\Sigma}) \sum_{S \in \mathbf{\Sigma}} \left| \mathcal{N}(S) - \mathcal{N}^\star(S) \right| .$$

So, the LHS is at least $\Theta(\alpha)$ times the absolute gap of the masses assigned by $\mathcal{N}$ and $\mathcal{N}^\star$ over a worst-case set that lies in a good partition (one with $\mathcal{N}^\star(U_{w,c,\mathbf{\Sigma}}) \geq \alpha/2$). This holds since the probability to draw a good partition is at least $\alpha/2$. The following optimization problem gives a lower bound on the mass gap of a worst-case set in a good partition and, consequently, a lower bound on the total variation distance between $\mathcal{N}_\pi^\star$ and $\mathcal{N}_\pi$.

$$\min_S \left| \int (\mathcal{N}(\boldsymbol{\mu}^\star; \boldsymbol{x}) - \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{x})) \mathbf{1}_S(\boldsymbol{x}) d\boldsymbol{x} \right| ,$$

$$\text{subj. to} \quad \int \mathcal{N}(\boldsymbol{\mu}^\star; \boldsymbol{x}) \mathbf{1}_S(\boldsymbol{x}) d\boldsymbol{x} \geq \alpha/2 .$$

We begin with a claim about the shape of the worst case set. Let $t = (\|\boldsymbol{\mu}\|_2^2 - \|\boldsymbol{\mu}^\star\|_2^2)/2$ be the hyperplane threshold.

Claim G.3. *Let $\mathcal{H}^+ = \{\boldsymbol{x} : \boldsymbol{x}^T(\boldsymbol{\mu} - \boldsymbol{\mu}^\star) < t\}$ and $\mathcal{H}^- = \{\boldsymbol{x} : \boldsymbol{x}^T(\boldsymbol{\mu} - \boldsymbol{\mu}^\star) > t\}$. The mass of the solution of the fractional knapsack is totally contained in either $\mathcal{H}^+$ or $\mathcal{H}^-$.*

Since the partition distribution satisfies Equation (G.1) with respect to the true Gaussian $\mathcal{N}(\boldsymbol{\mu}^\star)$ and since the set $\mathcal{H}$ is a hyperplane, the probability mass that is not cut by $\mathcal{H}$ is at least α. Hence, there exists a halfspace (either $\mathcal{H}^+$ or $\mathcal{H}^-$) with mass at least $\alpha/2$. Also, observe that the hyperplane $\mathcal{H}$ is the zero locus of the polynomial $q(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 - \|\boldsymbol{x} - \boldsymbol{\mu}^\star\|_2^2$ and, hence, it is the set of points where the two spherical Gaussians $\mathcal{N}(\boldsymbol{\mu})$ and $\mathcal{N}(\boldsymbol{\mu}^\star)$ assign equal mass. We have that

$$\mathcal{H}^+ = \left\{ \boldsymbol{x} : \mathcal{N}(\boldsymbol{\mu}^\star) > \mathcal{N}(\boldsymbol{\mu}) \right\} .$$

Hence, we can assume that the worst-case set lies totally in $\mathcal{H}^+$ and, then, the optimization problem can be written as

$$\min_S \int \left(1 - \frac{\mathcal{N}(\boldsymbol{\mu}; \boldsymbol{x})}{\mathcal{N}(\boldsymbol{0}; \boldsymbol{x})}\right) \mathcal{N}(\boldsymbol{0}; \boldsymbol{x}) \mathbf{1}_S(\boldsymbol{x}) d\boldsymbol{x}\,,$$

$$\text{subj. to} \quad \int \mathcal{N}(\boldsymbol{0}; \boldsymbol{x}) \mathbf{1}_S(\boldsymbol{x}) d\boldsymbol{x} \geq \alpha/2, \;\; S \in \mathcal{H}^+\,.$$

Without loss of generality, we assume that $\mathcal{N}^\star = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\mathcal{N} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I})$. In order to design the worst-case set, since the optimization has the structure of the fractional knapsack problem, we can think of each point $\boldsymbol{x} \in \mathcal{H}^+$ as having *weight* equal to its contribution to the mass gap $(\mathcal{N}(\boldsymbol{0}; \boldsymbol{x}) - \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{x}))$ and *value* equal to its density with respect to the true Gaussian $\mathcal{N}(\boldsymbol{0}; \boldsymbol{x})$. Hence, in order to design the worst-case set, the points $\boldsymbol{x} \in \mathcal{H}^+$ should be included in the set in order of increasing ratio of weight over value, until reaching a threshold T. So, we can define the worst-case set to be

$$S = \left\{\boldsymbol{x} \in \mathcal{H}^+ : 1 - \frac{\mathcal{N}(\boldsymbol{\mu}; \boldsymbol{x})}{\mathcal{N}(\boldsymbol{0}; \boldsymbol{x})} \leq T\right\} = \left\{\boldsymbol{x} \in \mathcal{H}^+ : 1 - \exp(p(\boldsymbol{x})) \leq T\right\},$$

where $p(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{x})^T(\boldsymbol{\mu} - \boldsymbol{x}) + \frac{1}{2}\boldsymbol{x}^T\boldsymbol{x} = -\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{x}$ and note that $p(\boldsymbol{x}) \leq 0$ for any $\boldsymbol{x} \in \mathcal{H}^+$. We will use the following anti-concentration result about the Gaussian mass of sets, defined by polynomials.

Lemma G.4 (Theorem 8 of Carbery and Wright (2001)). *Let $q, \gamma \in \mathbb{R}_+, \boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}$ in the positive semidefinite cone $\mathbb{S}^d_+$. Consider $p : \mathbb{R}^d \to \mathbb{R}$ a multivariate polynomial of degree at most ℓ and let*

$$\mathcal{Q} = \left\{\boldsymbol{x} \in \mathbb{R}^d : |p(\boldsymbol{x})| \leq \gamma\right\}.$$

Then, there exists an absolute constant C such that

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathcal{Q}) \leq \frac{Cq\gamma^{1/\ell}}{(\mathbf{E}_{\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[|p(\boldsymbol{z})|^{q/\ell}])^{1/q}}\,.$$

We can apply Lemma G.4 for the quadratic polynomial $p(\boldsymbol{x})$ by setting $\gamma = \frac{\alpha^2}{256C^2}\sqrt{\mathbf{E}_{\boldsymbol{x} \sim \mathcal{N}^\star}[p^2(\boldsymbol{x})]}$ with $q = 4$, where C is the absolute Carbery-Wright constant.

Hence, we get that the Gaussian mass of the set $\mathcal{Q} = \{\boldsymbol{x} : |p(\boldsymbol{x})| \leq \gamma\}$ is equal to

$$\mathcal{N}^\star(\mathcal{Q}) \leq \alpha/4 \,.$$

So, for any point $\boldsymbol{x}$ in the remaining $\alpha/4$ mass of the set S, it holds that $|p(\boldsymbol{x})| \geq \gamma$. We first observe that γ can lower bounded by the total variation distance of $\mathcal{N}^\star$ and $\mathcal{N}$. It suffices to lower bound the expectation $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{N}^\star}[p^2(\boldsymbol{x})]$. We have that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^\star}\left[p^2(\boldsymbol{x})\right] \geq \mathrm{Var}_{\boldsymbol{x}\sim\mathcal{N}^\star}\left[p(\boldsymbol{x})\right] = \mathrm{Var}_{\boldsymbol{x}\sim\mathcal{N}^\star}\left[-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{x}\right] = \|\boldsymbol{\mu}\|_2^2 \,,$$

and, hence

$$\gamma \geq \frac{\alpha^2}{256C^2} \cdot \|\boldsymbol{\mu}\|_2 \,.$$

We will use the following lemma for the total variation distance of two Normal distributions.

Lemma G.5 (see Corollaries 2.13 and 2.14 of Diakonikolas et al. (2016b))**.** *Let $N_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), N_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ be two Normal distributions. Then, it holds*

$$d_{\mathrm{TV}}(N_1, N_2) \leq \frac{1}{2}\left\|\boldsymbol{\Sigma}_1^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right\|_2 + \sqrt{2}\left\|\boldsymbol{I} - \boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2}\right\|_F \,.$$

Applying Lemma G.5 to the above inequality, we get

$$\gamma \geq \frac{\alpha^2}{256C^2} \cdot d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}), \mathcal{N}(\boldsymbol{\mu}^\star)) \,.$$

To conclude, we have to lower bound the L_1 gap between $\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x})\mathbf{1}_S(\boldsymbol{x})$ and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}; \boldsymbol{x})\mathbf{1}_S(\boldsymbol{x})$ and since S lies totally in $\mathcal{H}^+$

$$\int_S (\mathcal{N}(\mathbf{0}; \boldsymbol{x}) - \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{x}))d\boldsymbol{x} = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^\star}\left[1 - \exp(p(\boldsymbol{x}))\Big|\mathbf{1}_S(\boldsymbol{x})\right] \,.$$

To proceed, we distinguish two cases: First, assume that $\gamma \leq 1$ and recall that $\mathcal{Q} = \{\boldsymbol{x} : |p(\boldsymbol{x})| \leq \gamma\}$. Note that for $y \in [-1, 0]$, it holds that $1 - \exp(y) \geq |y|/2$

and, hence, we have that:

$$\int_S (\mathcal{N}(\mathbf{0};\boldsymbol{x}) - \mathcal{N}(\boldsymbol{\mu};\boldsymbol{x}))d\boldsymbol{x} \geq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^\star}\left[\frac{|p(\boldsymbol{x})|}{2}\mathbf{1}_{S\setminus\mathcal{Q}}(\boldsymbol{x})\right] \geq \gamma \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^\star}\left[\mathbf{1}_{S\setminus\mathcal{Q}}(\boldsymbol{x})\right] \geq \frac{\alpha\gamma}{4}\,,$$

and, by the lower bound for γ, we get

$$\int_S (\mathcal{N}(\mathbf{0},\boldsymbol{I};\boldsymbol{x}) - \mathcal{N}(\boldsymbol{\mu},\boldsymbol{I};\boldsymbol{x}))d\boldsymbol{x} \geq C_\alpha \cdot d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}),\mathcal{N}(\boldsymbol{\mu}^\star))\,,$$

for some $C_\alpha = \Omega(\alpha^3)$. Otherwise, let $\gamma > 1$. Note that for $y < -1$, it holds that $1 - \exp(y) \geq 1/2$. Hence, we get that

$$\int_S (\mathcal{N}(\mathbf{0};\boldsymbol{x}) - \mathcal{N}(\boldsymbol{\mu};\boldsymbol{x}))d\boldsymbol{x} \geq \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{N}^\star}\left[\frac{1}{2}\mathbf{1}_{S\setminus\mathcal{Q}}(\boldsymbol{x})\right] \geq \alpha/8\,.$$

In conclusion, we get that

$$d_{\mathrm{TV}}(\mathcal{N}_\pi^*,\mathcal{N}_\pi) \geq C_\alpha \cdot d_{\mathrm{TV}}(\mathcal{N}^\star,\mathcal{N})\,,$$

where $C_\alpha = \mathrm{poly}(\alpha)$ and depends only on α. □

G.4 Literature Overview on Partial Label Learning

The problem of learning from coarse labels falls in the regime of semi-supervised learning Chapelle et al. (2006) and it appears in various literature threads termed as (i) partial label learning Cour et al. (2011), (ii) ambiguous label learning Cour et al. (2009); Hüllermeier and Beringer (2006), (iii) superset label learning Hüllermeier and Cheng (2015) and (iv) soft label learning Côme et al. (2008). Closely related to these tasks are the problems of learning from complementary labels Ishida et al. (2017) and, more generally, learning from noisy and corrupted examples Angluin and Laird (1988); Scott et al. (2013); Blanchard and Scott (2014); Van Rooyen and Williamson (2017); Lukasik et al. (2020).

We stick with the term partial label learning for now since this is the most widely used. Many real-world learning tasks were solved under the framework of partial label learning such as multimedia content analysis Cour et al. (2009, 2011) and semantic image segmentation Papandreou et al. (2015).

We refer to Jin and Ghahramani (2002); Nguyen and Caruana (2008) and the references therein for some seminal papers in the area. Through the years, various approaches have been proposed to solve this challenging problem by utilizing major machine learning techniques, such as maximum likelihood estimation and Expectation-Maximization Jin and Ghahramani (2002), convex optimization Cour et al. (2011), *k*-nearest neighbors Hüllermeier and Beringer (2006) and error-correcting output codes Zhang (2014); Zhang et al. (2017a). For an overview of the practical treatment on the problem, we refer the interested reader to **?**Xu et al. (2021); Wen et al. (2021) (and the references therein) and more broadly to Triguero et al. (2015); Van Engelen and Hoos (2020).

Despite extensive studies on partial label learning from an industrial perspective (applied ML), our theoretical level of understanding is still limited. A fundamental line of research deals with the statistical consistency (see e.g., Cour et al. (2011); Cid-Sueiro et al. (2014); Feng et al. (2020); Cabannnes et al. (2020); Lv et al. (2020); Wen et al. (2021)) and the learnability Liu and Dietterich (2014) of partial label learning algorithms. Moreover, Cauchois et al. (2022) present a methodology between partial supervision and validation.

Closer to our learning from coarse labels approach are the works of Cid-Sueiro (2012) and Van Rooyen and Williamson (2017). In the former, the goal is to estimate the posterior class probabilities from partially labelled data while, in the latter, the authors study a more general problem of learning from corrupted labels and aim to “invert” the corruption. This technique is inspired by the work of Natarajan et al. (2013), where the authors proposed the method of unbiased estimators (which is close to the connection between random classification noise and the SQ framework of Kearns (1998)). This backward correction procedure of Natarajan et al. (2013); Cid-Sueiro (2012); Van Rooyen and Williamson (2017) recovers the information lost from the corrupted labels (under some structural assumptions) and results in an unbiased estimate of the risk with respect to true distribution. Crucially, these works have to assume that the corruption process (i.e., the coarsening mechanism) is known. This is also commented in Cabannnes et al. (2020). Our SQ reduction does not require to know the mechanism; in some sense, the algorithm uses rejection sampling and learning coarse discrete distributions (which is an unsupervised learning problem) in order to invert the coarsening in the sense of Van Rooyen and Williamson (2017) and obtain statistical queries with respect to the distribution over the finely-labeled examples.

REFERENCES

Agarwal, A., P. Patil, and S. Agarwal. 2018. Accelerated spectral ranking. In *International conference on machine learning*, 70–79. PMLR.

Ahamada, I., and E. Flachaire. 2010. Non-parametric econometrics. *OUP Catalogue*.

Ailon, Nir, Moses Charikar, and Alantha Newman. 2008. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)* 55(5):1–27.

Aledo, Juan A, José A Gámez, and David Molina. 2017. Tackling the supervised label ranking problem by bagging weak learners. *Information Fusion* 35:38–50.

Anderson, Joseph, Navin Goyal, and Luis Rademacher. 2013. Efficient learning of simplices. In *Conference on learning theory*, 1020–1045.

Angluin, D., and P. Laird. 1988. Learning from noisy examples. *Machine Learning* 2(4):343–370.

Arjevani, Y., Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. 2019. Lower bounds for non-convex stochastic optimization. `1912.02365`.

Aslam, Javed A, and Scott E Decatur. 1998. General bounds on statistical query learning and pac learning with noise via hypothesis boosting. *Information and Computation* 141(2):85–118.

Awasthi, P. 2018. Noisy pac learning of halfspaces. TTI Chicago, Summer Workshop on Robust Statistics, available at http://www.iliasdiakonikolas.org/tti-robust/Awasthi.pdf.

Awasthi, P., M. F. Balcan, N. Haghtalab, and R. Urner. 2015. Efficient learning of linear separators under bounded noise. In *Proceedings of the 28th conference on learning theory, COLT 2015*, 167–190.

Awasthi, P., M. F. Balcan, N. Haghtalab, and H. Zhang. 2016a. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th conference on learning theory, COLT 2016*, 152–192.

Awasthi, P., M. F. Balcan, and P. M. Long. 2017. The power of localization for efficiently learning linear separators with noise. *J. ACM* 63(6):50:1–50:27.

Awasthi, Pranjal, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. 2016b. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on learning theory*, 152–192. PMLR.

Awasthi, Pranjal, Avrim Blum, Or Sheffet, and Aravindan Vijayaraghavan. 2014. Learning mixtures of ranking models. *arXiv preprint arXiv:1410.8750*.

Bakshi, Ainesh, Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, Sushrut Karmalkar, and Pravesh K Kothari. 2020. Outlier-robust clustering of gaussians and other non-spherical mixtures. In *2020 ieee 61st annual symposium on foundations of computer science (focs)*, 149–159. IEEE Computer Society.

Balakrishnan, N, and Erhard Cramer. 2014. *The art of progressive censoring*. Springer.

Balcan, M.-F., A. Z. Broder, and T.Zhang. 2007. Margin based active learning. In *Learning theory, 20th annual conference on learning theory, COLT 2007*, vol. 4539 of *Lecture Notes in Computer Science*, 35–50. Springer.

Balcan, M. F., and N. Haghtalab. 2020a. Noise in classification. In *Beyond the worst-case analysis of algorithms*, ed. T. Roughgarden. Cambridge University Press.

Balcan, M.-F., and H. Zhang. 2017a. Sample and computationally efficient learning algorithms under s-concave distributions. In *Advances in neural information processing systems*, 4796–4805.

Balcan, Maria Florina, and Vitaly Feldman. 2015. Statistical active learning algorithms for noise tolerance and differential privacy. *Algorithmica* 72(1):282–315.

Balcan, Maria-Florina, and Nika Haghtalab. 2020b. Noise in classification.

Balcan, Maria-Florina F, and Hongyang Zhang. 2017b. Sample and computationally efficient learning algorithms under s-concave distributions. *Advances in Neural Information Processing Systems* 30.

Ball, Keith. 1993. The reverse isoperimetric problem for gaussian measure. *Discrete & Computational Geometry* 10(1):411–420.

Barron, Andrew R, Lhszl Gyorfi, and Edward C van der Meulen. 1992. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE transactions on Information Theory* 38(5):1437–1454.

Barron, Andrew R., and Chyong-Hwa Sheu. 1991. Approximation of density functions by sequences of exponential families. *The Annals of Statistics* 19(3): 1347–1369.

Bartlett, P. L., O. Bousquet, and S. Mendelson. 2005. Local rademacher complexities. *Ann. Statist.* 33(4):1497–1537.

Bartlett, P. L., M. I. Jordan, and J. D. Mcauliffe. 2006. Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 101(473):138–156.

Beigman, E., and B. B. Klebanov. 2009. Learning with annotation noise. In *Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp*, 280–287.

Ben-David, Shai, Nicolò Cesa-Bianchi, and Philip M Long. 1992. Characterizations of learnability for classes of {O,..., n}-valued functions. In *Proceedings of the fifth annual workshop on computational learning theory*, 333–340.

Ben-David, Shalev, Adam Bouland, Ankit Garg, and Robin Kothari. 2018. Classical lower bounds from quantum upper bounds. In *59th IEEE annual symposium on foundations of computer science, FOCS 2018, paris, france, october 7-9, 2018*, 339–349.

Bhattacharyya, Arnab, Rathin Desai, Sai Ganesh Nagarajan, and Ioannis Panageas. 2020. Efficient statistics for sparse graphical models from truncated samples. *arXiv preprint arXiv:2006.09735*.

Blanchard, Gilles, and Clayton Scott. 2014. Decontamination of mutually contaminated models. In *Artificial intelligence and statistics*, 1–9. PMLR.

Blum, A., A. Frieze, R. Kannan, and S. Vempala. 1997. A polynomial time algorithm for learning noisy linear threshold functions. *Algorithmica* 22(1/2):35–52.

———. 1998. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica* 22(1):35–52.

Blum, A., A. M. Frieze, R. Kannan, and S. Vempala. 1996. A polynomial-time algorithm for learning noisy linear threshold functions. In *37^{th} annual symposium on foundations of computer science, FOCS '96*, 330–338.

Blum, Avrim, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. 2005. Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth acm sigmod-sigact-sigart symposium on principles of database systems*, 128–138.

Börsch-Supan, Axel, and Vassilis A Hajivassiliou. 1993. Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of econometrics* 58(3):347–368.

Botev, Zdravko I, Joseph F Grotowski, Dirk P Kroese, et al. 2010. Kernel density estimation via diffusion. *The annals of Statistics* 38(5):2916–2957.

Boucheron, S., O. Bousquet, and G. Lugosi. 2005. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics* 1(9):323–375.

Bradley, Ralph Allan, and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39(3/4):324–345.

Braverman, Mark, and Elchanan Mossel. 2009. Sorting from noisy information. *arXiv preprint arXiv:0910.1191*.

Breen, Richard, et al. 1996. *Regression models: Censored, sample selected, or truncated data*, vol. 111. Sage.

Bukchin, Guy, Eli Schwartz, Kate Saenko, Ori Shahar, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. 2020. Fine-grained angular contrastive learning with coarse labels. *arXiv preprint arXiv:2012.03515*.

Busa-Fekete, Róbert, Dimitris Fotakis, Balázs Szörényi, and Manolis Zampetakis. 2019. Optimal learning of mallows block model. In *Conference on learning theory*, 529–532. PMLR.

Cabannnes, Vivien, Alessandro Rudi, and Francis Bach. 2020. Structured prediction with partial labelling through the infimum loss. In *International conference on machine learning*, 1230–1239. PMLR.

Canu, Stéphane, and Alex Smola. 2006. Kernel methods and the exponential family. *Neurocomputing* 69(7-9):714–720.

Caragiannis, Ioannis, Ariel D Procaccia, and Nisarg Shah. 2013. When do noisy votes reveal the truth? In *Proceedings of the fourteenth acm conference on electronic commerce*, 143–160.

Carbery, A., and J. Wright. 2001. Distributional and L^q norm inequalities for polynomials over convex bodies in R^n. *Mathematical Research Letters* 8(3):233–248.

Castro, R. M., and R. D. Nowak. 2008. Minimax bounds for active learning. *IEEE Transactions on Information Theory* 54(5):2339–2353.

Cauchois, Maxime, Suyash Gupta, Alnur Ali, and John Duchi. 2022. Predictive inference with weak supervision. *arXiv preprint arXiv:2201.08315*.

Chapelle, Olivier, Bernhard Schölkopf, and Alexander Zien. 2006. *Semi-supervised learning (adaptive computation and machine learning)*. The MIT Press.

Charikar, M., J. Steinhardt, and G. Valiant. 2017. Learning from untrusted data. In *Proceedings of stoc 2017*, 47–60.

Chen, S., F. Koehler, A. Moitra, and M. Yau. 2020a. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. In *Advances in neural information processing systems, neurips*.

Chen, Sitan, Frederic Koehler, Ankur Moitra, and Morris Yau. 2020b. Classification under misspecification: Halfspaces, generalized linear models, and connections to evolvability. *arXiv preprint arXiv:2006.04787*.

Chen, Yi-Chen, Vishal M Patel, Rama Chellappa, and P Jonathon Phillips. 2014. Ambiguously labeled learning using dictionaries. *IEEE Transactions on Information Forensics and Security* 9(12):2076–2088.

Chen, Zhuo, Ruizhou Ding, Ting-Wu Chin, and Diana Marculescu. 2018. Understanding the impact of label granularity on cnn-based image classification. In *2018 ieee international conference on data mining workshops (icdmw)*, 895–904. IEEE.

Cheng, Weiwei, Krzysztof Dembczynski, and Eyke Hüllermeier. 2010. Label ranking methods based on the plackett-luce model. In *Icml*.

Cheng, Weiwei, Jens Hühn, and Eyke Hüllermeier. 2009. Decision tree and instance-based learning for label ranking. In *Proceedings of the 26th annual international conference on machine learning*, 161–168.

Cheng, Weiwei, and Eyke Hüllermeier. 2008. Instance-based label ranking using the mallows model. In *Eccbr workshops*, 143–157.

———. 2012. Probability estimation for multi-class classification based on label ranking. In *Joint european conference on machine learning and knowledge discovery in databases*, 83–98. Springer.

Cheng, Yu, Ilias Diakonikolas, Rong Ge, and Mahdi Soltanolkotabi. 2020. High-dimensional robust mean estimation via gradient descent. In *International conference on machine learning*, 1768–1778. PMLR.

Chhikara, R. S., and J. McKeon. 1984. Linear discriminant analysis with misallocation in training samples. *Journal of the American Statistical Association* 79(388): 899–906.

Cid-Sueiro, Jesús. 2012. Proper losses for learning from partial labels. *Advances in neural information processing systems* 25.

Cid-Sueiro, Jesús, Darío García-García, and Raúl Santos-Rodríguez. 2014. Consistency of losses for learning from weak labels. In *Joint european conference on machine learning and knowledge discovery in databases*, 197–210. Springer.

Clémençon, Stephan, Anna Korba, and Eric Sibony. 2018. Ranking median regression: Learning to order through local consensus. In *Algorithmic learning theory*, 212–245. PMLR.

Clémençon, Stéphan, and Robin Vogel. 2020. A multiclass classification approach to label ranking. In *International conference on artificial intelligence and statistics*, 1421–1430. PMLR.

Cohen, A Clifford. 1991. *Truncated and censored samples: theory and applications*. CRC press.

———. 2016. *Truncated and censored samples: theory and applications*. CRC press.

Côme, Etienne, Latifa Oukhellou, Thierry Denœux, and Patrice Aknin. 2008. Mixture model estimation with soft labels. In *Soft methods for handling variability and imprecision*, 165–174. Springer.

Cour, Timothee, Ben Sapp, and Ben Taskar. 2011. Learning from partial labels. *The Journal of Machine Learning Research* 12:1501–1536.

Cour, Timothee, Benjamin Sapp, Chris Jordan, and Ben Taskar. 2009. Learning from ambiguously labeled images. In *2009 ieee conference on computer vision and pattern recognition*, 919–926. IEEE.

Dagan, I., O. Glickman, and B. Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, 177–190. Springer.

Daniely, A. 2015. A PTAS for agnostically learning halfspaces. In *Proceedings of the 28th conference on learning theory, COLT 2015*, 484–502.

———. 2016a. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th annual symposium on theory of computing, STOC 2016*, 105–117.

Daniely, Amit. 2016b. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the forty-eighth annual acm symposium on theory of computing*, 105–117.

Daniely, Amit, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. 2011. Multiclass learnability and the erm principle. In *Proceedings of the 24th annual conference on learning theory*, 207–232. JMLR Workshop and Conference Proceedings.

Daniely, Amit, and Shai Shalev-Shwartz. 2014. Optimal learners for multiclass problems. In *Conference on learning theory*, 287–316. PMLR.

Dasgupta, Sanjoy, Adam Tauman Kalai, and Claire Monteleoni. 2005. Analysis of perceptron-based active learning. In *Learning theory: 18th annual conference on learning theory, colt 2005, bertinoro, italy, june 27-30, 2005. proceedings 18*, 249–263. Springer.

Daskalakis, Constantinos, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. 2018. Efficient Statistics, in High Dimensions, from Truncated Samples. In *59th annual ieee symposium on foundations of computer science (focs)*, 639–649. IEEE.

———. 2019. Computationally and statistically efficient truncated regression. In *Conference on learning theory*, 955–960. PMLR.

Daskalakis, Constantinos, and Gautam Kamath. 2014. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of the 27th conference on learning theory, COLT 2014, barcelona, spain, june 13-15, 2014*, 1183–1213.

Daskalakis, Constantinos, Vasilis Kontonis, Christos Tzamos, and Emmanouil Zampetakis. 2021. A statistical taylor theorem and extrapolation of truncated densities. In *Conference on learning theory*, 1395–1398. PMLR.

Daskalakis, Constantinos, Dhruv Rohatgi, and Manolis Zampetakis. 2020. Truncated linear regression in high dimensions. *arXiv preprint arXiv:2007.14539*.

d'Aspremont, Alexandre. 2008. Smooth optimization with approximate gradient. *SIAM Journal on Optimization* 19(3):1171–1183.

De, Anindya, Ilias Diakonikolas, and Rocco A Servedio. 2014. Learning from satisfying assignments. In *Proceedings of the twenty-sixth annual acm-siam symposium on discrete algorithms*, 478–497. SIAM.

De, Anindya, Ryan O'Donnell, and Rocco Servedio. 2018. Learning sparse mixtures of rankings from noisy information. *arXiv preprint arXiv:1811.01216*.

Dekel, Ofer, Yoram Singer, and Christopher D Manning. 2003. Log-linear models for label ranking. *Advances in neural information processing systems* 16:497–504.

Deng, Jia, Jonathan Krause, and Li Fei-Fei. 2013. Fine-grained crowdsourcing for fine-grained recognition. In *Proceedings of the 2013 ieee conference on computer vision and pattern recognition*, 580–587. CVPR '13, USA: IEEE Computer Society.

Denis, François. 1998. Pac learning from positive statistical queries. In *International conference on algorithmic learning theory*, 112–126. Springer.

Devolder, Olivier, François Glineur, and Yurii Nesterov. 2014. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming* 146(1):37–75.

Devroye, L., and G. Lugosi. 2001. *Combinatorial methods in density estimation*. Springer: Springer Series in Statistics.

Devroye, Luc, and Gábor Lugosi. 2012. *Combinatorial methods in density estimation*. Springer Science & Business Media.

Diakonikolas, I., P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. 2010a. Bounded independence fools halfspaces. *SIAM Journal on Computing* 39(8):3441–3462.

Diakonikolas, I., T. Gouleakis, and C. Tzamos. 2019a. Distribution-independent pac learning of halfspaces with massart noise. In *Advances in neural information processing systems 32, neurips*.

Diakonikolas, I., G. Kamath, D. Kane, J. Li, J. Steinhardt, and Alistair Stewart. 2019b. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th international conference on machine learning, ICML 2019*.

Diakonikolas, I., G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. 2016a. Robust estimators in high dimensions without the computational intractability. In *Proceedings of focs'16*, 655–664.

———. 2017a. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th international conference on machine learning, ICML 2017*, 999–1008.

———. 2018a. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the twenty-ninth annual ACM-SIAM symposium on discrete algorithms, SODA 2018*, 2683–2702.

Diakonikolas, I., D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. 2021a. Efficiently learning halfspaces with tsybakov noise. *Symposium on Theory of Computing (STOC)*.

Diakonikolas, I., D. M. Kane, V. Kontonis, and N. Zarifis. 2020a. Algorithms and SQ lower bounds for pac learning one-hidden-layer ReLU networks. In *Conference on learning theory, COLT*, 1514–1539. PMLR.

Diakonikolas, I., D. M. Kane, and J. Nelson. 2010b. Bounded independence fools degree-2 threshold functions. In *Focs*, 11–20.

Diakonikolas, I., D. M. Kane, and A. Stewart. 2017b. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 ieee 58th annual symposium on foundations of computer science (focs)*, 73–84.

———. 2018b. Learning geometric concepts with nasty noise. In *Proceedings of the 50^{th} annual ACM SIGACT symposium on theory of computing, STOC 2018*, 1061–1073.

Diakonikolas, I., D. M. Kane, and N. Zarifis. 2020b. Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under Gaussian marginals. In *Advances in neural information processing systems, NeurIPS*.

Diakonikolas, I., W. Kong, and A. Stewart. 2019c. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the thirtieth annual ACM-SIAM symposium on discrete algorithms, SODA 2019*.

Diakonikolas, I., V. Kontonis, C. Tzamos, and N. Zarifis. 2020c. Learning halfspaces with massart noise under structured distributions. In *Conference on learning theory, COLT*.

———. 2021b. Learning halfspaces with tsybakov noise. *Symposium on Theory of Computing (STOC)*.

Diakonikolas, Ilias, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2016b. Robust estimators in high dimensions without the computational intractability. In *IEEE 57th annual symposium on foundations of computer science, FOCS 2016, 9-11 october 2016, hyatt regency, new brunswick, new jersey, USA*, 655–664.

———. 2018c. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the twenty-ninth annual ACM-SIAM symposium on discrete algorithms, SODA 2018, new orleans, la, usa, january 7-10, 2018*, 2683–2702.

Diakonikolas, Ilias, Daniel Kane, and Christos Tzamos. 2021c. Forster decomposition and learning halfspaces with noise. *Advances in Neural Information Processing Systems* 34.

Diakonikolas, Ilias, Daniel Kane, and Nikos Zarifis. 2020d. Near-optimal sq lower bounds for agnostically learning halfspaces and relus under gaussian marginals. *Advances in Neural Information Processing Systems* 33:13586–13596.

Diakonikolas, Ilias, and Daniel M Kane. 2020. Hardness of learning halfspaces with massart noise. *arXiv preprint arXiv:2012.09720*.

Diakonikolas, Ilias, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. 2021d. Learning general halfspaces with general massart noise under the gaussian distribution. *arXiv preprint arXiv:2108.08767*.

Diakonikolas, Ilias, Daniel M Kane, Thanasis Pittas, and Nikos Zarifis. 2021e. The optimality of polynomial regression for agnostic learning under gaussian marginals. *arXiv preprint arXiv:2102.04401*.

Diakonikolas, Ilias, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. 2020e. Learning halfspaces with massart noise under structured distributions. In *Conference on learning theory*, 1486–1513. PMLR.

Djuric, Nemanja, Mihajlo Grbovic, Vladan Radosavljevic, Narayan Bhamidipati, and Slobodan Vucetic. 2014. Non-linear label ranking for large-scale prediction of long-term user interests. In *Twenty-eighth aaai conference on artificial intelligence*.

Drori, Y., and O. Shamir. 2019. The complexity of finding stationary points with stochastic gradient descent. `1910.01845`.

Dunagan, John, and Santosh Vempala. 2008. A simple polynomial-time rescaling algorithm for solving linear programs. *Mathematical Programming* 114(1):101–114.

Eldan, Ronen. 2011. A polynomial number of random points does not determine the volume of a convex body. *Discrete & Computational Geometry* 46(1):29–47.

Feldman, V., P. Gopalan, S. Khot, and A. Ponnuswami. 2006a. New results for learning noisy parities and halfspaces. In *Proc. focs*, 563–576.

Feldman, V., C. Guzman, and S. Vempala. 2015a. Statistical query algorithms for stochastic convex optimization. *CoRR* abs/1512.09170.

Feldman, V., W. Perkins, and S. Vempala. 2015b. 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*, 77–86.

Feldman, Vitaly. 2017. A general characterization of the statistical query complexity. In *Conference on learning theory*, 785–830. PMLR.

Feldman, Vitaly, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. 2006b. New results for learning noisy parities and halfspaces. In *2006 47th annual ieee symposium on foundations of computer science (focs'06)*, 563–574. IEEE.

Feldman, Vitaly, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. 2017. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)* 64(2):1–37.

Feller, William. 1957. An introduction to probability theory and its applications. *John Wiley*.

Feng, Lei, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. 2020. Provably consistent partial-label learning. *Advances in Neural Information Processing Systems* 33:10948–10960.

Fisher, RA. 1931. Properties and applications of Hh functions. *Mathematical tables* 1:815–852.

Fleming, Noah, Pravesh Kothari, Toniann Pitassi, et al. 2019. Semialgebraic proofs and efficient algorithm design. *Foundations and Trends® in Theoretical Computer Science* 14(1-2):1–221.

Fotakis, D., A. Kalavasis, and C. Kontonis, V. Tzamos. 2022. Noisy linear label ranking. In *Advances in neural information processing systems 35, neurips*.

Fotakis, D., A. Kalavasis, V. Kontonis, and C. Tzamos. 2021a. Efficient algorithms for learning from coarse labels. In *Conference on learning theory, (colt)*.

Fotakis, Dimitris, Alkis Kalavasis, and Eleni Psaroudaki. 2021b. Label ranking through nonparametric regression. *arXiv preprint arXiv:2111.02749*.

Fotakis, Dimitris, Alkis Kalavasis, and Konstantinos Stavropoulos. 2021c. Aggregating incomplete and noisy rankings. In *International conference on artificial intelligence and statistics*, 2278–2286. PMLR.

Fotakis, Dimitris, Alkis Kalavasis, and Christos Tzamos. 2020. Efficient parameter estimation of truncated boolean product distributions. In *Conference on learning theory*, 1586–1600. PMLR.

Frénay, B., and M. Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25(5):845–869.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani, et al. 2001. *The elements of statistical learning*, vol. 1. Springer series in statistics New York.

Frieze, Alan, Mark Jerrum, and Ravi Kannan. 1996. Learning linear transformations. In *Foundations of computer science, 1996. proceedings., 37th annual symposium on*, 359–368. IEEE.

Fürnkranz, Johannes, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. 2008. Multilabel classification via calibrated label ranking. *Machine learning* 73(2): 133–153.

Gajek, Leslaw, et al. 1988. On the minimax value in the scale model with truncated data. *The Annals of Statistics* 16(2):669–677.

Galton, Francis. 1897. An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London* 62(379-387):310–315.

Gasca, Mariano, and Thomas Sauer. 2000. Polynomial interpolation in several variables. *ADV. COMPUT. MATH* 12:377–410.

Ghadimi, S., G. Lan, and H. Zhang. 2016. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming* 155(1-2):267–305.

Gill, Richard D, Mark J Van Der Laan, and James M Robins. 1997. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the first seattle symposium in biostatistics*, 255–294. Springer.

Goel, S., A. Gollakota, Z. Jin, S. Karmalkar, and A. Klivans. 2020a. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International conference on machine learning, ICML*.

Goel, S., A. Gollakota, and A. R. Klivans. 2020b. Statistical-query lower bounds via functional gradients. In *Advances in neural information processing systems, NeurIPS*.

Goel, Surbhi, Aravind Gollakota, and Adam Klivans. 2020c. Statistical-query lower bounds via functional gradients. *Advances in Neural Information Processing Systems* 33:2147–2158.

Good, Irving J. 1963. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics* 34(3): 911–934.

Gourieroux, Christian. 2000. *Econometrics of qualitative dependent variables*. Cambridge university press.

Goyal, Navin, and Luis Rademacher. 2009. Learning convex bodies is hard. *arXiv preprint arXiv:0904.1227*.

Grbovic, Mihajlo, Nemanja Djuric, Shengbo Guo, and Slobodan Vucetic. 2013. Supervised clustering of label ranking data using label preference information. *Machine learning* 93(2-3):191–225.

Grbovic, Mihajlo, Nemanja Djuric, and Slobodan Vucetic. 2012. Learning from pairwise preference data using gaussian mixture model. *Preference Learning: Problems and Applications in AI* 33.

Guionnet, Alice. 2009. *Large random matrices*, vol. 1957. Springer Science & Business Media.

Guo, Yanming, Yu Liu, Erwin M Bakker, Yuanhao Guo, and Michael S Lew. 2018. Cnn-rnn: a large-scale hierarchical image classification framework. *Multimedia tools and applications* 77(8):10251–10271.

Guruswami, V., and P. Raghavendra. 2006. Hardness of learning halfspaces with noise. In *Proc. 47th ieee symposium on foundations of computer science (focs)*, 543–552. IEEE Computer Society.

Guruswami, Venkatesan, and Prasad Raghavendra. 2009. Hardness of learning halfspaces with noise. *SIAM Journal on Computing* 39(2):742–765.

Hanneke, S. 2011. Rates of convergence in active learning. *Ann. Statist.* 39(1): 333–361.

Hanneke, S., and L. Yang. 2015. Minimax analysis of active learning. *J. Mach. Learn. Res.* 16:3487–3602.

Har-Peled, Sariel, Dan Roth, and Dav Zimak. 2003. Constraint classification for multiclass classification and ranking. *Advances in neural information processing systems* 809–816. URL: `https://proceedings.neurips.cc/paper/2002/file/16026d60ff9b54410b3435b403afd226-Paper.pdf`.

Håstad, Johan. 2001. Some optimal inapproximability results. *Journal of the ACM (JACM)* 48(4):798–859.

Haussler, D. 1992. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation* 100:78–150.

Haussler, David. 2018. Decision theoretic generalizations of the pac model for neural net and other learning applications. In *The mathematics of generalization*, 37–116. CRC Press.

Hazan, E. 2016. Introduction to online convex optimization. *Foundations and Trends® in Optimization* 2(3-4):157–325.

Heckman, James J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, 475–492. NBER.

Hopkins, M., D. M. Kane, S. Lovett, and G. Mahajan. 2020. Noise-tolerant, reliable active classification with comparison queries. In *Colt*.

Hopkins, Samuel B, and Jerry Li. 2019. How Hard is Robust Mean Estimation? In *Conference on learning theory*, 1649–1682.

Hsu, Daniel, Clayton Sanford, Rocco Servedio, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. 2022. Near-optimal statistical query lower bounds for agnostically learning intersections of halfspaces with gaussian marginals. *arXiv preprint arXiv:2202.05096*.

Huber, Peter J. 2004. *Robust statistics*, vol. 523. John Wiley & Sons.

Hüllermeier, Eyke, and Jürgen Beringer. 2006. Learning from ambiguously labeled examples. *Intelligent Data Analysis* 10(5):419–439.

Hüllermeier, Eyke, and Weiwei Cheng. 2015. Superset learning based on generalized loss minimization. In *Joint european conference on machine learning and knowledge discovery in databases*, 260–275. Springer.

Hüllermeier, Eyke, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. 2008. Label ranking by learning pairwise preferences. *Artificial Intelligence* 172(16): 1897–1916.

Hunter, David R. 2004. Mm algorithms for generalized bradley-terry models. *The annals of statistics* 32(1):384–406.

Ilyas, Andrew, Emmanouil Zampetakis, and Constantinos Daskalakis. 2020. A theoretical and practical framework for regression and classification from truncated samples. In *International conference on artificial intelligence and statistics*, 4463–4473. PMLR.

Ishida, Takashi, Gang Niu, Weihua Hu, and Masashi Sugiyama. 2017. Learning from complementary labels. *Advances in neural information processing systems* 30.

Jiao, Qihan, Zhi Liu, Gongyang Li, Linwei Ye, and Yang Wang. 2020. Fine-grained image classification with coarse and fine labels on one-shot learning. In *2020 ieee international conference on multimedia & expo workshops (icmew)*, 1–6. IEEE.

Jiao, Qihan, Zhi Liu, Linwei Ye, and Yang Wang. 2019. Weakly labeled fine-grained classification with hierarchy relationship of fine and coarse labels. *Journal of Visual Communication and Image Representation* 63:102584.

Jin, Rong, and Zoubin Ghahramani. 2002. Learning with multiple labels. *Advances in neural information processing systems* 15.

Kalai, A., A. Klivans, Y. Mansour, and R. Servedio. 2008. Agnostically learning halfspaces. *SIAM Journal on Computing* 37(6):1777–1805. Special issue for FOCS 2005.

Kalai, Adam Tauman, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. 2005. Agnostically learning halfspaces. In *46th annual IEEE symposium on foundations of computer science (FOCS 2005), 23-25 october 2005, pittsburgh, pa, usa, proceedings*, 11–20.

Kane, Daniel M. 2011. The gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. *computational complexity* 20(2):389–412.

Kearns, M., R. Schapire, and L. Sellie. 1994a. Toward Efficient Agnostic Learning. *Machine Learning* 17(2/3):115–141.

Kearns, Michael. 1998. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)* 45(6):983–1006.

Kearns, Michael J, Robert E Schapire, and Linda M Sellie. 1994b. Toward efficient agnostic learning. *Machine Learning* 17(2):115–141.

Kenyon-Mathieu, Claire, and Warren Schudy. 2006. How to rank with few errors–a ptas for weighted feedback arc set on tournaments. In *Electronic colloquium on computational complexity, report no. 144 (2006)*. Citeseer.

Klebanov, B. B., and E. Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics* 35(4):495–503.

———. 2010. 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*, 438–446.

Klivans, A., P. Long, and R. Servedio. 2009a. Learning Halfspaces with Malicious Noise. *Journal of Machine Learning Research* 10:2715–2740.

Klivans, A. R., and P. Kothari. 2014. Embedding hard learning problems into gaussian space. In *Approximation, randomization, and combinatorial optimization. algorithms and techniques, APPROX/RANDOM 2014*, 793–809.

Klivans, A. R., P. K. Kothari, and R. Meka. 2018. Efficient algorithms for outlier-robust regression. In *Conference on learning theory, COLT 2018*, 1420–1430.

Klivans, A. R., P. M. Long, and A. K. Tang. 2009b. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *13th international workshop, RANDOM 2009*, 588–600.

Klivans, Adam R., Ryan O'Donnell, and Rocco A. Servedio. 2008. Learning geometric concepts via gaussian surface area. In *49th annual IEEE symposium on foundations of computer science, FOCS 2008, october 25-28, 2008, philadelphia, pa, USA*, 541–550.

Kontonis, V., C. Tzamos, and M. Zampetakis. 2019. Efficient truncated statistics with unknown truncation. In *2019 ieee 60th annual symposium on foundations of computer science (focs)*, 1578–1595. IEEE.

Korba, Anna, Stephan Clémençon, and Eric Sibony. 2017. A learning theory of ranking aggregation. In *Artificial intelligence and statistics*, 1001–1010. PMLR.

Korba, Anna, Alexandre Garcia, and Florence d'Alché Buc. 2018. A structured prediction approach for label ranking. *arXiv preprint arXiv:1807.02374*.

Lai, K. A., A. B. Rao, and S. Vempala. 2016a. Agnostic estimation of mean and covariance. In *Proceedings of focs'16*.

Lai, Kevin A., Anup B. Rao, and Santosh Vempala. 2016b. Agnostic estimation of mean and covariance. In *IEEE 57th annual symposium on foundations of computer science, FOCS 2016, 9-11 october 2016, hyatt regency, new brunswick, new jersey, USA*, 665–674.

Lai, Tze Leung, and Zhiliang Ying. 1991. Estimating a distribution function with truncated and censored data. *The Annals of Statistics* 417–442.

Ledoux, Michel. 1994. Semigroup proofs of the isoperimetric inequality in euclidean and gauss space. *Bulletin des sciences mathématiques* 118(6):485–510.

Lee, Alice. 1914. Table of the gaussian" tail" functions; when the" tail" is larger than the body. *Biometrika* 10(2/3):208–214.

Lee, Y. T., and S. S. Vempala. 2017. Eldan's stochastic localization and the kls hyperplane conjecture: An improved lower bound for expansion. In *2017 ieee 58th annual symposium on foundations of computer science (focs)*, 998–1007.

Lei, Jie, Zhenyu Guo, and Yang Wang. 2017. Weakly supervised image classification with coarse and fine labels. In *2017 14th conference on computer and robot vision (crv)*, 240–247. IEEE.

Letouzey, Fabien, François Denis, and Rémi Gilleron. 2000. Learning from positive and unlabeled examples. In *International conference on algorithmic learning theory*, 71–85. Springer.

Li, Qi, and Jeffrey Scott Racine. 2007. *Nonparametric econometrics: theory and practice*. Princeton University Press.

Liu, Allen, and Ankur Moitra. 2018. Efficiently learning mixtures of mallows models. In *2018 ieee 59th annual symposium on foundations of computer science (focs)*, 627–638. IEEE.

———. 2021. Robust voting rules from algorithmic robust statistics. *arXiv preprint arXiv:2112.06380*.

Liu, Liping, and Thomas Dietterich. 2014. Learnability of the superset label learning problem. In *International conference on machine learning*, 1629–1637. PMLR.

Lovász, L., and S. Vempala. 2007. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms* 30(3):307–358.

Lovász, L., and S. Vempala. 2007. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms* 30(3):307–358.

Lu, Tyler, and Craig Boutilier. 2011. Learning mallows models with pairwise preferences. In *Icml*.

Luce, R Duncan. 2012. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Lukasik, Michal, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. 2020. Does label smoothing mitigate label noise? In *International conference on machine learning*, 6448–6458. PMLR.

Lv, Jiaqi, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. 2020. Progressive identification of true labels for partial-label learning. In *International conference on machine learning*, 6500–6510. PMLR.

Maass, W., and G. Turan. 1994. How fast can a threshold gate learn? In *Computational learning theory and natural learning systems*, ed. S. Hanson, G. Drastal, and R. Rivest, 381–414. MIT Press.

Maddala, Gangadharrao S. 1986. *Limited-dependent and qualitative variables in econometrics*. 3, Cambridge university press.

———. 1987. Limited dependent variable models using panel data. *Journal of Human resources* 307–338.

Mallows, Colin L. 1957. Non-null ranking models. i. *Biometrika* 44(1/2):114–130.

Mammen, E., and A. B. Tsybakov. 1999. Smooth discrimination analysis. *Ann. Statist.* 27(6):1808–1829.

Mangoubi, O., and N. K. Vishnoi. 2019a. Nonconvex sampling with the metropolis-adjusted langevin algorithm. In *Conference on learning theory, COLT 2019*.

Mangoubi, Oren, and Nisheeth K Vishnoi. 2019b. Nonconvex sampling with the metropolis-adjusted langevin algorithm. In *Conference on learning theory*, 2259–2293. PMLR.

Mao, Cheng, and Yihong Wu. 2020. Learning mixtures of permutations: Groups of pairwise comparisons and combinatorial method of moments. *arXiv preprint arXiv:2009.06784*.

Martin Grötschel, Alexander Schrijver (auth.), László Lovász. 1993. *Geometric algorithms and combinatorial optimization*. 2nd ed. Algorithms and Combinatorics 2, Springer-Verlag Berlin Heidelberg.

Mason, J. C., and D. C. Handscomb. 2002. *Chebyshev polynomials*. CRC press.

Massart, P., and E. Nedelec. 2006. Risk bounds for statistical learning. *Ann. Statist.* 34(5):2326–2366.

Massart, Pascal. 2007. *Concentration inequalities and model selection*, vol. 6. Springer.

Massart, Pascal, and Élodie Nédélec. 2006. Risk bounds for statistical learning. *The Annals of Statistics* 34(5):2326–2366.

McDonald, Daniel. 2017. Minimax density estimation for growing dimension. In *Artificial intelligence and statistics*, 194–203.

Menon, A. K., B. Van Rooyen, and N. Natarajan. 2018. Learning from binary labels with instance-dependent noise. *Machine Learning* 107(8-10):1561–1595.

Nagarajan, Sai Ganesh, and Ioannis Panageas. 2019. On the Analysis of EM for truncated mixtures of two Gaussians. In *31st international conference on algorithmic learning theory (alt)*, 955–960.

Nasser, Rajai, and Stefan Tiegel. 2022. Optimal sq lower bounds for learning halfspaces with massart noise. *arXiv preprint arXiv:2201.09818*.

Natarajan, Balas K. 1989. On learning sets and functions. *Machine Learning* 4(1): 67–97.

Natarajan, Nagarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. *Advances in neural information processing systems* 26.

Nazarov, F. 2003a. On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric aspects of functional analysis (2001-2002)*, 169–187. Lecture Notes in Math., Vol. 1807, Springer.

Nazarov, Fedor. 2003b. On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a gaussian measure. In *Geometric aspects of functional analysis*, 169–187. Springer.

Negahban, Sahand, Sewoong Oh, and Devavrat Shah. 2017. Rank centrality: Ranking from pairwise comparisons. *Operations Research* 65(1):266–287.

Neyman, Jerzy. 1937. Smooth test for goodness of fit. *Scandinavian Actuarial Journal* 1937(3-4):149–199.

Nguyen, Nam, and Rich Caruana. 2008. Classification with partial labels. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining*, 551–559.

O'Donnell, R. 2014. *Analysis of boolean functions*. Cambridge University Press.

Owen, Art, et al. 1990. Empirical likelihood ratio confidence regions. *The Annals of Statistics* 18(1):90–120.

Owen, Art B. 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75(2):237–249.

———. 2001. *Empirical likelihood*. CRC press.

Padgett, WJ, and Diane T McNichols. 1984. Nonparametric density estimation from censored data. *Communications in Statistics-Theory and Methods* 13(13):1581–1611.

Paouris, G. 2006. Concentration of mass on convex bodies. *Geometric & Functional Analysis GAFA* 16(5):1021–1049.

Papadimitriou, Christos H. 1981. On the complexity of integer programming. *Journal of the ACM (JACM)* 28(4):765–768.

Papandreou, George, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the ieee international conference on computer vision*, 1742–1750.

Pearson, Karl. 1902. On the systematic fitting of frequency curves. *Biometrika* 2: 2–7.

Pearson, Karl, and Alice Lee. 1908. On the generalised probable error in multiple normal correlation. *Biometrika* 6(1):59–68.

Pisier, Gilles. 1986. Probabilistic methods in the geometry of banach spaces. In *Probability and analysis*, 167–241. Springer.

Qin, Zengyi, Jiansheng Chen, Zhenyu Jiang, Xumin Yu, Chunhua Hu, Yu Ma, Suhua Miao, and Rongsong Zhou. 2020. Learning fine-grained estimation of physiological states from coarse-grained labels by distribution restoration. *Scientific Reports* 10(1):1–10.

Renegar, James. 1992a. On the computational complexity and geometry of the first-order theory of the reals, part III: quantifier elimination. *J. Symb. Comput.* 13(3):329–352.

———. 1992b. On the computational complexity of approximating solutions for real algebraic formulae. *SIAM J. Comput.* 21(6):1008–1025.

Ristin, M., J. Gall, M. Guillaumin, and L. Van Gool. 2015. From categories to subcategories: Large-scale image classification with partial class label refinement. In *2015 ieee conference on computer vision and pattern recognition (cvpr)*, 231–239.

Rivest, R., and R. Sloan. 1994a. A formal model of hierarchical concept learning. *Information and Computation* 114(1):88–114.

Rivest, Ronald L, and Robert Sloan. 1994b. A formal model of hierarchical concept-learning. *Information and Computation* 114(1):88–114.

Rosenblatt, F. 1958. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65:386–407.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3):211–252.

Rebelo de Sá, Cláudio, Carla Rebelo, Carlos Soares, and Arno Knobbe. 2015. Distance-based decision tree algorithms for label ranking. In *Portuguese conference on artificial intelligence*, 525–534. Springer.

de Sá, Cláudio Rebelo, Carlos Soares, Arno Knobbe, and Paulo Cortez. 2017. Label ranking forests. *Expert systems* 34(1):e12166.

Schneider, Helmut. 1986. *Truncated and censored samples from normal populations*. Marcel Dekker, Inc.

Schrijver, Alexander. 1998. *Theory of linear and integer programming*. John Wiley & Sons.

Scott, Clayton, Gilles Blanchard, and Gregory Handy. 2013. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference on learning theory*, 489–511. PMLR.

Scott, David W. 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Shah, SM, and MC Jaiswal. 1966. Estimation of parameters of doubly truncated normal distribution from first four sample moments. *Annals of the Institute of Statistical Mathematics* 18(1):107–111.

Shalev-Shwartz, S., and S. Ben-David. 2014a. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, Shai. 2007. Online learning: Theory, algorithms, and applications. Ph.D. thesis, The Hebrew University of Jerusalem.

Shalev-Shwartz, Shai, and Shai Ben-David. 2014b. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

———. 2014c. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shamir, Ohad. 2018. Distribution-specific hardness of learning neural networks. *The Journal of Machine Learning Research* 19(1):1135–1163.

Shawe-Taylor, J., and N. Cristianini. 2000. *An introduction to support vector machines*. Cambridge University Press.

Simonoff, Jeffrey S. 2012. *Smoothing methods in statistics*. Springer Science & Business Media.

Sloan, R. H. 1988. Types of noise in data for concept learning. In *Proceedings of the first annual workshop on computational learning theory*, 91–96. COLT '88, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

———. 1992. Corrigendum to types of noise in data for concept learning. In *Proceedings of the fifth annual ACM conference on computational learning theory, COLT 1992*, 450.

———. 1996. *Pac learning, noise, and geometry*, 21–41. Boston, MA: Birkhäuser Boston.

Stute, Winfried, et al. 1993. Almost sure representations of the product-limit estimator for truncated data. *The Annals of Statistics* 21(1):146–156.

Szegö, G. 1967. *Orthogonal polynomials*. American Mathematical Society colloquium publications τ. 23, American Mathematical Society.

Taherkhani, Fariborz, Hadi Kazemi, Ali Dabouei, Jeremy Dawson, and Nasser M Nasrabadi. 2019. A weakly supervised fine label classifier enhanced by coarse supervision. In *Proceedings of the ieee international conference on computer vision*, 6459–6468.

Thomas, David R, and Gary L Grunkemeier. 1975. Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association* 70(352):865–871.

Tobin, James. 1958. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society* 24–36.

Touvron, Hugo, Alexandre Sablayrolles, Matthijs Douze, Matthieu Cord, and Hervé Jégou. 2020. Grafit: Learning fine-grained image representations with coarse labels. *arXiv preprint arXiv:2011.12982*.

Triguero, Isaac, Salvador García, and Francisco Herrera. 2015. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems* 42(2):245–284.

Tsybakov, A. 2004. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics* 32(1):135–166.

Tsybakov, Alexandre B. 2008. *Introduction to nonparametric estimation*. Springer Science & Business Media.

Van Engelen, Jesper E, and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine Learning* 109(2):373–440.

Van Rooyen, Brendan, and Robert C Williamson. 2017. A theory of learning with corrupted labels. *J. Mach. Learn. Res.* 18(1):8501–8550.

Van Zuylen, Anke, and David P Williamson. 2009. Deterministic pivoting algorithms for constrained ranking and clustering problems. *Mathematics of Operations Research* 34(3):594–620.

Vapnik, V. 1982. *Estimation of dependences based on empirical data: Springer series in statistics*. Berlin, Heidelberg: Springer-Verlag.

Vembu, Shankar, and Thomas Gärtner. 2010. Label ranking algorithms: A survey. In *Preference learning*, 45–64. Springer.

Vempala, Santosh, and John Wilmes. 2019. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and sq lower bounds. In *Conference on learning theory*, 3115–3117. PMLR.

Vershynin, R. 2018a. *High-dimensional probability: An introduction with applications in data science*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Vershynin, Roman. 2018b. *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge university press.

Vishnoi, Nisheeth K. 2021. *Algorithms for convex optimization*. Cambridge University Press.

Wand, Matt P, and M Chris Jones. 1994. *Kernel smoothing*. CRC press.

Wasserman, Larry. 2006. *All of nonparametric statistics*. Springer Science & Business Media.

Wen, Hongwei, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. 2021. Leveraged weighted loss for partial label learning. In *International conference on machine learning*, 11091–11100. PMLR.

Wolynetz, MS. 1979. Algorithm as 139: Maximum likelihood estimation in a linear model from confined and censored normal data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28(2):195–206.

Woodroofe, Michael, et al. 1985. Estimating a distribution function with truncated data. *The Annals of Statistics* 13(1):163–177.

Wu, Shanshan, Alexandros G. Dimakis, and Sujay Sanghavi. 2019. Learning distributions generated by one-layer relu networks. In *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, neurips 2019, december 8-14, 2019, vancouver, bc, canada*, ed. Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, 8105–8115.

Wu, Ximing. 2010. Exponential series estimator of multivariate densities. *Journal of Econometrics* 156(2):354–366.

Xu, Ning, Congyu Qiao, Xin Geng, and Min-Ling Zhang. 2021. Instance-dependent partial label learning. *Advances in Neural Information Processing Systems* 34.

Yan, S., and C. Zhang. 2017a. 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*, 1056–1066.

Yan, Songbai, and Chicheng Zhang. 2017b. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. *Advances in Neural Information Processing Systems* 30.

Yu, Fei, and Min-Ling Zhang. 2016. Maximum margin partial label learning. In *Asian conference on machine learning*, 96–111. PMLR.

Zhang, C., J. Shen, and P. Awasthi. 2020a. Efficient active learning of sparse halfspaces with arbitrary bounded noise. In *Advances in neural information processing systems, NeurIPS*.

Zhang, Chicheng, and Yinan Li. 2021. Improved algorithms for efficient active learning halfspaces with massart and tsybakov noise. In *Proceedings of thirty fourth conference on learning theory*, ed. Mikhail Belkin and Samory Kpotufe, vol. 134 of *Proceedings of Machine Learning Research*, 4526–4527. PMLR.

Zhang, Chicheng, Jie Shen, and Pranjal Awasthi. 2020b. Efficient active learning of sparse halfspaces with arbitrary bounded noise. In *Advances in neural information processing systems*, ed. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, vol. 33, 7184–7197. Curran Associates, Inc.

Zhang, Min-Ling. 2014. Disambiguation-free partial label learning. In *Proceedings of the 2014 siam international conference on data mining*, 37–45. SIAM.

Zhang, Min-Ling, Fei Yu, and Cai-Zhi Tang. 2017a. Disambiguation-free partial label learning. *IEEE Transactions on Knowledge and Data Engineering* 29(10):2155–2167.

Zhang, Y., P. Liang, and M. Charikar. 2017b. A hitting time analysis of stochastic gradient langevin dynamics. In *Proceedings of the 30th conference on learning theory, COLT 2017*, 1980–2022.

Zhang, Yuchen, Percy Liang, and Moses Charikar. 2017c. A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on learning theory*, 1980–2022. PMLR.

Zhou, Yangming, Yangguang Liu, Xiao-Zhi Gao, and Guoping Qiu. 2014a. A label ranking method based on gaussian mixture model. *Knowledge-Based Systems* 72: 108–113.

Zhou, Yangming, Yangguang Liu, Jiangang Yang, Xiaoqi He, and Liangliang Liu. 2014b. A taxonomy of label ranking algorithms. *J. Comput.* 9(3):557–565.

Zhou, Yangming, and Guoping Qiu. 2018. Random forest for label ranking. *Expert Systems with Applications* 112:99–109.