Efficient Statistical Inference
Under Sampling and Computational Constraints

by

Ankit Pensia

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: 04/24/2023

The dissertation is approved by the following members of the Final Oral Committee:
Po-Ling Loh, Professor, Pure Mathematics and Mathematical Statistics
Varun Jog, Assistant Professor, Pure Mathematics and Mathematical Statistics
Ilias Diakonikolas, Associate Professor, Computer Sciences
Jerry Zhu, Professor, Computer Sciences
Dimitris Papailiopoulos, Associate Professor, Electrical and Computer Engineering

Dedicated to my sister.

ACKNOWLEDGMENTS

दिल ना-उमीद तो नहीं नाकाम ही तो है
लम्बी है ग़म की शाम मगर शाम ही तो है

— फ़ैज़ अहमद फ़ैज़

I would like to express my deepest gratitude to Po-Ling Loh and Varun Jog, both of whom have played an indispensable role in my Ph.D. journey since the beginning. Their unwavering support, generosity with their time, and expert advice have been invaluable to me, both in technical and non-technical matters. Moreover, they have led by example and demonstrated how fulfilling a career in research can be. I will always cherish the fond memories of our research meetings where I was constantly in awe of Po-Ling's ability to spot fundamental connections between seemingly unrelated areas and Varun's ability to understand, explain, and then extend any result visually.[1] I would like to conclude with a quote from Po-Ling (in a response to my question in my first semester about what topic to focus on), which faithfully captures their curiosity-driven approach to research: "*It'll be fun to see where the research meandering takes us next ~~semester~~.*"

I consider myself incredibly fortunate to also have Ilias Diakonikolas as an advisor and mentor. He has been incredibly generous with his time, expertise, and mentorship, which have been instrumental in shaping my research trajectory. His breadth of knowledge and technical expertise, coupled with his ability to ask fundamental questions, is inspiring. Throughout the last four years, I have enjoyed collaborating with him on a range of topics and look forward to collaborations in the future.

Furthermore, I was privileged to have had the opportunity to collaborate closely with Daniel Kane, who has been an oracle personified. His expertise and insights have

[1] In the last three years, two extremely early-career researchers, A and S, have actively participated in our meetings; I heartily thank them for adding a sense of excitement, joy, and uncertainty to these meetings.

been critical in advancing my research, and I have learned a great deal from him.

I also thank Dimitris Papailiopoulos and Jerry Zhu for serving in my committee and providing feedback and support throughout my Ph.D. journey. Along with other members of UW-Madison such as Steve Wright, Rob Nowak, Kangwook Lee, Jelena Diakonikolas, and Sebastien Roch, their collaborative attitude, welcoming persona, and willingness to help have made UW-Madison a wonderful place to learn.

I am deeply grateful to Pranjal Awasthi and Satyen Kale for hosting me during Summer 2022 for the internship at Google. Their guidance, technical expertise, and patience were invaluable in providing me with a rich research experience. I am also appreciative of their insightful advice on professional career.

I am indebted to Rakesh K. Bansal and Piyush Rai for their kind and patient mentorship during my undergraudate days. I would like to specially thank Rakesh K. Bansal for introducing me to the beautiful world of probability and encouraging me to pursue research in this field.

Throughout the last six years, I have been fortunate to be surrounded by a wonderful group of students and postdocs. I would like to extend my heartfelt thanks to my first labmates Adrian, Amir, Ashley, Duzhe, Jinnian, Muni, Xiaomin, Zheng, and Zhili for creating a welcoming and warm environment from the very beginning. Being part of two research groups, I also interacted and consistently learned from the learning theorists: Jasper, Lisheng, Nikos, Sushrut, Thanasis, Vasilis, and Yuxin. I have also deeply enjoyed collaborating with Jasper, Sushrut, and Thanasis; I am grateful for their support and friendship through out the last two years and look forward to continued collaborations. Thanasis has been my most frequent student collaborator, and I have found all of our collaborations joyful and full of learning, especially the hours-long discussions on the virtual yellowboard.

In addition to the above collaborators, I would like to acknowledge my wonderful

coauthors on other projects: Rajat, Nikhil, Mingyuan, Alankrita, Alistair, Shashank, Alliot, and Harit.

I would also like to acknowledge the important role that reading groups have played in my learning experience. In addition to the people above at Madison, I thank Akshay and Alankrita for reading random stuff with me virtually.

Madison would not have felt a home without the friendship of Aditya, Anant, and Shashank. Shashank in particular has left a positive mark on virtually every part of my Ph.D. journey as an amazing friend, housemate, and collaborator. I also thank Anuja, Amrita, Ashish, Bhumesh, Gaurav, Kaushik, Muni, Pradyot, Rishabh, Shantanu, Swati, Tuan, and Vishnu for their invaluable friendship. I also thank many of them for numerous thrilling road trips.

I am also grateful to people from my undergraduate college and beyond who continue to be an important part of my life: Akshay, Anand, Devesh, Jitender, Prashun, Rajat, Ram, Rahul, Shubham, and Supratim.

Finally, I am indebted to my parents and my sister for all their love and sacrifices.

CONTENTS

ABSTRACT

Statistical inference has a long history of established algorithms with theoretical guarantees, but modern machine learning applications impose new statistical and computational constraints. These constraints include constraints on sampling such as poor quality datasets that deviate from idealized assumptions and constraints on computational resources such as time, memory, communication bandwidth, and privacy. These constraints can lead to a significant decrease in the performance of classical inference techniques, calling for new algorithmic solutions. In this thesis, we focus on fundamental statistical inference tasks such as mean estimation, linear regression, and hypothesis testing in the presence of aforementioned constraints.

The first part of the thesis focuses on statistical constraints on sampling and the challenges posed by real-world datasets that often do not conform to idealized assumptions. Many such datasets contain heavy tails, arbitrary outliers, and heterogeneity as opposed to the idealistic assumption of i.i.d. (sub-)Gaussian data. We develop practical statistical inference algorithms for mean estimation and linear regression with provable guarantees that are robust to these deviations. We achieve these results by developing algorithms that work under minimal structures on the data and proving that these structures hold with exponential probability, even under heavy-tailed data. In regimes where the existence of efficient algorithms is unknown, we give concrete evidence that efficient algorithms might indeed not exist by showing average-case computational lower bounds for a restricted family of algorithms.

The second part of the thesis focuses on computational constraints and the need to optimize algorithms for limited memory, communication bandwidth, and privacy in large-scale, distributed machine learning pipelines (in addition to optimizing for runtime). We begin by considering the space complexity of efficient algorithms for high-dimensional robust statistics, where we develop the first streaming algorithms with

near-optimal space complexity. Finally, we consider simple hypothesis testing under communication bandwidth and local privacy constraints, where we characterize the minmax optimal sample complexity and develop computationally-efficient algorithms.

1 INTRODUCTION

मैं जिसे ओढ़ता बिछाता हूँ
वो ग़ज़ल आप को सुनाता हूँ

— दुष्यंत कुमार

Statistical inference is a well-studied field with applications spanning engineering, operations research, and machine learning. With over a century of active research, it has well-known classical algorithms and associated theoretical guarantees. However, modern data science applications impose constraints, both statistical (e.g., quality and quantity of training data) and computational (e.g., limited time, memory, communication bandwidth, and privacy). Unfortunately, these constraints often lead to significant degradation of the performance of classical inference techniques, highlighting the need for novel algorithmic solutions.

This thesis aims to address this need by proposing practical statistical inference algorithms that offer provable guarantees under various resource constraints. Specifically, we will focus on two types of constraints: statistical and computational resources, which we will describe in more detail below.

- (Constraints on Sampling) Modern datasets are often so large that it is no longer possible to curate them carefully. This lack of curation causes many real-world datasets to be of poor quality, unlike what is assumed in theory. To elaborate on this point, much of classical statistical theory rests on data being "i.i.d." and "subgaussian", assumptions that rarely hold, if ever. Indeed, outliers, heterogeneity, and heavy tails are all fixtures of modern real-world datasets. This vast gap in data quality between theory and practice has dire consequences: algorithms developed for the idealized data fail dramatically upon even a slight quality degradation. This

brittleness, a consequence of "model misspecification", is systematically studied under the umbrella of "robust statistics"[HR09].

- (Constraints on Computational Resources) The design of inference algorithms under computational constraints has traditionally focused on optimizing the runtime. However, the emergence of large-scale, distributed machine learning pipelines has now brought to the fore the importance of optimizing for limited memory and communication bandwidth. In settings like distributed learning on mobile phones and edge devices, constraints on memory, communication bandwidth, or privacy render previously developed estimators completely inapplicable. Consequently, there is a need to rethink the algorithmic design philosophy for these situations.

This thesis aims to develop practical statistical inference algorithms with provable guarantees under these constraints. Our focus in this thesis is on fundamental statistical inference tasks that are both widely used and capture the challenges of these constraints.

1.1 Organization

This thesis is divided in two parts. The first part, "Constraints on Sampling: Outliers, Heavy-Tails, and Heterogeneity" (Part I), focuses on topics pertaining to constraints on sampling. This part is structured as follows:

- Chapter 2 serves as an introduction on the challenges posed by constraints on sampling. Assuming minimal background knowledge, we describe the inference tasks that we consider and the constraints on sampling. For each of these tasks and the constraints, we also provide a brief summary of our results.

- Chapter 3 is based on "Outlier Robust Mean Estimation with Subgaussian Rates via Stability " [DKP20] (published in NeurIPS 2020), joint with Ilias Diakonikolas

and Daniel M. Kane.

- Chapter 4 is based on " Outlier-Robust Sparse Mean Estimation for Heavy-Tailed Distributions" [DKLP22] (published in NeurIPS 2022), joint with Ilias Diakonikolas, Daniel M. Kane, and Jasper C. H. Lee.

- Chapter 5 is based on "Robust regression with covariate filtering: Heavy tails and adversarial contamination" [PJL20b], joint with Varun Jog and Po-Ling Loh.

- Chapter 6 is based on "Statistical Query Lower Bounds for List-Decodable Linear Regression" [DKPPS21] (published in NeurIPS 2021), joint with Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Alistair Stewart.

- Chapter 7 is based on "Mean estimation for entangled single-sample distributions" [PJL19b] (published in ISIT 2019) and "Estimating location parameters in sample-heterogeneous distributions" [PJL19a] (published in Information and Inference: a Journal of the IMA), joint with Varun Jog and Po-Ling Loh.

The second part of thesis, "Constraints on Computational Resources: Communication, Memory, and Privacy" (Part II), focuses on topics related to constraints on computational resources. The content in this part is structured as follows:

- Chapter 8 provides an overview of the constraints on computational constraints that we consider. The chapter is designed to be accessible to readers with minimal background knowledge, and we begin by introducing the inference tasks that we consider and the constraints on sampling that are relevant to these tasks. In addition, we provide a concise summary of the results we obtained for each task and constraint.

- Chapter 9 is based on "Streaming Algorithms for High-Dimensional Robust Statistics" [DKPP22] (published in ICML 2022), joint with Ilias Diakonikolas, Daniel M. Kane, and Thanasis Pittas.

- Chapter 10 is based on "Simple Binary Hypothesis Testing under Communication Constraints" [PLJ22] (published in ISIT 2022) and "Communication-constrained hypothesis testing: Optimality, robustness, and reverse data processing inequalities " [PJL22], joint with Varun Jog and Po-Ling Loh.

- Chapter 11 is based on "Simple Binary Hypothesis Testing under Local Differential Privacy and Communication Constraints" [PAJL23], joint with Amir Asadi, Varun Jog, and Po-Ling Loh.

In addition to these works, the author also worked on topics that are not part of this thesis: generalization error of sequential noisy algorithms [PJL18], robustness to test-time attacks [PJL20a], Bayesian approaches to multi-label learning [PPMZR19], concentration inequalities [BP22], distribution testing [DKP23], and sum-of-squares approaches to robust sparse estimation [DKKPP22a; DKKPP22b].

Part I

Constraints on Sampling:

Outliers, Heavy-Tails, and Heterogeneity

2 OVERVIEW OF RESULTS: CONSTRAINTS ON SAMPLING

All models are wrong, but some are useful.

—George E. P. Box

This chapter gives a comprehensive introduction to the difficulties that arise when the sampling process results in low-quality data, such as outliers, heavy tails, and heterogeneity. We explore fundamental inference tasks, such as mean estimation and linear regression, under these sampling constraints.

To provide a thorough understanding of each inference task and the corresponding constraints, we outline the problem statement, present a brief review of relevant literature, and conclude with a discussion of our contributions. Through our analysis, we aim to highlight the unique statistical and computational challenges faced in each scenario and the research questions that motivated us.

We divide this chapter into two sections: Section 2.1, which addresses the typical situation where outliers represent a small fraction of the data, and Section 2.2, which focuses on cases where low-quality data constitutes the majority.

2.1 Outliers and Heavy-Tailed Distribution

Adversarial outliers and heavy-tailed distributions constitute two significant challenges typical to real-world datasets.

Outliers *Outliers* are data points significantly different from the rest. Errors in measurement, data entry, or data analysis can cause outliers. Let us briefly discuss outliers' central role in statistical inference using the following motivations.

Consider a scientific experiment where we collect measurements and fit a model to these measurements to identify the parameters of interest. In such a setting, outliers

may correspond to measurement errors corresponding to malfunctioning pieces of equipment, incorrect experiment procedures, or errors in data logging. These errors are not "random" but rather systematic. As a result, they can severely bias the estimate if not adequately handled.

Alternatively, one may argue that the perfect dataset is unlikely to exist. A perfect dataset is when all observations come from a particular model in our model class; this setting is known as "well-specified" in the literature. Well-specification is a strong assumption because the real-world dataset is often too complicated and may not *exactly* satisfy the clean, simplified statistical model. As George Box famously said, "All models are wrong, but some are useful".

We now formally define the contamination model that places no restriction on outliers except the fact that the number of outliers is small:

Definition 2.1.1 (Strong Contamination Model)**.** *Given a parameter $0 < \epsilon < 1/2$ and a family of distributions $\mathcal{D}$ on $\mathbb{R}^d$, the* adversary *operates as follows: The algorithm specifies the number of samples n, and n samples are drawn from some unknown $D \in \mathcal{D}$. The (computationally unbounded) adversary can inspect the samples, remove up to ϵn of them and replace them with arbitrary points. This modified set of n points is then given as input to the algorithm. We say that a set of samples is ϵ-*corrupted *if generated by the above process.*

Observe that the corrupted dataset is neither independent nor identically distributed from D (the dataset is not independent because the outliers can depend on the inliers.)

We will be primarily interested in the regime where ϵ, the fraction of outliers, is a small constant independent of the dimension d.

It is easy to see that many commonly used algorithms are not robust to even a single outlier (e.g., the sample mean to estimate the mean, ordinary least squares algorithm for linear regression). Developing algorithms that are robust to outliers have been systematically studied in the field of "robust statistics", with pioneering

contributions from Huber [Hub64; Hub65]. Since then, many statistically-efficient procedures have been developed. Until recently, these procedures were computationally prohibitive, leading to an unfortunate dichotomy: existing robust algorithms were either computationally-efficient or statistically efficient, but not simultaneously.

In a series of papers, the first robust polynomial-time algorithms for several high-dimensional estimation tasks were developed in [LRV16; DKKLMS16; DKKLMS17]. In particular, they attained near-optimal rates for light-tailed distributions on many tasks, but it was unclear if these algorithms were optimal for heavy-tailed distributions (defined below).

Heavy-Tailed Distributions One of the most common assumptions in the literature on non-asymptotic statistical inference is that the data has sub-Gaussian tails.

However, many natural distributions, for example, power-law distributions, do not satisfy this assumption. Since sub-gaussianity is equivalent to all moments of a distribution being controlled appropriately, a natural way to define heavy-tailed distributions is to require only a small number of low-degree moments being controlled, motivating the following definition:

Definition 2.1.2 (Heavy-tailed Distributions (informal))**.** *We say a distribution D over $\mathbb{R}^d$ is heavy-tailed if the low-degree moments of the distribution exist along each unit vector.*

Standard estimators, which are optimal for subgaussian distributions, could be severely suboptimal for heavy-tailed distributions. This phenomenon was highlighted in a seminal paper by Catoni [Cat12] for the problem of univariate mean estimation. In particular, suppose the data distribution D is a univariate distribution with mean μ and variance σ^2. If D were a Gaussian (or a subgaussian) distribution, then the sample mean of n i.i.d. points from D, $\widehat{\mu}_{\text{sample-mean}}$, would have been an optimal estimator, obtaining

the following guarantee: with probability $1-\tau$,

$$|\widehat{\mu}_{\text{sample-mean}} - \mu| \lesssim \sigma\sqrt{\log(1/\tau)/n}\ .$$

However, there is a heavy-tailed distribution D with mean μ and variance σ^2, such that with probability, $1-\tau$,

$$|\widehat{\mu}_{\text{sample-mean}} - \mu| \lesssim \sigma\sqrt{1/(\tau n)}\ .$$

This sub-optimal performance of the sample mean on heavy-tailed distributions begs whether this poor performance of $\widehat{\mu}_{\text{sample-mean}}$ is inherent in the mean estimation of heavy-tailed distributions. Surprisingly, the answer is no! There are multiple estimators $\widehat{\mu}$ (see, for example, [Cat12; LV20; BCL13; OO19]) that achieve the rate $|\widehat{\mu} - \mu| \lesssim \sigma\sqrt{\log(1/\tau)/n}$ for heavy-tailed distributions.

A major thrust of this thesis is on obtaining similar performance guarantees for high-dimensional estimation tasks. As we will see shortly, new challenges (both statistical and computational) appear in the high-dimensional regime.

2.1.1 Inference Task 1: Multivariate Mean Estimation

Let D be an (unknown) heavy-tailed distribution over $\mathbb{R}^d$ with (unknown) mean μ and (unknown) covariance Σ. Consider the problem of estimating the mean μ from the samples of D. As discussed earlier, it is unrealistic in many settings to assume that the dataset is i.i.d. and the distribution D is light-tailed. We are thus led to the following problem:

Inference Task 1 (Outlier-Robust Heavy-Tailed Mean Estimation (in Euclidean Norm))**.** *Let D be an (unknown) heavy-tailed distribution over $\mathbb{R}^d$ with mean μ and covariance Σ. Given*

a set of ϵ-corrupted samples in $\mathbb{R}^d$ from D, ϵ, and τ, compute an estimate $\widehat{\mu}$ such that with probability $1-\tau$, we have that $\|\widehat{\mu}-\mu\|_2$ is small.

Observe that there are no further restrictions on D beyond finite covariance, and there could be ϵ-fraction of outliers in the data. For simplicity, and without loss of generality, we will assume that D is supported on a bounded domain of size $\sqrt{\mathrm{tr}(\Sigma)/\left(\epsilon+\frac{\log(1/\tau)}{n}\right)}$.

Inference Task 1 is a fundamental problem in robust statistics with far-reaching consequences. For example, robust mean estimation is a crucial sub-routine for general robust stochastic optimization (cf. [PSBR20; DKKLSS19]).

Before discussing our contributions, let us first discuss the information-theoretic best-possible result. It can be shown that the optimal rate is the following: given n samples, with probability $1-\tau$,

$$\|\widehat{\mu}-\mu\|_2 \asymp \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}+\sqrt{\frac{\|\Sigma\|_{\mathrm{op}}\log(1/\tau)}{n}}+\sqrt{\|\Sigma\|_{\mathrm{op}}\epsilon}\,. \tag{2.1}$$

That is, a (computationally-inefficient) algorithm achieves this rate, and all algorithms must have at least this much error rate.

Regarding historical development, two families of computationally-efficient algorithms tried to match Equation (2.1).

Stability-based algorithms First were the *stability-based algorithms*, developed in the field of *algorithmic robust statistics*, to optimize the dependence on ϵ [DKKLMS17; DK19]. We now define the notion of stability:

Definition 2.1.3 (Stability)**.** *We say a dataset $T\subset\mathbb{R}^d$ is (ϵ,δ)-stable with respect to $\mu\in\mathbb{R}^d$ and $\sigma^2\in\mathbb{R}_+$, if for all $T'\subseteq T$ with $|T'|\geq(1-\epsilon)|T|$, the following holds:*

- *(First moment)* $\|(1/|T'|)\sum_{x\in T'}x-\mu\|_2\leq\sigma\delta$,
- *(Second moment)* $\|(1/|T'|)\sum_{x\in T'}(x-\mu)(x-\mu)^\top-\sigma^2 I\|_{\mathrm{op}}\leq\sigma^2\delta^2/\epsilon$.

By stability-based algorithms, we mean the following:

Definition 2.1.4 (Stability-based algorithms, informal)**.** *We say an algorithm $\mathcal{A}$ is stability-based if it satisfies the following guarantee: Let S be an (ϵ, δ)-stable set with respect to μ and σ^2, and let T be an $O(\epsilon)$-corrupted version of S. Then, given T and ϵ, the algorithm $\mathcal{A}$ efficiently computes*[2] *$\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 \lesssim \sigma\delta$.*

Many different kinds of stability-based algorithms have been developed in literature: convex programming, iterative filtering, gradient descent, and nearly-linear time algorithms. These algorithms are not just polynomial-time but also practically implementable, as many of these are spectral-based.

Going back to Inference Task 1 and Equation (2.1), we see that stability-based algorithms can be used as follows: Let S be a set of n i.i.d. samples from a heavy-tailed distribution and suppose that S is $(\epsilon', \delta_{n,\epsilon,\tau,\Sigma})$-stable with respect to μ and $\|\Sigma\|_{\text{op}}$ with probability $1 - \tau$. Then, as long as $\epsilon' > \epsilon$, stability-based algorithms on Inference Task 1 obtain the rate of $O\left(\sqrt{\|\Sigma\|_{\text{op}}}\delta_{n,\epsilon,\tau,\Sigma}\right)$. Thus, a very promising way to achieve Equation (2.1) is to show that $\sqrt{\|\Sigma\|_{\text{op}}}\delta_{n,\epsilon,\tau,\Sigma}$ matches the right-hand side of Equation (2.1) up to constants. This is a reasonable wish since it was shown in [DKKLMS17] for constant τ and given $\tilde{O}(d/\epsilon)$ samples, the dependence on ϵ was correct, i.e., $\sqrt{\epsilon}$.

However, as we saw earlier, the dependence of δ on the failure probability τ was unclear. The best-known upper-bound at that time was: with probability $1 - \tau$, S is (ϵ, δ)-stable with respect to μ and $\|\Sigma\|_{\text{op}}$ with

$$\sqrt{\|\Sigma\|_{\text{op}}}\delta \lesssim \sqrt{\frac{\text{tr}(\Sigma)\log(d/\tau)}{n}} + \sqrt{\|\Sigma\|_{\text{op}}\epsilon}\,. \tag{2.2}$$

Thus, the existing upper bound on the statistical error rate for the stability-based algorithms in Equation (2.2) was far from the optimal rate in Equation (2.1). This leads to

[2]For simplicity, we limit our discussion to deterministic algorithms, but randomized algorithms are also applicable if their failure probability is exponentially small in ϵT.

the following question:

Question 1. *Do the stability-based algorithms achieve (near)-optimal error guarantees for Inference Task 1?*

In particular, does the following hold?

Question 2. *Is a set of i.i.d. data from a heavy-tailed distribution stable with high probability?*

We now turn to the second family of the algorithms that we will call *Median-of-means algorithms* (for lack of a better word).

Median-of-means algorithms This algorithm family was initially focused on the uncontaminated (i.i.d.) heavy-tailed data, i.e., with $\epsilon = 0$. First, the data is randomly divided into k blocks of equal size (discarding the data if needed), and we take the sample mean of each of these blocks. Let $\{z_1, \dots, z_k\}$ be these data points. Then, the algorithm computes an (appropriately defined) high-dimensional median of $z_1, \dots, z_k$, thus termed "median-of-means". The univariate median of means has been known to achieve the optimal rate in the (uncontaminated) univariate setting for decades [NY83].

In a seminal paper, Lugosi and Mendelson [LM19d] proposed a (computationally-inefficient) high-dimensional analog of the median that achieves the rate in Equation (2.1) (in the uncontaminated setting). Later, Hopkins [Hop20] (see also [CFB19]) proposed a computationally-efficient version of the estimator in [LM19d], achieving the following: For $\epsilon = 0$,

$$\|\widehat{\mu}_{\text{median-of-means}} - \mu\|_2 \lesssim \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_{\text{op}} \log(1/\tau)}{n}}\,. \tag{2.3}$$

Later, [DL22b] observed that the median-of-means algorithms naturally inherit robustness properties from the median step (if the number of blocks, k, is sufficiently larger than the number of outliers). Thus, combining their observation, we obtain that even in

the presence of outliers, the estimator in Equation (2.3) continues to obtain the following rate:

$$\|\widehat{\mu}_{\text{median-of-means}} - \mu\|_2 \lesssim \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_{\text{op}} \log(1/\tau)}{n}} + \sqrt{\|\Sigma\|_{\text{op}} \epsilon}\,. \tag{2.4}$$

Although this rate matches the information-theoretic rate in Equation (2.1) in polynomial-time, these algorithms are still impractical as they rely on large semidefinite programs[3] which are slow in practice (Recall that stability-based algorithms were practical). This leads to the following question:

Question 3. *Is achieving the rate in Equation (2.1) possible using practical (not just polynomial-time) algorithms?*

At a more fundamental level, the question arises whether these two families of algorithms, developed disjointly, are related.

Question 4. *Are stability-based algorithms related to the median-of-means family of algorithms? If so, how?*

Moreover, the stability-based algorithms are known to be adaptive to the distribution's tails. For example, for distributions with bounded k-th moments for $k \geq 2$, the information-theoretic dependence on ϵ is $\epsilon^{1-1/k}$, which is much better than $\sqrt{\epsilon}$ for $k > 2$. Moreover, when the covariance is spherical, the stability-based algorithms were known to achieve this rate (in the constant failure probability regime). However, even for Gaussian distributions, median-of-means algorithms are inherently stuck at $\sqrt{\epsilon}$.

[3] We mention that a spectral algorithm within the median-of-means framework was developed in [LLVZ20]. Still, the proposed algorithm, while spectral, is more complicated than the known stability-based algorithms, and we are unaware of any practical implementation of this algorithm. As we will see later, one can use more practical (stability-based) algorithms instead.

Question 5. *Is there a computationally-efficient algorithm (hopefully, practical as well) that adapts to the tails of the input distribution in Inference Task 1, with better dependence on ϵ as the tails get progressively lighter?*

Our Contributions In Chapter 3, we will answer Questions 1 to 5 as follows:

1. (Question 2) Unfortunately, if S is a set of i.i.d. data points from a heavy-tailed distribution, then with high probability, S may not be stable (with the optimal or near-optimal parameter δ). In particular, the rate in Equation (2.2) is roughly tight in the worst-case; see Example 3.2.2. Thus, the answer to Question 2 is negative.

2. (Question 1) However, perhaps surprisingly, we will show that, with high probability, S contains *a large subset* S' that is stable with a near-optimal parameter δ. Moreover, the existence of a large stable subset is sufficient for stability-based algorithms to succeed, and thus the answer to Question 1 is affirmative. Thus, stability-based algorithms achieve near-optimal error for Inference Task 1.

3. (Question 4) Next, we study the relationship between the median-of-means and stability. All of the existing median-of-means algorithms rely on a particular structure on the input data, say $\mathcal{E}$, which holds with high probability. We show that this same structure $\mathcal{E}$ implies that the data is stable with the optimal parameters. That is, median-of-means algorithms are also using stability in disguise.

4. (Question 3) Leveraging this freshly-established connection, we show that applying stability-based algorithms (after a simple preprocessing) achieves the optimal error Equation (2.1) practically.

5. (Question 5) We show that the stability parameter of the data improves as the tails of the distribution get lighter (using the technical insights gained by proving

Question 1). Thus, stability-based algorithms strictly improve over median-of-means algorithms as the tails of the distributions get progressively lighter (for spherical distributions).

2.1.2 Inference Task 2: Structured (Sparse) Mean Estimation

In the context of statistical inference, additional information about the unknown mean of the distribution is often available. For example, many natural signals are sparse in the appropriate basis, such as images in wavelet basis [EK12; HTW15]. This observation motivates the problem of estimating the mean parameter under *structured* assumptions, where the parameter of interest, denoted as μ, satisfies some additional structure. One of the most commonly studied structures in theory and practice is sparsity [EK12; HTW15]. Specifically, the goal is to estimate μ accurately while utilizing the prior knowledge that only a few coordinates of μ are non-zero, defined formally below:

Definition 2.1.5 (Sparsity)**.** *We say a vector $x \in \mathbb{R}^d$ is k-sparse if at most k of the coordinates of x are non-zero.*

The benefits of sparsity in the (sub)-Gaussian i.i.d. data regime are widely recognized in statistics. Specifically, given n samples in $\mathbb{R}^d$ from $\mathcal{N}(\mu, I)$ with a k-sparse mean μ, the soft-thresholding estimator gives an estimate $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 \lesssim \sqrt{\frac{k \log(d/k)}{n}} + \sqrt{\log(1/\tau)/n}$. This error rate is much better than the unstructured setting, where the information-theoretic error is much higher, $\Theta\left(\sqrt{d/n} + \sqrt{\log(1/\tau)/n}\right)$.

Despite the promising results achieved by the aforementioned estimator, it is known to be sensitive to outliers and heavy-tailed data. This observation has motivated the problem of outlier-robust heavy-tailed sparse mean estimation, formulated as follows:

Inference Task 2 (Outlier-Robust Heavy-Tailed Sparse Mean Estimation)**.** *Let D be an (unknown) heavy-tailed distribution over $\mathbb{R}^d$ with a k-*sparse *mean μ and covariance Σ. Given*

a set of ϵ-corrupted samples in $\mathbb{R}^d$ from D, sparsity parameter k, ϵ and τ, compute an estimate $\widehat{\mu}$ such that with probability $1-\tau$, we have that $\|\widehat{\mu}-\mu\|_2$ is small.

Information-theoretically, it is possible to achieve the following rate[4]: for $n \gtrsim k\log(d/k)+\log(1/\tau)$, there is a (computationally-inefficient) estimator $\widehat{\mu}$ such that

$$\|\widehat{\mu}-\mu\|_2 \leq \sqrt{\frac{k\log(d/k)\|\Sigma\|_{\text{op}}}{n}} + \sqrt{\frac{\|\Sigma\|_{\text{op}}\log(1/\tau)}{n}} + \sqrt{\|\Sigma\|_{\text{op}}\epsilon}. \tag{2.5}$$

This leads to the question of whether this rate can be achieved computationally efficiently. Efficient algorithms for robust sparse mean estimation were developed in [BDLS17; DKKPS19; CDKGGS22; DKKPP22b]. However, those algorithms focused on light-tailed distributions (As we will show later, these algorithms failed on heavy-tailed data).

The best-known computationally-efficient algorithm at the time was rather naive: apply a robust univariate algorithm coordinate-wise and then threshold the resulting coordinate-wise to obtain a k-sparse vector $\widehat{\mu}_{\text{coordinate-wise}}$. This algorithm achieved the following rate: with probability $1-\tau$ for n large enough,

$$\|\widehat{\mu}_{\text{coordinate-wise}}-\mu\|_2 \lesssim \sqrt{\frac{k\log(d)\|\Sigma\|_{\text{op}}}{n}} + \sqrt{\frac{k\|\Sigma\|_{\text{op}}\log(1/\tau)}{n}} + \sqrt{k\|\Sigma\|_{\text{op}}\epsilon}\,. \tag{2.6}$$

This rate is much worse than the rate in Equation (2.5) because the dependence on both ϵ and $\log(1/\tau)$ worsens as k increases. Unfortunately, it might not be possible to achieve the rate in Equation (2.5) in a computationally-efficient manner. For example, Equation (2.5) implies that for $\epsilon=\Theta(1)$ and $\tau=\Theta(1)$, it $\Theta\left(k\log(d)\right)$ samples suffice to get constant error. However, [DKS17; BB20] have given evidence that even if D is isotropic Gaussian, computationally-efficient algorithms need at least roughly k^2 samples; this phenomenon is known as *information-computation gap* in the literature. Thus, this leads to the following

[4]To the best of our knowledge, the optimal statistical rate for sparse mean estimation has not been established for heavy-tailed distributions (even without outliers).

question:

Question 6. *What can computationally-efficient algorithms achieve for Inference Task 2? In particular, are there computationally-efficient algorithms that uses* $\mathrm{poly}(k, \log d, \log(1/\tau), 1/\epsilon)$ *samples and achieves error* $O(\sqrt{\epsilon})$*?*

We now describe our contributions.

Our Contributions In Chapter 4, we answer Question 6 by developing the first computationally-efficient algorithm for Inference Task 2 that improves upon Equation (2.6). We will show that under an additional (mild) assumption of bounded fourth moments along axis directions, there is a computationally-efficient algorithm $\widehat{\mu}$ with the following guarantee: for $n \gtrsim (k^2 \log d + \log(1/\tau))/\epsilon$, then with probability $1 - \tau$,

$$\|\widehat{\mu} - \mu\|_2 \lesssim \sqrt{\|\Sigma\|_{\mathrm{op}} \epsilon}\,. \tag{2.7}$$

This guarantee should be compared with the error rate in Equation (2.5), which requires $(k \log(d/k) + \log(1/\tau))/\epsilon$ many samples to achieve the same error. Thus, the proposed algorithm requires k^2 samples instead of the statistically optimal rate of k samples. However, as mentioned earlier, the information-computation tradeoffs established in [DKS17; BB20] require that computationally-efficient algorithms have a quadratic dependence on k. Thus, the proposed algorithm achieves the near-optimal error among computationally-efficient algorithms qualitatively.

2.1.3 Inference Task 3: Linear Regression

Our focus now turns to a supervised learning problem involving heavy-tailed data and outliers. One of the most fundamental supervised learning problems is linear regression, which aims to learn a linear function that maps input features (covariates) to output

labels (responses). In line with the theme of this section, we consider the following variant of the problem:

Inference Task 3 (Outlier-Robust Heavy-Tailed Linear Regression)**.** *Let (X, y) be jointly distributed on $(\mathbb{R}^d, \mathbb{R})$ according to the distribution D as follows: X has mean zero, identity covariance, and bounded low-degree moments, and conditioned on $X = x$, Y is distributed as $x^\top \beta^* + Z$ for an unknown vector $\beta^* \in \mathbb{R}^d$ and independent zero-mean unit-variance noise Z from a heavy-tailed distribution. Given a set of ϵ-corrupted samples from D (both covariates (x) and responses (y) can be corrupted), ϵ, failure probability τ, compute an estimate $\widehat{\beta}$ such that $\|\widehat{\beta} - \beta^*\|_2$ is small.*

It is worth noting that in this problem, both the covariates and the responses could be heavy-tailed and corrupted. For simplicity of presentation, we will consider only the heavy-tailed setting without any contamination for the present discussion. Despite both the covariates and responses being heavy-tailed, it is possible to achieve the following subgaussian performance: for $n \gtrsim d + \log(1/\tau)$, with probability $1 - \tau$

$$\|\widehat{\beta} - \beta^*\|_2 \asymp \sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} \tag{2.8}$$

However, the standard ordinary least squares (OLS) estimator, $\|\widehat{\beta}_{\text{OLS}}\|$, has much worse performance: there exists a heavy-tailed distribution such that under the same settings, $\|\widehat{\beta}_{\text{OLS}} - \beta^*\|_2 \gtrsim \sqrt{\frac{d}{n\tau}}$

To bridge this performance gap, [CHKRT20] used sum-of-squares hierarchy to develop an estimator $\widehat{\beta}$ that achieved the rate in Equation (2.8) provided $n \gtrsim d\sqrt{\log(1/\tau)}$. While this represents a substantial improvement, the dependence on the failure probability on the sample complexity is still multiplicative, raising the question of whether further improvements can be made.

Question 7. *Are there computationally-efficient algorithms that attain the (near)-optimal rate in Equation (2.8)?*

We note that the main challenge in answering Question 7 is handling the lax assumptions on the covariates [HR09]. Indeed, consider the weaker corruption model, which we call the *label corruption model*, where only the responses (Y) are allowed to be corrupted and/or heavy-tailed, but the covariates (X) are required to be uncontaminated and subgaussian. Label corruption model is a much easier corruption model, and in fact, several classic estimators are known to be robust to this model, for example, Huber loss regression [Hub73], least trimmed squares [Rou84], et cetera.

Given the recent advances in algorithmic robust statistics for high-dimensional data (for example, Inference Task 1), a natural question is whether we can combine the insights from these recent advances and the classical literature on the label corruption model to solve the strong contamination model. We arrive at the following conceptual question:

Question 8. *Can we reduce the strong contamination model to the label corruption model for robust regression using recent advances in algorithmic robust statistics?*

In addition to being a conceptual question, answering Question 8 in the affirmative is likely to have practical implications: it would offer modularity and simplicity in the algorithms addressing the strong contamination model.

We now describe our contributions:

Our Contributions In Chapter 5, we answer Question 7 by developing a computationally efficient algorithm with near-optimal error guarantees. Our proposed algorithm is surprisingly simple and also addresses Question 8. Specifically, our approach involves running stability-based algorithms on the covariates to remove a certain fraction of points deemed outliers. We then apply classical estimators that are robust to label corruption,

such as Huber regression, to the remaining data points. This modular approach allows us to leverage existing algorithmic tools for robust statistics and reduce the problem of strong contamination to the simpler label corruption model.

2.2 Inference when Good Quality Data is in Minority

In the previous section, we discussed the challenges posed by outliers and heavy-tailed distributions in designing inference algorithms for poor-quality statistical resources. However, our discussion was limited to the setting where the fraction of outliers, ϵ, is smaller than $1/2$, and the fraction of "good data" was larger than $1/2$. However, there are scenarios where the fraction of "good data" is less than $1/2$, for example, in crowdsourcing, where most participants could be unreliable.

In such scenarios, without further assumptions or modifications of the problem, it is impossible to output a reasonable estimate because the model is not even identifiable: The input distribution might be a mixture of k-many multiple components, with each component being far from all other components. There is no way to identify the "true" component of the mixture. Thus, one must modify the problem statement to handle the scenarios where the fraction of "good data" is less than $1/2$. The following two subsections will consider two separate modifications that make the problem well-posed.

2.2.1 Inference Task 4: List-decodable Linear Regression

As mentioned earlier, the inference task is not identifiable when the inliers are in the minority because if there are only α-fraction of inliers for $\alpha \leq 1/2$, then there can be multiple hypotheses, $\lfloor 1/\alpha \rfloor$-many in fact, such that each hypothesis is consistent with some α-fraction of the data. To address the identifiability issue, the concept of *list-decodable learning* was proposed and studied in [BBV08; CSV17]. In this setting, the

algorithm to allowed to output a small list of hypotheses such that at least one is close to the correct one. Formally, the list-decodable learning model is defined as follows:

Definition 2.2.1 (List-Decodable Learning)**.** *Given a parameter $0 < \alpha < 1/2$ and a distribution family $\mathcal{D}$ on $\mathbb{R}^d$, the algorithm specifies $n \in \mathbb{Z}_+$ and observes n i.i.d. samples from a distribution $E = \alpha D + (1-\alpha)N$, where D is an unknown distribution in $\mathcal{D}$ and N is arbitrary. We say D is the distribution of inliers, N is the distribution of outliers, and E is an $(1-\alpha)$-corrupted version of D. Given sample access to an $(1-\alpha)$-corrupted version of D, the goal is to output a "small" list of hypotheses $\mathcal{L}$ at least one of which is (with high probability) close to the target parameter of D.*

Many inference tasks have been studied in this model, for example, mean estimation, linear regression, and covariance estimation. Our focus in this section is on the list-decodable linear regression, defined below:

Inference Task 4 (List-Decodable Linear Regression)**.** *Fix $\sigma > 0$. For $\beta \in \mathbb{R}^d$, let D_β be the distribution over (X, y), $X \in \mathbb{R}^d$, $y \in \mathbb{R}$, such that $X \sim \mathcal{N}(0, I_d)$ and $y = \beta^\top X + \eta$, where $\eta \sim \mathcal{N}(0, \sigma^2)$ independently of X. Given sample access to an $(1-\alpha)$-corrupted version of D_{β^*} (for an unknown β^* with norm less than 1), the goal is to output a "small" list of vectors $\mathcal{L}$ at least one of which is (with high probability) close to the β^*, i.e., $\min_{w \in \mathcal{L}} \|w - \beta^*\|_2$ is small.*

List-decodable linear regression was first studied in [KKK19; RY20a]. It can be shown that the information-theoretic error rate for this problem is $\tilde{\Theta}(\sigma/\alpha)$, which can be achieved with an $(1-\alpha)$-corrupted set of size $\mathrm{poly}(d/\alpha)$ [DKPPS21; KKK19]. However, the known algorithms for this task, developed in [KKK19; RY20a] required $d^{\mathrm{poly}(1/\alpha)}$ many samples to succeed, which is much larger than $\mathrm{poly}(d/\alpha)$. This gap in the sample complexity of the existing computationally-efficient algorithms and the information-theoretic sample naturally leads to the question of whether there are better algorithms:

Question 9. *Are there computationally-efficient algorithms for Inference Task 4 that use* $\mathrm{poly}(d/\alpha)$ *samples?*

As we show below, the answer is likely to be no.

Our Contributions In Chapter 6, we show that the answer to Question 9 is negative (in a restricted family of algorithms) and Inference Task 4 exhibits information-computation tradeoff. That is, any computationally-efficient statistical query algorithm for Inference Task 4 must use $d^{\mathrm{poly}(1/\alpha)}$ samples (even to get accuracy less than a small enough constant). Thus, our results imply that the algorithms in [KKK19; RY20a] are qualitatively the best possible.

2.2.2 Inference Task 5: Mean Estimation under Sample-Heterogeneity

We now consider a different learning model that focuses on the regime when the quality data is in the minority, known as sample-heterogeneity. We will focus on the problem of mean estimation. In this model, the algorithm outputs a single estimate (as opposed to a list of hypotheses in list-decodable learning). As highlighted earlier, one needs to restrict the outliers to make the problem well-posed. To this end, we will assume that (i) the i-th data point x_i is sampled independently from a distribution D_i, and (ii) and each distribution has the same mean μ, but the variances could be different. Here, the variance of the distribution determines its quality: the lower the variance, the higher the quality. We will focus on the regime when a vanishing fraction of points, $o_n(1)$, are good quality (low variance), but the remaining data points could have infinite variance. Primary attention will be given to the case when each D_i is a Gaussian distribution, but the results hold more generally for symmetric unimodal distributions.

Inference Task 5 (Location Estimation under Sample-Heterogeneity)**.** *Let $\mu \in \mathbb{R}$ and $\sigma_1, \ldots, \sigma_n \in \mathbb{R}_+$ be arbitrary and unknown. Let S be a set of n samples where each $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$ independently. Given S as input, compute an estimate $\widehat{\mu}$ such that $|\widehat{\mu} - \mu|$ is small with high probability.*

Inference Task 5 has long history in statistics dating back to the 1960s [Wei69; HM97]. Recently, it was studied in the theoretical computer science literature by [CDKL14] motivated by crowdsourcing applications. However, optimal rates were unknown in many important settings.

It is easy to see that even if a single data point has a large variance, the sample mean performs poorly (since the variance of the sample mean depends on the largest variance). Even sample median requires at least $\Omega(\sqrt{n})$ variances to be small. The central regime of interest is when $o(\sqrt{n})$-many samples have small variances.

Question 10. *Are there efficient algorithms for Inference Task 5 that work when the fraction of high-quality samples is $o(1/\sqrt{n})$?*

As it might be challenging to know a priori the level of heterogeneity in the data, we like an estimator that is adaptive to the heterogeneity in the data. In particular, we would like our algorithm to recover the $O(1/\sqrt{n})$-convergence in the i.i.d. regime (without knowing beforehand that the data is i.i.d.).

Question 11. *Are there efficient algorithms for Inference Task 5 that adapt to the level of heterogeneity in the data?*

Our Contributions In Chapter 7, we make progress on Questions 10 and 11 by developing the first estimator that attains near-optimal error guarantees for Inference Task 5 in levels of heterogeneity. The proposed estimator adapts to the heterogeneity of the data in many settings, achieving near-optimal performance in the i.i.d. regime and in the regime

where only $O(\log n)$ points are of high quality. Our proposed estimator combines several classical estimators, median, shorth, and modal interval, to obtain the near-optimal algorithms; as a result, the proposed estimator is also computationally-efficient and practical.

3 ROBUST MEAN ESTIMATION

A common mistake that people make when trying to design something completely foolproof is to underestimate the ingenuity of complete fools.

—Douglas Adams

We study the problem of outlier robust high-dimensional mean estimation under a finite covariance assumption, and more broadly under finite low-degree moment assumptions. We consider a standard stability condition from the recent robust statistics literature and prove that, except with exponentially small failure probability, there exists a large fraction of the inliers satisfying this condition. As a corollary, it follows that a number of recently developed algorithms for robust mean estimation, including iterative filtering and non-convex gradient descent, give optimal error estimators with (near-)subgaussian rates. Previous analyses of these algorithms gave significantly suboptimal rates. As a corollary of our approach, we obtain computationally efficient spectral algorithms with subgaussian rate for outlier-robust mean estimation in the strong contamination model under a finite covariance assumption.

3.1 Introduction

3.1.1 Background and Motivation

Consider the following problem: For a given family $\mathcal{F}$ of distributions on $\mathbb{R}^d$, estimate the mean of an unknown $D \in \mathcal{F}$, given access to i.i.d. samples from D. This is the problem of (multivariate) mean estimation and is arguably *the* most fundamental statistical task. In the most basic setting where $\mathcal{F}$ is the family of high-dimensional Gaussians, the empirical mean is well-known to be an optimal estimator — in the sense that it achieves

the best possible accuracy-confidence tradeoff and is easy to compute. Unfortunately, the empirical mean is known to be highly suboptimal if we relax the aforementioned modeling assumptions. In this work, we study high-dimensional mean estimation *in the high confidence regime* when the underlying family $\mathcal{F}$ is only assumed to satisfy bounded moment conditions (e.g., finite covariance). Moreover, we relax the "i.i.d. assumption" and aim to obtain estimators that are robust to a constant fraction of adversarial outliers.

Throughout this paper, we focus on the following data contamination model (see, e.g., [DKKLMS16]) that generalizes several existing models, including Huber's contamination model [Hub64].

Definition 3.1.1 (Strong Contamination Model)**.** *Given a parameter $0 < \epsilon < 1/2$ and a distribution family $\mathcal{F}$ on $\mathbb{R}^d$, the* adversary *operates as follows: The algorithm specifies the number of samples n, and n samples are drawn from some unknown $D \in \mathcal{F}$. The adversary is allowed to inspect the samples, remove up to ϵn of them and replace them with arbitrary points. This modified set of n points is then given as input to the algorithm. We say that a set of samples is* ϵ-corrupted *if it is generated by the above process.*

The parameter ϵ in Definition 3.1.1 is the fraction of outliers and quantifies the power of the adversary. Intuitively, among our input samples, an unknown $(1-\epsilon)$ fraction are generated from a distribution of interest and are called *inliers*, and the rest are called *outliers*.

We note that the strong contamination model is strictly stronger than Huber's contamination model. Recall that in Huber's contamination model [Hub64], the adversary generates samples from a mixture distribution P of the form $P = (1-\epsilon)D + \epsilon N$, where $D \in \mathcal{F}$ is the unknown target distribution and N is an adversarially chosen noise distribution. That is, in Huber's model the adversary is oblivious to the inliers and is only allowed to add outliers.

In the context of robust mean estimation, we want to design an algorithm (estimator) with the following performance: Given any ϵ-corrupted set of n samples from an unknown distribution $D \in \mathcal{F}$, the algorithm outputs an estimate $\widehat{\mu} \in \mathbb{R}^d$ of the target mean μ of D such that *with high probability* the ℓ_2-norm $\|\widehat{\mu} - \mu\|_2$ is small. The ultimate goal is to obtain a *computationally efficient estimator with optimal confidence-accuracy tradeoff*. For concreteness, in the proceeding discussion we focus on the case that $\mathcal{F}$ is the family of all distributions on $\mathbb{R}^d$ with bounded covariance, i.e., any $D \in \mathcal{F}$ has covariance matrix $\Sigma \preceq I$. (We note that the results of this paper apply for the more general setting where $\Sigma \preceq \sigma^2 I$, where $\sigma > 0$ is unknown to the algorithm.)

Perhaps surprisingly, even for the special case of $\epsilon = 0$ (i.e., without adversarial contamination), designing an optimal mean estimator in the high-confidence regime is far from trivial. In particular, it is well-known (and easy to see) that the empirical mean achieves highly sub-optimal rate. A sequence of works in mathematical statistics (see, e.g., [Cat12; Min15; DLLO16; LM19d]) designed novel estimators with improved rates, culminating in an optimal estimator [LM19d]. See [LM19a] for a survey on the topic. The estimator of [LM19d] is based on the median-of-means framework and achieves a "subgaussian" performance guarantee:

$$\|\widehat{\mu} - \mu\|_2 = O(\sqrt{d/n} + \sqrt{\log(1/\tau)/n}) \,, \tag{3.1}$$

where $\tau > 0$ is the failure probability. The error rate (3.1) is information-theoretically optimal for any estimator and matches the error rate achieved by the empirical mean on Gaussian data. Unfortunately, the estimator of [LM19d] is not efficiently computable. In particular, known algorithms to compute it have running time exponential in the dimension d. Related works [Min15; PBR19] provide computationally efficient estimators alas with suboptimal rates. The first polynomial time algorithm achieving the optimal rate (3.1) was given in [Hop20], using a convex program derived from the Sums-of-Squares

method. Efficient algorithms with improved asymptotic runtimes were subsequently given in [CFB19; DL22b; LLVZ20].

We now turn to the outlier-robust setting ($\epsilon > 0$) for the constant confidence regime, i.e., when the failure probability τ is a small universal constant. The statistical foundations of outlier-robust estimation were laid out in early work by the robust statistics community, starting with the pioneering works of [Tuk60] and [Hub64]. For example, the minimax optimal estimator satisfies:

$$\|\widehat{\mu} - \mu\|_2 = O(\sqrt{\epsilon} + \sqrt{d/n}) \; . \tag{3.2}$$

Until fairly recently however, all known polynomial-time estimators attained sub-optimal rates. Specifically, even in the limit when $n \to \infty$, known polynomial time estimators achieved error of $O(\sqrt{\epsilon d})$, i.e., scaling polynomially with the dimension d. Recent work in computer science, starting with [DKKLMS16; LRV16], gave the first efficiently computable outlier-robust estimators for high-dimensional mean estimation. For bounded covariance distributions, [DKKLMS17; SCV18] gave efficient algorithms with the right error guarantee of $O(\sqrt{\epsilon})$. Specifically, the filtering algorithm of [DKKLMS17] is known to achieve a near-optimal rate of $O(\sqrt{\epsilon} + \sqrt{d \log d/n})$ (with high constant probability).

In this paper, we aim to achieve the best of both worlds. In particular, we ask the following question:

Can we design computationally efficient *estimators with subgaussian rates and optimal dependence on the contamination parameter ϵ?*

Recent work [LM21b] gave an *exponential time* estimator with optimal rate in this setting. Specifically, [LM21b] showed that a multivariate extension of the trimmed-mean

achieved the optimal error of

$$\|\widehat{\mu} - \mu\|_2 = O\left(\sqrt{\epsilon} + \sqrt{d/n} + \sqrt{\log(1/\tau)/n}\right) . \tag{3.3}$$

We note that [LM21b] posed as an open question the existence of a computationally efficient estimator achieving the optimal rate (3.3). Two recent works [DL22b; LLVZ20] gave efficient estimators with subgaussian rates that are outlier-robust in the *additive* contamination model — a *weaker* model than that of Definition 3.1.1. Prior to this work, no polynomial time algorithm with optimal (or near-optimal) rate was known in the strong contamination model of Definition 3.1.1. As a corollary of our approach, we answer the question of [LM21b] in the affirmative (see Proposition 3.1.6). In the following subsection, we describe our results in detail.

3.1.2 Our Contributions

At a high-level, the main conceptual contribution of this work is in showing that several previously developed computationally efficient algorithms for high-dimensional robust mean estimation achieve near-subgaussian rates or subgaussian rates (after a simple pre-processing). A number of these algorithms are known to succeed under a standard *stability* condition (Definition 3.1.2) – a simple deterministic condition on the empirical mean and covariance of a finite point set. We will call such algorithms *stability-based.*

Our contributions are as follows:

- We show (Theorem 3.1.4) that given a set of i.i.d. samples from a finite covariance distribution, except with exponentially small failure probability, there exists a large fraction of the samples satisfying the stability condition. As a corollary, it follows (Proposition 3.1.5) that *any* stability-based robust mean estimation algorithm achieves optimal error with (near-)subgaussian rates.

- We show an analogous probabilistic result (Theorem 3.1.8) for known covariance distributions (or, more generally, spherical covariance distributions) with bounded k-th moment, for some $k \geq 4$. As a corollary, we obtain that *any* stability-based robust mean estimator achieves optimal error with (near-)subgaussian rates (Proposition 3.1.9.)

- For the case of finite covariance distributions, we show (Proposition 3.1.6) that a simple pre-processing step followed by any stability-based robust mean estimation algorithm yields optimal error and subgaussian rates.

To formally state our results, we require some terminology and background.

Basic Notation For a vector $v \in \mathbb{R}^d$, we use $\|v\|_2$ to denote its ℓ_2-norm. For a square matrix M, we use $\mathrm{tr}(M)$ to denotes its trace, and $\|M\|_{\mathrm{op}}$ to denote its spectral norm. We say a symmetric matrix A is PSD (positive semidefinite) if $x^\top A x \geq 0$ for all vectors x. For a PSD matrix M, we use $\mathrm{r}(M)$ to denote its stable rank (or intrinsic dimension), i.e., $\mathrm{r}(M) := \mathrm{tr}(M)/\|M\|_{\mathrm{op}}$. For two symmetric matrices A and B, we use $\langle A, B \rangle$ to denote the trace inner product $\mathrm{tr}(AB)$ and say $A \preceq B$ when $B - A$ is PSD.

We use $[n]$ to denote the set $\{1, \ldots, n\}$ and $\mathcal{S}^{d-1}$ to denote the d-dimensional unit sphere. We use Δ_n to denote the probability simplex on $[n]$, i.e., $\Delta_n = \{w \in \mathbb{R}^n : w_i \geq 0, \sum_{i=1}^n w_i = 1\}$. For a multiset $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ of cardinality n and $w \in \Delta_n$, we use μ_w to denote its weighted mean $\mu_w = \sum_{i=1}^n w_i x_i$. Similarly, we use $\overline{\Sigma}_w$ to denote its weighted second moment matrix (centered with respect to μ) $\overline{\Sigma}_w = \sum_{i=1}^n w_i (x_i - \mu)(x_i - \mu)^\top$. For a set $S \subset \mathbb{R}^d$, we denote $\mu_S = (1/|S|) \sum_{x \in S} x$ and $\overline{\Sigma}_S = (1/|S|) \sum_{x \in S} (x - \mu)(x - \mu)^\top$ to denote the mean and (central) second moment matrix with respect to the uniform distribution on S.

For a set E, we use $\mathbb{I}(x \in E)$ to denote the indicator function for event E. For simplicity, we use $\mathbb{I}(x \geq t)$ to denote the indicator function for the event $E = \{x : x \geq t\}$.

For a random variable Z, we use $\mathbf{Var}(Z)$ to denote its variance. We use $d_{\mathrm{TV}}(p, q)$ to denote the total variation distance between distributions p and q.

Stability Condition and Robust Mean Estimation. We can now define the stability condition:

Definition 3.1.2 (see, e.g., [DK19])**.** *Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. A finite set $S \subset \mathbb{R}^d$ is (ϵ, δ)-stable with respect to mean $\mu \in \mathbb{R}^d$ and σ^2 if for every $S' \subseteq S$ with $|S'| \geq (1-\epsilon)|S|$, the following conditions hold: (i) $\|\mu_{S'} - \mu\|_2 \leq \sigma\delta$, and (ii) $\|\overline{\Sigma}_{S'} - \sigma^2 I\|_{\mathrm{op}} \leq \sigma^2\delta^2/\epsilon$.*

The aforementioned condition or a variant thereof is used in every known outlier-robust mean estimation algorithm. Definition 3.1.2 requires that after restricting to a $(1-\epsilon)$-density subset S', the sample mean of S' is within $\sigma\delta$ of the mean μ, and the sample variance of S' is $\sigma^2(1 \pm \delta^2/\epsilon)$ in every direction. (We note that Definition 3.1.2 is intended for distributions with covariance $\Sigma \preceq \sigma^2 I$). We will omit the parameters μ and σ^2 when they are clear from context. In particular, our proofs will focus on the case $\sigma^2 = 1$, which can be achieved by scaling the datapoints appropriately.

A number of known algorithmic techniques previously used for robust mean estimation, including convex programming based methods [DKKLMS16; SCV18; CDG19], iterative filtering [DKKLMS16; DKKLMS17; DHL19], and even first-order methods [CDGS20; ZJS22b], are known to succeed under the stability condition. Specifically, prior work has established the following theorem:

Theorem 3.1.3 (Robust Mean Estimation Under Stability, see, e.g., [DK19])**.** *Let $T \subset \mathbb{R}^d$ be an ϵ-corrupted version of a set S with the following properties: S contains a subset $S' \subseteq S$ such that $|S'| \geq (1-\epsilon)|S|$ and S' is $(C\epsilon, \delta)$ stable with respect to $\mu \in \mathbb{R}^d$ and $\sigma^2 \in \mathbb{R}_+$, for a sufficiently large constant $C > 0$. Then there is a polynomial-time algorithm, that on input ϵ, T, computes $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 = O(\sigma\delta)$.*

We note in particular that the iterative filtering algorithm [DKKLMS17; DK19] (see also Section 2.4.3 of [DK19]) is a very simple and practical stability-based algorithm. While previous works made the assumption that the upper bound parameter σ^2 is known to the algorithm, we point out in Appendix A.1.2 that essentially the same algorithm and analysis works for unknown σ^2 as well.

Our Results Our first main result establishes the stability of a subset of i.i.d. points drawn from a distribution with bounded covariance.

Theorem 3.1.4. *Fix any $0 < \tau < 1$. Let S be a multiset of n i.i.d. samples from a distribution on $\mathbb{R}^d$ with mean μ and covariance Σ . Let $\epsilon' = \Theta(\log(1/\tau)/n + \epsilon) \leq c$, for a sufficiently small constant $c > 0$. Then, with probability at least $1 - \tau$, there exists a subset $S' \subseteq S$ such that $|S'| \geq (1 - \epsilon')n$ and S' is $(2\epsilon', \delta)$-stable with respect to μ and $\|\Sigma\|_{\mathrm{op}}$, where $\delta = O(\sqrt{(\mathrm{r}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\epsilon} + \sqrt{\log(1/\tau)/n})$.*

Theorem 3.1.4 significantly improves the probabilistic guarantees in prior work on robust mean estimation. This includes the *resilience* condition of [SCV18; ZJS22a] and the *goodness* condition of [DHL19].

As a corollary, it follows that any stability-based algorithm for robust mean estimation achieves near-subgaussian rates.

Proposition 3.1.5. *Let T be an ϵ-corrupted set of n samples from a distribution in $\mathbb{R}^d$ with mean μ and covariance Σ. Let $\epsilon' = \Theta(\log(1/\tau)/n + \epsilon) \leq c$ be given, for a constant $c > 0$. Then any stability-based algorithm on input T and ϵ', efficiently computes $\widehat{\mu}$ such that with probability at least $1 - \tau$, we have $\|\widehat{\mu} - \mu\|_2 = O(\sqrt{(\mathrm{tr}\,(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\|\Sigma\|_{\mathrm{op}}\epsilon} + \sqrt{\|\Sigma\|_{\mathrm{op}}\log(1/\tau)/n})$.*

We note that the above error rate is minimax optimal in both ϵ and τ, and the restriction of $\log(1/\tau)/n = O(1)$ is information-theoretically required [DLLO16]. In particular, the term $\sqrt{\log(1/\tau)/n}$ is *additive* as opposed to multiplicative. The first term is near-optimal, up to the $\sqrt{\log \mathrm{r}(\Sigma)}$ factor, which is at most $\sqrt{\log d}$ (recall that $\mathrm{r}(\Sigma)$ denotes

the stable rank of Σ, i.e., $\mathrm{r}(\Sigma) = \mathrm{tr}(\Sigma)/\|\Sigma\|_{\mathrm{op}}$). Prior to this work, the existence of a polynomial-time algorithm achieving the above near-subgaussian rate in the strong contamination model was open. Proposition 3.1.5 shows that any stability-based algorithm suffices for this purpose, and in particular it implies that the iterative filtering algorithm [DK19] achieves this rate *as is*.

Given the above, a natural question is whether stability-based algorithms achieve subgaussian rates *exactly*, i.e., whether they match the optimal bound (3.3) attained by the computationally inefficient estimator of [LM21b]. While the answer to this question remains open, we show that after a simple pre-processing of the data, stability-based estimators are indeed subgaussian.

The pre-processing step follows the median-of-means principle [NY83; JVV86; AMS99]. Given a multiset of n points $x_1, \ldots, x_n$ in $\mathbb{R}^d$ and $k \in [n]$, we proceed as follows:

1. First randomly bucket the data into k disjoint buckets of equal size (if k does not divide n, remove some samples) and compute their empirical means $z_1, \ldots, z_k$.

2. Output an (appropriately defined) multivariate median of $z_1, \ldots, z_k$.

Notably, for the case of $\epsilon = 0$, all known efficient mean estimators with subgaussian rates use the median-of-means framework [Hop20; DL22b; CFB19; LLVZ20].

To obtain the desired computationally efficient robust mean estimators with subgaussian rates, we proceed as follows:

1. Given a multiset S of n ϵ-corrupted samples, randomly group the data into $k = \lfloor \epsilon' n \rfloor$ disjoint buckets, where $\epsilon' = \Theta(\log(1/\tau)/n + \epsilon)$, and let $z_1, \ldots, z_k$ be the corresponding empirical means of the buckets.

2. Run any stability-based robust mean estimator on input $\{z_1, \ldots, z_k\}$.

Specifically, we show:

Proposition 3.1.6. (*informal*) *Consider the same setting as in Proposition 3.1.5. Let* $k = \lfloor \epsilon' n \rfloor$ *and* $z_1, \ldots, z_k$ *be the points after median-of-means pre-processing on the corrupted set* T*. Then any stability-based algorithm, on input* $\{z_1, \ldots, z_k\}$*, computes* $\widehat{\mu}$ *such that with probability at least* $1 - \tau$*, it holds* $\|\widehat{\mu} - \mu\|_2 = O(\sqrt{\mathrm{tr}(\Sigma)/n} + \sqrt{\|\Sigma\|_{\mathrm{op}}\epsilon} + \sqrt{\|\Sigma\|_{\mathrm{op}} \log(1/\tau)/n})$.

Proposition 3.1.6 yields the first computationally efficient algorithm with subgaussian rates in the strong contamination model, answering the open question of [LM21b].

To prove Proposition 3.1.6, we establish a connection between the median-of-means principle and stability. In particular, we show that the key probabilistic lemma from the median-of-means literature [LM19d; DL22b] also implies stability.

Theorem 3.1.7. (*informal*) *Consider the setting of Theorem 3.1.4 and set* $k = \lfloor \epsilon' n \rfloor$*. The set* $\{z_1, \ldots, z_k\}$*, with probability* $1 - \tau$*, contains a subset of size at least* $0.99k$ *which is* $(0.1, \delta)$*-stable with respect to* μ *and* $k\|\Sigma\|_{\mathrm{op}}/n$*, where* $\delta = O(\sqrt{\mathrm{r}(\Sigma)/k} + 1)$.

A drawback of the median-of-means framework is that the error dependence on ϵ does not improve if we impose stronger assumptions on the distribution. Even if the underlying distribution is an identity covariance Gaussian, the error rate would scale as $O(\sqrt{\epsilon})$, whereas the stability-based algorithms achieve error of $O(\epsilon\sqrt{\log(1/\epsilon)})$ [DKKLMS16]. Our next result establishes tighter error bounds for distributions with identity covariance and bounded central moments.

We say that a distribution has a bounded k-th central moment σ_k, if for all unit vectors v, it holds $(\mathbb{E}(v^\top(X - \mu))^k)^{1/k} \leq \sigma_k(\mathbb{E}(v^\top(X - \mu))^2)^{1/2}$. For such distributions, we establish the following stronger stability condition.

Theorem 3.1.8. *Let* S *be a multiset of* n *i.i.d. samples from a distribution on* $\mathbb{R}^d$ *with mean* μ*, covariance* $\Sigma = I$*, and bounded central moment* σ_k*, for some* $k \geq 4$*. Let* $\epsilon' = \Theta(\log(1/\tau)/n + \epsilon) \leq c$*, for a sufficiently small constant* $c > 0$*. Then, with probability at least* $1 - \tau$*, there exists a*

subset $S' \subseteq S$ such that $|S'| \geq (1-\epsilon')n$ and $|S'|$ is $(2\epsilon', \delta)$-stable with respect to μ and $\sigma^2 = 1$, where $\delta = O(\sqrt{d \log d/n} + \sigma_k \epsilon^{1-\frac{1}{k}} + \sigma_4 \sqrt{\log(1/\tau)/n})$.

As a corollary, we obtain the following result for robust mean estimation with high probability in the strong contamination model:

Proposition 3.1.9. *Let T be an ϵ-corrupted set of n points from a distribution on $\mathbb{R}^d$ with mean μ, covariance $\sigma^2 I$, and k-th bounded central moment σ_k, for some $k \geq 4$. Let $\epsilon' = \Theta(\log(1/\tau)/n + \epsilon) \leq c$ be given, for some $c > 0$. Then any stability-based algorithm, on input T and ϵ', efficiently computes $\widehat{\mu}$ such that with probability at least $1-\tau$, we have $\|\widehat{\mu} - \mu\|_2 = O(\sigma(\sqrt{d \log d/n} + \sigma_k \epsilon^{1-\frac{1}{k}} + \sigma_4 \sqrt{\log(1/\tau)/n}))$.*

We note that the above error rate is near-optimal up to the $\log d$ factor and the dependence on σ_4. Prior to this work, no polynomial-time estimator achieving this rate was known. Finally, recent computational hardness results [HL19] suggest that the assumption on the covariance above is inherent to obtain computationally efficient estimators with error rate better than $\Omega(\sqrt{\epsilon})$, even in the constant confidence regime.

3.1.3 Related Work

Since the initials works [DKKLMS16; LRV16], there has been an explosion of research activity on algorithmic aspects of outlier-robust high dimensional estimation by several communities. See, e.g., [DK19] for a recent survey on the topic. In the context of outlier-robust mean estimation, a number of works [DKKLMS17; SCV18; CDG19; DHL19] have obtained efficient algorithms under various assumptions on the distribution of the inliers. Notably, efficient high-dimensional outlier-robust mean estimators have been used as primitives for robustly solving machine learning tasks that can be expressed as stochastic optimization problems [PSBR20; DKKLSS19]. The above works typically

focus on the constant probability error regime and do not establish subgaussian rates for their estimators.

Two recent works [DL22b; LLVZ20] studied the problem of outlier-robust mean estimation in the additive contamination model (when the adversary is only allowed to add outliers) and gave computationally efficient algorithms with subgaussian rates. Specifically, [DL22b] gave an SDP-based algorithm, which is very similar to the algorithm of [CDG19]. The algorithm of [LLVZ20] is a fairly sophisticated iterative spectral algorithm, building on [CFB19]. In the strong contamination model, non-constructive outlier-robust estimators with subgaussian rates were established very recently. Specifically, [LM21b] gave a an exponential time estimator achieving the optimal rate. Our Proposition 3.1.6 implies that a very simple and practical algorithm – pre-processing followed by iterative filtering [DKKLMS17; DK19] – achieves this guarantee.

In an independent and concurrent work, Hopkins, Li, and Zhang [HLZ20] also studied the relation between median-of-means and stability for the case of bounded covariance.

3.1.4 Organization

In Section 3.2, we prove Theorem 3.1.4 that establishes the stability of points sampled from a finite covariance distribution. In Section 3.3, we establish the connection between median-of-means principle and stability to prove Theorem 3.1.7. Finally, Section 3.4 contains our results for distributions with identity covariance and finite central moments.

3.2 Robust Mean Estimation for Finite Covariance Distributions

Problem Setting Consider a distribution P in $\mathbb{R}^d$ with unknown mean μ and unknown covariance Σ. We first note that it suffices to consider the distributions such that $\|\Sigma\|_{\text{op}} = 1$. Note that for covariance matrices Σ with $\|\Sigma\|_{\text{op}} = 1$, we have $\mathrm{r}(\Sigma) = \mathrm{tr}(\Sigma)$. In the remainder of this section, we will thus establish the (ϵ, δ) stability with respect to μ and $\sigma^2 = 1$, where $\delta = O(\sqrt{\mathrm{tr}(\Sigma)\log(\mathrm{r}(\Sigma))/n} + \sqrt{\epsilon} + \sqrt{\log(1/\tau)/n})$.

Let S be a multiset of n i.i.d. samples from P. For the ease of exposition, we will assume that the support of P is bounded, i.e., for each i, $\|x_i - \mu\|_2 = O(\sqrt{\mathrm{tr}(\Sigma)/\epsilon})$ almost surely. As we show in Section 3.2.3, we can simply consider the points violating this condition as outliers.

We first relax the conditions for stability in the Definition 3.1.2 in the following Claim 3.2.1, proved in Appendix A.4.1, at an additional cost of $O(\sqrt{\epsilon})$.

Claim 3.2.1. (*Stability for bounded covariance*) *Let $R \subset \mathbb{R}^d$ be a finite multiset such that $\|\mu_R - \mu\|_2 \leq \delta$, and $\|\overline{\Sigma}_R - I\|_{\text{op}} \leq \delta^2/\epsilon$ for some $0 \leq \epsilon \leq \delta$. Then R is $(\Theta(\epsilon), \delta')$ stable with respect to μ (and $\sigma^2 = 1$), where $\delta' = O(\delta + \sqrt{\epsilon})$.*

Given Claim 3.2.1, our goal in proving Theorem 3.1.4 is to show that with probability $1 - \tau$, there exists a set $S' \subseteq S$ such that $|S'| \geq (1 - \epsilon')n$, $\|\mu_{S'} - \mu\|_2 \leq \delta$ and $\|\overline{\Sigma}_S - I\|_{\text{op}} \leq \delta^2/\epsilon'$, for some value of $\delta = O(\sqrt{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)/n} + \sqrt{\epsilon} + \sqrt{\log(1/\tau)/n})$ and $\epsilon' = \Theta(\epsilon + \log(1/\tau)/n)$.

We first remark that the original set S of n i.i.d. data points does not satisfy either of the conditions in Claim 3.2.1. It does not satisfy the first condition because the sample mean is highly sub-optimal for heavy-tailed data [Cat12]. For the second condition, we note that the known concentration results for $\overline{\Sigma}_S$ are not sufficient. For example, consider the case of $\Sigma = I$ in the parameter regime of ϵ, τ, and n such that $\epsilon = O(\log(1/\tau)/n)$ and

$n = \Omega(d \log d/\epsilon)$ so that $\delta = O(\sqrt{\epsilon})$. For S to be (ϵ, δ) stable, we require that $\|\overline{\Sigma}_S - I\|_{\text{op}} = O(1)$ with probability $1 - \tau$. However, the Matrix-Chernoff bound (see, e.g., [Tro15, Theorem 5.1.1]) only guarantees that with probability at least $1 - \tau$, $\|\overline{\Sigma}_S - I\|_{\text{op}} = \tilde{O}(d)$.

To further elaborate that the set S of n i.i.d. data points does not satisfy either of the conditions in Claim 3.2.1, we give the following concrete example. The following example shows that the lack of concentration outlined in the previous graph is not simply an artifact of analysis but that it is inherent to heavy-tailed distributions, at least in some parameter regimes. In particular, consider the special case when ϵ and $\log(1/\tau)/n$ are both small positive constants and $n \geq d \log d$, implying that ϵ' is also a small absolute constant. Then, the question is whether the set S satisfies the following: with probability at least $1 - 2^{-\Omega(n)} \geq 1 - e^{-5d}$, the set S satisfies that $\|\mu_S - \mu\|_2 = O(\delta)$ and $\|\Sigma_S - \Sigma\|_{\text{op}} = O(\delta^2)$ for δ, the stability parameter, equal to $O(1)$. The following example will show that this is false: in fact, the actual value of δ, the stability parameter, will be roughly $\sqrt{d}$ times larger than the desired value (as long as $n = O(d^2)$).

Example 3.2.2 (Empirical covariance matrix does not concentrate fast enough)**.** *Let $u \in \mathbb{R}^d$ be an arbitrary vector with norm $\sqrt{d}$. Let P be the distribution over $\mathbb{R}^d$ of the random vector X defined as follows: with probability $1/2d$ each, X takes the values u and $-u$. With the remaining probability of $1 - \frac{1}{d}$, X is equal to the origin. Then $\mu := \mathbb{E}[X] = 0$ and $\Sigma := \mathbb{E}[(X - \mu)(X - \mu)^\top] = \frac{1}{d} uu^\top$, and thus $\|\Sigma\|_{\text{op}} = (1/d)\|u\|_2^2 = 1$. Observe that the distribution P also satisfies the bounded support condition.*

Let S be the set of n i.i.d. samples from the distribution P. Then, the probability that we observe $m_1 \geq 1$ number of u, $n - m_1$ number of origin is

$$\left(\frac{1}{2d}\right)^{m_1} \left(1 - \frac{1}{d}\right)^{n - m_1} \geq d^{-2m_1} e^{-2\frac{n}{d}},$$

where we use that for $x \in (0, 0.5)$, $1 - x \geq e^{-2x}$. Plugging in $m_1 = \frac{d}{\log d}$, we obtain that the expression above is at least $e^{-2d - 2\frac{n}{d}}$. Let $\mathcal{E}'$ be this event, which holds with probability at least

e^{-4d} when $n \in [d, d^2]$. On the event $\mathcal{E}'$, the mean of S has norm $\|\mu_S\|_2 = \frac{m_1}{n}\|u\|_2 = \frac{\sqrt{d}}{\log d}$ and the second moment matrix of S satisfies $\|\overline{\Sigma}_S\|_{\mathrm{op}} = \frac{m_1}{n}\|u\|_2^2 = \frac{d}{\log d}$. Thus, the resulting stability parameter of the set S on $\mathcal{E}'$ is of the order of $\sqrt{d/\log d}$, which is roughly $\sqrt{d}$ times the desired value of $O(1)$.

The rest of this section is devoted to showing that, with high probability, it is possible to remove $\epsilon' n$ points from S such that both conditions in Claim 3.2.1 are satisfied for the subset.

3.2.1 Controlling the Variance

As a first step, we show that it is possible to remove an ϵ-fraction of points so that the second moment matrix concentrates. Since finding a subset is a discrete optimization problem, we first perform a continuous relaxation: instead of finding a large subset, we find a suitable distribution on points. Define the following set of distributions:

$$\Delta_{n,\epsilon} = \left\{ w \in \mathbb{R}^n : 0 \le w_i \le 1/((1-\epsilon)n); \sum_{i=1}^{n} w_i = 1 \right\}.$$

Note that $\Delta_{n,\epsilon}$ is the convex hull of all the uniform distributions on $S' \subseteq S : |S'| \ge (1-\epsilon)n$. In Appendix A.4.2, we show how to recover a subset S' from the w. Although we use the set $\Delta_{n,\epsilon}$ for the sole purpose of theoretical analysis, the object $\Delta_{n,\epsilon}$ has also been useful in the design of computationally efficient algorithms [DKKLMS16; DK19]. We will now show that, with high probability, there exists a $w \in \Delta_{n,\epsilon}$ such that $\overline{\Sigma}_w$ has small spectral norm.

Our proof technique has three main ingredients: (i) minimax duality, (ii) truncation, and (iii) concentration of truncated empirical processes. Let $\mathcal{M}$ be the set of all PSD matrices with trace norm 1, i.e., $\mathcal{M} = \{M : M \succeq 0, \mathrm{tr}(M) = 1\}$. Using minimax duality [Sio58] and the variational characterization of spectral norm, we obtain the following

reformulation:

$$\begin{aligned}\min_{w\in\Delta_{n,\epsilon}} \|\overline{\Sigma}_w - I\|_{\text{op}} &\leq 1 + \min_{w\in\Delta_{n,\epsilon}} \|\overline{\Sigma}_w\|_{\text{op}} \\ &= 1 + \min_{w\in\Delta_{n,\epsilon}} \max_{M\in\mathcal{M}} \left\langle \sum_{i=1}^n w_i (x_i-\mu)(x_i-\mu)^\top, M \right\rangle \\ &= 1 + \max_{M\in\mathcal{M}} \min_{w\in\Delta_{n,\epsilon}} \left\langle \sum_{i=1}^n w_i (x_i-\mu)(x_i-\mu)^\top, M \right\rangle. \end{aligned} \tag{3.4}$$

This dual reformulation plays a fundamental role in our analysis. Lemma 3.2.3 below, proved in Appendix A.3.2, states that, with high probability, all the terms in the dual reformulation are bounded.

Lemma 3.2.3. *Let $x_1, \dots, x_n$ be n i.i.d. points from a distribution in $\mathbb{R}^d$ with mean μ and covariance $\Sigma \preceq I$. Let $Q = \Theta(1/\sqrt{\epsilon} + (1/\epsilon)\sqrt{\text{tr}(\Sigma)/n})$. For $M \in \mathcal{M}$, let $S_M = \{i \in [n] : (x_i-\mu)^\top M (x_i-\mu) \leq Q^2\}$. Let $\mathcal{E}$ be the event $\mathcal{E} = \{\forall M \in \mathcal{M}, |S_M| \geq (1-\epsilon)n\}$. There exists a constant $c > 0$ such that the event $\mathcal{E}$ happens with probability at least $1 - \exp(-c\epsilon n)$.*

Lemma 3.2.3 draws on the results by Lugosi and Mendelson [LM21b, Proposition 1] and Depersin and Lecué [DL22b, Proposition 1]. The proof is given in Appendix A.3. Importantly, given $n = \Omega(\text{tr}(\Sigma)/\epsilon)$ samples, the threshold Q is $O(1/\sqrt{\epsilon})$. Approximating the empirical process in Eq. (3.4) with a truncated process allows us to use the powerful inequality for concentration of bounded empirical processes due to Talagrand [Tal96a]. Formally, we show the following lemma:

Lemma 3.2.4. *Let $x_1, \dots, x_n$ be n i.i.d. points from a distribution in $\mathbb{R}^d$ with mean μ and covariance $\Sigma \preceq I$. Further assume that for each i, $\|x_i - \mu\|_2 = O(\sqrt{\text{tr}(\Sigma)/\epsilon})$. There exists $c, c' > 0$ such that for $\epsilon \in (0, c')$, with probability $1 - 2\exp(-cn\epsilon)$, we have that $\min_{w\in\Delta_{n,\epsilon}} \left\|\overline{\Sigma}_w - I\right\|_{\text{op}} \leq \delta^2/\epsilon$, where $\delta = O(\sqrt{(\text{tr}(\Sigma)\log \text{r}(\Sigma))/n} + \sqrt{\epsilon})$.*

Proof. Throughout the proof, assume that the event $\mathcal{E}$ from Lemma 3.2.3 holds. Without

loss of generality, also assume that $\mu = 0$. Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be the following function:

$$f(x) := \begin{cases} x, & \text{if } x \le Q^2 \\ Q^2, & \text{otherwise.} \end{cases} \tag{3.5}$$

It follows directly that f is 1-Lipschitz and $0 \le f(x) \le x$. Using minimax duality,

$$\min_{w \in \Delta_{n,\epsilon}} \|\overline{\Sigma}_w - I\|_{\text{op}} \le 1 + \max_{M \in \mathcal{M}} \min_{w \in \Delta_{n,\epsilon}} \sum w_i x_i^\top M x_i \le 1 + \max_{M \in \mathcal{M}} \sum_{i=1}^n f(x_i^\top M x_i)/((1-\epsilon)n),$$

where the second inequality uses that on event $\mathcal{E}$, for every $M \in \mathcal{M}$, the set $S_M = \{[i] \in n : x_i^\top M x_i \le Q^2\}$ has cardinality larger than $(1-\epsilon)n$, and thus, the uniform distribution on the set S_M belongs to $\Delta_{n,\epsilon}$. Define the following empirical processes R and R':

$$R = \sup_{M \in \mathcal{M}} \sum_{i=1}^n f(x_i^\top M x_i), \qquad R' = \sup_{M \in \mathcal{M}} \sum_{i=1}^n f(x_i^\top M x_i) - \mathbb{E}\, f(x_i^\top M x_i).$$

As $0 \le f(x) \le x$, we have that $0 \le \mathbb{E}\, f(x_i^\top M x) \le \mathbb{E}\, x_i^\top M x \le 1$, which gives that $|R - R'| \le n$. Overall, we obtain the following bound:

$$\min_{w \in \Delta_{n,\epsilon}} \|\overline{\Sigma}_w - I\|_{\text{op}} \le 1 + R/((1-\epsilon)n) \le 1 + 2(R' + n\epsilon)/n \le (2R')/n + 3.$$

Note that $3 \le \delta^2/\epsilon$ when $\delta \ge \sqrt{3\epsilon}$. We now apply Talagrand's concentration inequality on R', as each term is bounded by Q^2. We defer the details to Lemma 3.2.5 below, showing that $R'/n = O(\delta^2/\epsilon)$ with probability $1 - \exp(-cn\epsilon)$. By taking a union bound, we get that both $R'/n = O(\delta^2/\epsilon)$ and $\mathcal{E}$ hold with high probability. □

We provide the details of concentration of the empirical process, related to the variance in Lemma 3.2.4, which was omitted above.

Lemma 3.2.5. *Consider the setting in the proof of Lemma 3.2.4. Then, with probability* $1 - \exp(-n\epsilon)$, $R'/n \le \delta^2/\epsilon$, *where* $\delta = O(\sqrt{(\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\epsilon})$.

Proof. We will apply Talagrand's concentration inequality for the bounded empirical process, see Theorem A.2.1. We first calculate the quantity σ^2, the wimpy variance, required in Theorem A.2.1 below

$$\begin{aligned}\sigma^2 &= \sup_{M\in M} \sum_{i=1}^n \mathbf{Var}(f(x_i^\top M x_i)) \le \sup_{M\in M} \sum_{i=1}^n \mathbb{E}(f(x_i^\top M x_i))^2 \le \sup_{M\in M} \sum_{i=1}^n Q^2\, \mathbb{E}\, f(x_i^\top M x_i) \\ &\le nQ^2,\end{aligned}$$

where we use that $f(x) \le Q^2$, $f(x) \le x$, and $\mathbb{E}\, x^\top M x \le 1$. We now focus our attention to $\mathbb{E}\, R'$. Let ξ_i be n i.i.d. Rademacher random variables, independent of $x_1, \ldots, x_n$. We use contraction and symmetrization properties for Rademacher averages [LT91; BLM13] to get

$$\begin{aligned}\mathbb{E}\, R' &= \mathbb{E} \sup_{M\in\mathcal{M}} \sum_{i=1}^n f(x_i^\top M x_i) - \mathbb{E}\, f(x_i^\top M x_i) \le 2\,\mathbb{E} \sup_{M\in\mathcal{M}} \sum_{i=1}^n \xi_i f(x_i^\top M x_i) \\ &\le 2\,\mathbb{E} \sup_{M\in\mathcal{M}} \sum_{i=1}^n \xi_i x_i^\top M x_i = 2\,\mathbb{E} \left\| \sum_{i=1}^n \xi_i x_i x_i^\top \right\|_{\mathrm{op}} \\ &= O\left(\sqrt{\frac{n\,\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)}{\epsilon}} + \frac{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)}{\epsilon} \right),\end{aligned} \tag{3.6}$$

where the last step uses the refined version of matrix-Bernstein inequality [Min17], stated in Theorem A.2.2, with $L = O(\mathrm{tr}(\Sigma)/\epsilon)$.

Note that the empirical process R' is bounded by Q^2. By applying Talagrand's concentration inequality for bounded empirical processes (Theorem A.2.1), with probability at least $1 - \exp(-n\epsilon)$, we have

$$R' = O\left(\mathbb{E}\, R' + \sqrt{nQ^2}\sqrt{n\epsilon} + Q^2 n\epsilon \right)$$

$$\implies \frac{R'}{n} = O\left(\frac{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)}{n\epsilon} + \sqrt{\frac{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)}{n\epsilon}} + Q\sqrt{\epsilon} + \epsilon Q^2\right)$$
$$= \frac{1}{\epsilon} O\left(\frac{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)}{n} + \sqrt{\frac{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)}{n}}\sqrt{\epsilon} + Q\epsilon\sqrt{\epsilon} + (\epsilon Q)^2\right)$$
$$= \frac{1}{\epsilon}\left(O\left(\sqrt{\frac{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)}{n}} + \sqrt{\epsilon} + \epsilon Q\right)\right)^2$$
$$= \frac{\delta^2}{\epsilon},$$

where $\delta = O(\sqrt{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)/n} + \sqrt{\epsilon} + \epsilon Q) = O(\sqrt{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)/n} + \sqrt{\epsilon})$, where we use the fact that $\epsilon Q = O(\sqrt{\epsilon} + \sqrt{\mathrm{tr}(\Sigma)/n})$. □

3.2.2 Controlling the Mean

Suppose $u^* \in \Delta_{n,\epsilon}$ achieves the minimum in Lemma 3.2.4, i.e., $\|\overline{\Sigma}_{u^*} - I\|_{\mathrm{op}} \leq \delta^2/\epsilon$. It is not necessary that $\|\mu_{u^*} - \mu\|_2 \leq \delta$. Recall that our aim is to find a $w \in \Delta_{n,\epsilon}$ that satisfies the conditions: (i) $\|\mu_w - \mu\|_2 \leq \delta$, and (ii) $\|\overline{\Sigma}_w - I\|_{\mathrm{op}} \leq \delta^2/\epsilon$. Given u^*, we will remove additional $O(\epsilon)$-fraction of probability mass from u^* to obtain a $w \in \Delta_n$ such that $\|\mu_w - \mu\|_2 \leq \delta$. For $u \in \Delta_n$, consider the following set of distributions:

$$\Delta_{n,\epsilon,u} := \left\{w : \sum_{i=1}^{n} w_i = 1, w_i \leq u_i/(1-\epsilon)\right\}.$$

For any $w \in \Delta_{n,\epsilon,u^*}$, we directly obtain that $\Sigma_w \preceq \Sigma_{u^*}/(1-\epsilon)$. Our main result in this subsection is that, with high probability, there exists a $w^* \in \Delta_{n,4\epsilon,u^*}$ such that $\|\mu_{w^*} - \mu\|_2 \leq \delta$. We first prove an intermediate result, Lemma 3.2.6 below, that uses the truncation (Lemma 3.2.3) and simplifies the constraint $\Delta_{n,4\epsilon,u^*}$. Let $g : \mathbb{R} \to \mathbb{R}$ be the

following thresholding function:

$$g(x) = \begin{cases} x, & \text{if } x \in [-Q, Q], \\ Q, & \text{if } x > Q, \\ -Q, & \text{if } x < -Q. \end{cases} \tag{3.7}$$

Lemma 3.2.6. *Let $w \in \Delta_{n,\epsilon}$ for some $\epsilon \leq 1/2$. Suppose that the following event $\mathcal{E}$ holds:*

$$\mathcal{E} := \left\{ \sup_{M \in \mathcal{M}} |\{i : (x_i - \mu)^\top M (x_i - \mu) \geq Q^2\}| \leq \epsilon n \right\}.$$

For a unit vector v, let $S_v \in [n]$ be the following multiset: $S_v = \{x_i : x_i \in S, |x_i^\top v| \leq Q\}$. For a unit vector v, let $\overline{w}^{(v)}$ be the following distribution:

$$\tilde{w}_i^{(v)} := \min\left(w_i, \frac{\mathbb{I}\{x_i \in S_v\}}{|S_v|}\right), \qquad \overline{w}^{(v)} := \frac{\tilde{w}^{(v)}}{\|\tilde{w}^{(v)}\|_1}. \tag{3.8}$$

Let $g(\cdot)$ be defined as in Eq. (3.7). Then, for all unit vectors v, $\overline{w}^{(v)} \in \Delta_{n,4\epsilon,w}$. Moreover, the following inequalities hold:

$$\left|\sum_{i=1}^n \overline{w}_i^{(v)} v^\top (x_i - \mu)\right| \leq 4\epsilon Q + \left|\frac{\sum_{i \in S_v} v^\top (x_i - \mu)}{|S_v|}\right| \leq 5\epsilon Q + \left|\frac{\sum_{i \in S} g(v^\top (x_i - \mu))}{(1-\epsilon)n}\right|.$$

Proof. On the event $\mathcal{E}$, we have that $|S_v| \geq (1-\epsilon)n$ for all $v \in \mathcal{S}^{d-1}$. In order to show that $\overline{w}^{(v)} \in \Delta_{n,4\epsilon,w}$, it suffices to show that for all v, $\overline{w}_i^{(v)} \leq w_i/(1-4\epsilon)$. By the definition of $\overline{w}_i^{(v)}$, it is sufficient to show that $\|\tilde{w}^{(v)}\|_1 \geq 1 - 4\epsilon$. Let u_S and u_{S_v} denote the uniform distributions on the multi-sets S and S_v respectively. Let $d_{\mathrm{TV}}(p, q)$ denote the total

variation distance between the distributions p and q. First note that

$$d_{\mathrm{TV}}(w, u_{S_v}) \leq d_{\mathrm{TV}}(w, u_S) + d_{\mathrm{TV}}(u_S, u_{S_v}) \leq \frac{\epsilon}{1-\epsilon} + \frac{\epsilon}{1-\epsilon} \leq \frac{2\epsilon}{1-\epsilon} \leq 4\epsilon. \tag{3.9}$$

We now use the alternative characterization of total variation distance (see, e.g., [Tsy09, Lemma 2.1]):

$$d_{\mathrm{TV}}(p, q) = (1/2)\sum_{i=1}^{n} |p_i - q_i| = 1 - \sum_{i=1}^{n} \min(p_i, q_i).$$

Observe that $\tilde{w}^{(v)} = \min(w, u_{S_v})$; combining this observation with Eq. (3.9), we get the following lower bound on $\|\tilde{w}^{(v)}\|_1$:

$$\|\tilde{w}^{(v)}\|_1 = 1 - d_{\mathrm{TV}}(w, u_{S_v}) \geq 1 - 4\epsilon.$$

This concludes that $\overline{w}^{(v)} \in \Delta_{n,4\epsilon,w}$. We now focus our attention on the second result in the theorem statement. The first inequality follows from the fact that both distributions $\overline{w}^{(v)}$ and u_{S_v} have total variation distance less than 4ϵ, and supported on $[-Q, Q]$. The second inequality follows from the fact that (i) $|S_v| \geq (1-\epsilon)n$, (ii) $g(\cdot)$ is identity on S_v, and bounded by Q outside $[-Q, Q]$, and (iii) at most ϵ-fraction of the points are outside S_v. This completes the proof. □

Using Lemma 3.2.6, we prove the following:

Lemma 3.2.7. *Let $x_1, \ldots, x_n$ be n i.i.d. points from a distribution in $\mathbb{R}^d$ with mean μ and covariance $\Sigma \preceq I$. Let $0 < \epsilon < 1/2$ and $u \in \Delta_{n,\epsilon}$. Then, for a constant $c > 0$, the following holds with probability $1 - \exp(-cn\epsilon)$: $\min_{w \in \Delta_{n,4\epsilon,u}} \|\mu_w - \mu\|_2 \leq \delta$, where $\delta = O\left(\sqrt{\epsilon} + \sqrt{\mathrm{tr}(\Sigma)/n}\right)$.*

At a high-level, the proof of Lemma 3.2.7 proceeds as follows: We use duality and the variational characterization of the ℓ_2 norm to reduce our problem to an empirical

process over projections. We then use Lemma 3.2.6 to simplify the domain constraint $\Delta_{n,4\epsilon,u^*}$ and obtain a bounded empirical process, with an overhead of $O(\epsilon Q) = O(\delta)$.

Proof. (Proof of Lemma 3.2.7) Let Δ be the set $\Delta_{n,4\epsilon,u}$ and assume that $\mu = 0$ without loss of generality. On the event $\mathcal{E}$ (defined in Lemma 3.2.3), using minimax duality and Claim 3.2.6, we get

$$\min_{w\in\Delta}\max_{v\in\mathcal{S}^{d-1}}\sum_{i=1}^{n} w_i x_i^\top v = \max_{v\in\mathcal{S}^{d-1}}\min_{w\in\Delta}\sum_{i=1}^{n} w_i x_i^\top v \le 5\epsilon Q + \max_{v\in\mathcal{S}^{d-1}}|\sum_{i\in[n]} 2g(v^\top x_i)/n|. \tag{3.10}$$

We define the following empirical processes:

$$N = \sup_{v\in\mathcal{S}^{d-1}}\sum_{i=1}^{n} g(v^\top x_i), \qquad N' = \sup_{v\in\mathcal{S}^{d-1}}\sum_{i=1}^{n} g(v^\top x_i) - \mathbb{E}[g(v^\top x_i)].$$

As $g(\cdot)$ is an odd function and $\mathcal{S}^{d-1}$ is an even set, we get that both N and N' are non-negative. For any $v \in \mathcal{S}^{d-1}$, note that $v^\top x$ has variance at most 1 and $\mathbb{P}(|v^\top x| \ge Q) = O(\epsilon)$. We can thus bound $\mathbb{E}\, g(v^\top x)$ as $O(\sqrt{\epsilon}) = O(\epsilon Q)$ (see Proposition A.2.3). This gives us that $|N - N'| = O(n\epsilon Q)$. Using the variational form of the ℓ_2 norm with Eq. (3.10) leads to the following inequality in terms of N':

$$\min_{w\in\Delta}\|\mu_w\|_2 = \max_{v\in\mathcal{S}^{d-1}}\min_{w\in\Delta}\sum_{i=1}^{n} w_i x_i^\top v \le 5\epsilon Q + N/((1-\epsilon)n) = O(\epsilon\, Q) + (2N')/n\ .$$

Note that the term ϵQ is small as $\epsilon Q = O(\delta)$. As N' is a bounded empirical process, with the bound Q, we can apply Talagrand's concentration inequality. We defer the details to Lemma 3.2.8 below, showing that $N'/n = O(\sqrt{\mathrm{tr}(\Sigma)/n} + \sqrt{\epsilon}) = O(\delta)$. Taking a union bound over concentration of N' and the event $\mathcal{E}$, we get that the desired result holds with high probability. □

Lemma 3.2.8. *Consider the setting in Lemma 3.2.7. Then, with probability,* $1 - \exp(-n\epsilon)$, $R'/n = O(\sqrt{\mathrm{tr}(\Sigma)/n} + \sqrt{\epsilon})$.

Proof. We will use Talagrand's concentration inequality for bounded empirical processes, stated in Theorem A.2.1. We first calculate the wimpy variance required for Theorem A.2.1,

$$\sigma^2 = \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n \mathbf{Var}(g(x_i^\top v)) \le \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n \mathbb{E}\, g(v^\top x_i)^2 \le \sup_{v \in \mathcal{S}^{d-1}} n\, \mathbb{E}(v^\top x_i)^2 \le n. \tag{3.11}$$

We also bound the quantity $\mathbb{E}\, R'$ using symmetrization and contraction [LT91; BLM13] properties of Rademacher averages. We have that

$$\begin{aligned}\mathbb{E}\, R' &= \mathbb{E} \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n g(v^\top x_i) - \mathbb{E}\, g(v^\top x_i) \le 2\, \mathbb{E} \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n \epsilon_i g(v^\top x_i) \\ &\le 2\, \mathbb{E} \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n \epsilon_i v^\top x_i = 2\, \mathbb{E} \left\| \sum_{i=1}^n \epsilon_i x_i \right\|_2 \le 2\sqrt{n \operatorname{tr}(\Sigma)},\end{aligned}$$

where the last step uses that $\epsilon_i x_i$ has covariance Σ. By applying Talagrand's concentration inequality for bounded empirical processes (Theorem A.2.1), we get that with probability at least $1 - \exp(-n\epsilon)$,

$$R'/n = O(\mathbb{E}\, R'/n + \sqrt{n\epsilon} + Q\epsilon) = O(\sqrt{\operatorname{tr}(\Sigma)/n} + \sqrt{\epsilon}).$$

□

3.2.3 Proof of Theorem 3.1.4

We first state a result stating that deterministic rounding of weights suffice, proved in Appendix A.4.2.

Lemma 3.2.9. *For $\epsilon \le \frac{1}{3}$, let $w \in \Delta_{n,\epsilon}$ be such that for $\epsilon \le \delta$, we have (i) $\|\mu_w - \mu\|_2 \le \delta$ and (ii) $\|\overline{\Sigma}_w - I\|_{\text{op}} \le \delta^2/\epsilon$. Then there exists a subset $S_1 \subseteq S$ such that*

1. $|S_1| \ge (1 - 2\epsilon)|S|$.

2. *S_1 is (ϵ', δ') stable with respect to μ and $\sigma^2 = 1$, where $\delta' = O(\delta + \sqrt{\epsilon} + \sqrt{\epsilon'})$.*

In the following, we combine the results in the previous lemmas to obtain the stability of a subset with high probability. We first give a proof sketch.

Proof Sketch of Theorem 3.1.4 By Lemma 3.2.4, we get that there exists a $u^* \in \Delta_{n,\epsilon}$ such that $\|\overline{\Sigma}_{u^*} - I\|_{\text{op}} \leq \delta^2/\epsilon$. Applying Lemma 3.2.7 with this u^*, we get that there exists a $w^* \in \Delta_{n,4\epsilon,u^*}$ such that $\|\mu_{u^*} - \mu\|_2 \leq \delta$. $v^\top \overline{\Sigma}_{w^*} v \leq (1/(1-4\epsilon)) v^\top \overline{\Sigma}_{u^*} v = O(\delta^2/\epsilon)$, for small enough ϵ. To obtain a discrete set, we show that rounding w^* to a discrete set only leads to slightly worse constants.

We are now ready to prove our main theorem, which we restate for completeness.

Theorem 3.2.10 (Theorem 3.1.4)**.** *Let $x_1, \ldots, x_n$ be n i.i.d. points in $\mathbb{R}^d$ from a distribution with mean μ and covariance Σ. Let $\epsilon' = O(\log(1/\tau)/n + \epsilon) \leq c$ for a sufficiently small positive constant c. Then, with probability at least $1 - \tau$, there exists a subset $S' \subseteq S$ s.t. $|S|' \geq (1-\epsilon')n$ and $|S'|$ is $(C\epsilon', \delta)$-stable with respect to μ and $\|\Sigma\|_{\text{op}}$ with $\delta = O(\sqrt{(\mathrm{r}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{C\epsilon'})$.*

Proof. Note that we can assume without loss of generality that $\mu = 0$ and $\|\Sigma\|_{\text{op}} = 1$, upper bound δ by $\delta = O(\sqrt{\mathrm{tr}(\Sigma)\log(\mathrm{r}(\Sigma))/n} + \sqrt{C\epsilon'})$; otherwise, apply the following arguments to the random variable $(x_i - \mu)/\sqrt{\|\Sigma\|_{\text{op}}}$ (the result holds trivially if $\|\Sigma\|_{\text{op}} = 0$).

We first prove a simpler version of the theorem for distributions with bounded support. The reason we make this assumption is to apply the matrix concentration results in Theorem A.2.2.

Base case: Bounded support Assume that $\|x_i - \mu\|_2 = O(\sqrt{\mathrm{tr}(\Sigma)/\epsilon'})$ almost surely.

Note that the bounded support assumption allows us to apply Lemma 3.2.4. Set $\tilde{\epsilon} = \epsilon'/c'$ for a large constant c' to be determined later. Let $u^* \in \Delta_{n,\tilde{\epsilon}}$ achieve the minimum in Lemma 3.2.4. For this u^*, let $w^* \in \Delta_{n,4\tilde{\epsilon},u^*}$ be the distribution achieving the minimum

in Lemma 3.2.7. Note that the probability of error is at most $2\exp(-\Omega(n\tilde{\epsilon}))$. We can choose ϵ' large enough, $\tilde{\epsilon} = \epsilon'/c = \Omega(\log(1/\tau)/n)$, so that the probability of failure is at most $1-\tau$. Let $\delta = C\sqrt{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)/n} + C\sqrt{\tilde{\epsilon}}$ for a large enough constant C to be determined later. We first look at the variance of w^* using the guarantee of u^* in Lemma 3.2.4:

$$\sum_{i=1}^n w_i^* x_i x_i^\top \preceq \sum_{i=1}^n \frac{1}{1-\epsilon'} u_i^* x_i x_i^\top \preceq 2\sum_{i=1}^n u_i^* x_i x_i^\top \leq \frac{1}{\tilde{\epsilon}}(C\sqrt{\mathrm{tr}(\Sigma)\log \mathrm{r}(\Sigma)/n} + C\sqrt{\tilde{\epsilon}})^2. \quad (3.12)$$

By choosing C to be a large enough constant, we get that $\|\sum_{i=1}^n w^* x_i x_i^\top - I\|_{\mathrm{op}} \leq \delta^2/\tilde{\epsilon}$. Now, we look at the mean. Lemma 3.2.7 states that

$$\left\|\sum_{i=1}^n w^* x_i\right\|_2 = O\left(\sqrt{\tilde{\epsilon}} + C\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}\right) \leq \delta. \quad (3.13)$$

Since $w^* \in \Delta_{n,4\tilde{\epsilon},u*}$ and $u^* \in \Delta_{n,\tilde{\epsilon}}$, we have that $w^* \in \Delta_{n,5\tilde{\epsilon}}$. Therefore, we have a $w^* \in \Delta_{n,5\tilde{\epsilon}}$ that satisfies the requirements of Lemma A.4.2. Applying Lemma A.4.2, we get the desired statement for a set $S' \subseteq S$. Finally, we can choose the constant c' in the definition of $\tilde{\epsilon}$ large enough, so that the set has cardinality $|S'| \geq (1-\epsilon')n$. This completes the proof for the case of bounded support.

General case We first do a simple truncation. For a large enough constant C', let E be the following event:

$$E = \left\{X : \|X-\mu\|_2 \leq C'\sqrt{\frac{\mathrm{tr}(\Sigma)}{\epsilon'}}\right\}. \quad (3.14)$$

Let Q be the distribution of X conditioned on E. Note that P can be written as a convex combination of two distributions: Q and some distribution R,

$$P = (1-\mathbb{P}(E))Q + \mathbb{P}(E^c)R. \quad (3.15)$$

Let $Z \sim Q$. By Chebyshev's inequality, we get that $\mathbb{P}(E^c) \leq \epsilon'/C'^2$. Using arguments similar to Lemma A.2.5, we get that $\|\mathbb{E}\, Z - \mu\|_2 = O(\sqrt{\epsilon'})$ and $\mathrm{Cov}(Z) \preceq (1/(1-\epsilon))I$. The distribution Q satisfies the assumptions of the base case analyzed above after scaling by $(1/(1-\epsilon)) = \Theta(1)$. Let S_E be the set $\{i : x_i \in E\}$ and let E_1 be the following event:

$$E_1 = \{|S_E| \geq (1 - \epsilon'/2)n\}. \tag{3.16}$$

A Chernoff bound implies that given n samples from P, for a $c > 0$, with probability at least $1 - \exp(-cn\epsilon'/C'^2) \geq 1 - \tau/2$ (by choosing C' large enough and $\epsilon' = \Omega(\log(1/\tau)/n)$), E_1 holds.

For a fixed $m \geq (1 - \epsilon'/2)n$, let $z_1, \ldots, z_m$ be m i.i.d. draws from the distribution Q. Applying the theorem statement of the base case for each such m, we get that, except with probability $\tau/2$, there exists an $S' \subseteq [m] \subseteq [n]$ with $|S'| \geq (1 - \epsilon'/2)m \geq (1 - \epsilon'/2)^2 n \geq (1 - \epsilon')n$, such that $|S'|$ is $(C\epsilon', O(\sqrt{d \log d/n} + \sqrt{C\epsilon'}))$-stable.

As mentioned above (event E_1), $m \geq (1 - \epsilon'/2)n$ with probability at least $1 - \tau/2$. We can now marginalize over m to say that with probability at least $1 - \tau$, there exists a $(C\epsilon', \delta)$ stable set S' of cardinality at least $(1 - \epsilon')n$.

However, we are still not done. We have the guarantee that S' is stable with respect to $\mathbb{E}\, Z$. Using the triangle inequality and Cauchy-Schwarz, we get that the set is also $(C\epsilon', \delta')$ stable with respect to μ as well, where $\delta' = \delta + \|\mu - \mathbb{E}\, Z\|_2 = \delta + O(\sqrt{\epsilon'})$. This completes the proof. □

3.3 Robust Mean Estimation using Median-of-Means Principle

In this section, we again consider distributions with finite covariance matrix Σ. We now turn our attention to the proof of Theorem 3.1.7 that removes the additional logarithmic

factor $\sqrt{\log(r(\Sigma))}$. In Section 3.3.1, we show a result stating that pre-processing on i.i.d. points yields a set that contains a large stable subset (after rescaling). Then, in Section 3.3.2, we use a coupling argument to show a similar result in the strong contamination model.

We recall the median of means principle. Let $k \in [n]$.

1. First randomly bucket the data into k disjoint buckets of equal size (if k does not divide n, remove some samples) and compute their empirical means $z_1, \ldots, z_k$.
2. Output (appropriately defined) multivariate median of $z_1, \ldots, z_k$.

3.3.1 Stability of Uncorrupted Data

We first recall the result (with different constants) from Depersin and Lecué [DL22b] in a slightly different notation.

Theorem 3.3.1 ([DL22b, Proposition 1])**.** *Let $z_1, \ldots, z_k$ be k points in $\mathbb{R}^d$ obtained by the median-of-means preprocessing on n i.i.d. data $x_1, \ldots, x_n$ from a distribution with mean μ and covariance Σ. Let $\mathcal{M}$ be the set of PSD matrices with trace at most 1. Then, there exists a constant $c > 0$, such that with probability at least $1 - \exp(-ck)$, we have that for all $M \in \mathcal{M}$, $\left|\{i \in [k] : (z_i - \mu)^\top M (z_i - \mu) > (k\|\Sigma\|/n)\delta^2\}\right| \leq \frac{k}{100}$, where $\delta = O(\sqrt{\mathrm{r}(\Sigma)/k} + 1)$.*

We now state our main result in this section, proved using minimax duality, that Theorem 3.3.1 implies stability. We first consider the case of i.i.d. data points, as it conveys the underlying idea clearly.

Theorem 3.3.2. *Let $x_1, \ldots, x_n$ be n i.i.d. random variables from a distribution with mean μ and covariance $\Sigma \preceq I$. For $k \in [n]$, let $z_1, \ldots, z_k$ be the variables obtained by median-of-means preprocessing. Then, with probability $1 - \exp(-ck)$, where c is a positive universal constant, there exists a set $S_1 \subseteq [k]$ and $|S_1| \geq 0.95k$ such that S_1 is $(0.1, \delta)$-stable with respect to μ and $k\|\Sigma\|/n$, where $\delta = O(\sqrt{\mathrm{r}(\Sigma)/k} + 1)$.*

Proof. For brevity, let $\sigma = \sqrt{k\|\Sigma\|_{\mathrm{op}}/n}$. Suppose that the conclusion in Theorem 3.3.1 holds with $\delta = O(\sqrt{\mathrm{r}(\Sigma)/k} + 1)$ such that $\delta \geq 1$, i.e., for every $M \in \mathcal{M}$, for at least $0.99k$ points $(z_i - \mu)^\top M (z_i - \mu) \leq \sigma^2\delta^2$. Using minimax duality, we get that

$$\begin{aligned}
\min_{w \in \Delta_{k,0.01}} \left\| \sum_{i=1}^k w_i (z_i - \mu)(z_i - \mu)^\top \right\|_{\mathrm{op}} &= \min_{w \in \Delta_{k,0.01}} \max_{M \in \mathcal{M}} \left\langle M, \sum_{i=1}^k w_i (z_i - \mu)(z_i - \mu)^\top \right\rangle \\
&= \max_{M \in \mathcal{M}} \min_{w \in \Delta_{k,0.01}} \left\langle M, \sum_{i=1}^k w_i (z_i - \mu)(z_i - \mu)^\top \right\rangle \\
&\leq \sigma^2\delta^2,
\end{aligned}$$

where the last step uses the conclusion of Theorem 3.3.1. As $\delta^2 \geq 1$, we also get that $\| \sum_{i=1}^k w_i^*(z_i - \mu)(z_i - \mu)^\top - \sigma^2 I\| \leq \sigma^2\delta^2$. Let w^* be the distribution that achieves the minimum in the above statement. We can also upper bound the first moment of w^* using the bound on the second moment of w^* as follows:

$$\sum_{i=1}^k w_i^* v^\top (z_i - \mu) \leq \sqrt{\sum_{i=1}^k w_i^* (v^\top (z_i - \mu))^2} \leq \sqrt{\left\| \sum_{i=1} w_i^* (z_i - \mu)(z_i - \mu)^\top \right\|_{\mathrm{op}}} \leq \sqrt{\sigma^2\delta^2} = \sigma\delta.$$

Given this $w^* \in \Delta_{k,0.01}$, we will now obtain a subset of $\{z_1, \ldots, z_k\}$ that satisfies the stability condition. In particular, Lemma A.4.2 shows that we can deterministically round w^* such that there exists a large stable subset of $\{z_1, \ldots, z_k\}$ which is $(0.1, \delta)$ stable with respect to μ and σ^2. □

3.3.2 Stability Under Strong Contamination Model

We now prove Theorem 3.1.7, i.e., stability of a subset after corruption, using Theorem 3.3.2. The following result shares the same principle as [DHL19, Lemma B.1]: we add a coupling argument because the pre-processing step (random bucketing) introduces an additional source of randomness.

Theorem 3.3.3. (*Formal statement of Theorem 3.1.7*) *Let T be an ϵ-corrupted version of the set S, where S is a set of n i.i.d. points from a distribution P with mean μ and covariance Σ. Set $\epsilon' = \Theta(\epsilon + \log(1/\tau)/n)$ and set $k = \lfloor \epsilon' n \rfloor$. Let T_k be the set of k points obtained by median-of-means preprocessing on the set T. Then, with probability $1 - \tau$, T_k is 0.01-corruption of a set S_k such that there exists a $S'_k \subseteq S_k$, $|S'_k| \geq 0.95k$ and S'_k is $(0.1, \delta)$ stable with respect to μ and $k\|\Sigma\|_{\text{op}}/n$, where $\delta = O(\sqrt{\text{r}(\Sigma)/k} + 1)$.*

Proof. For simplicity, assume k divides n and let $m = n/k$.

Let $S = \{x_1, \ldots, x_n\}$ be the multiset of n i.i.d. points in $\mathbb{R}^d$ from P. We can write T as $T = \{x'_1, \ldots, x'_n\}$ such that $|\{i : x'_i \neq x_i\}| \leq \epsilon n$.

As the algorithm only gets a multiset, we first order them arbitrarily. Let $r'_1, \ldots, r'_n$ be any arbitrary labelling of points and let $\sigma_1(\cdot)$ be the permutation such that $r'_i = x'_{\sigma_1(i)}$. We now split the points randomly into buckets by randomly shuffling them. Let $\sigma(\cdot)$ be a uniformly random permutation of $[n]$ independent of T (and S). Define $w'_i = r'_{\sigma(i)} = x'_{\sigma_1(\sigma(i))}$. For $i \in [k]$, define the bucket B'_i to be the multiset $B'_i := \{w'_{(i-1)m+1}, \ldots, w'_{im}\}$. For $i \in [k]$, define z'_i to be the mean of the set B'_i, i.e., $z_i = \mu_{B'_i}$. That is, the input to the stable algorithm would be the multiset T_k, where $T_k = \{z'_1, \ldots, z'_k\}$.

We now couple the corrupted points with the original points. For σ and σ_1, define their composition σ' as $\sigma'(i) := \sigma_1(\sigma(i))$. Define $r_i := x_{\sigma_1(i)}$ and $w_i := r_{\sigma(i)} = x_{\sigma'(i)}$. Importantly, Proposition 3.3.4 below states that w_i's are i.i.d. from P. The analogous bucket for uncorrupted samples is $B_i := \{w_{(i-1)m+1}, \ldots, w_{im}\}$. For $i \in [k]$, define $z_i := \mu_{B_i}$ and define S_k to be $\{z_1, \ldots, z_k\}$. Therefore, $z_1, \ldots, z_k$ are obtained from the median-of-means processing of i.i.d. data $w_1, \ldots, w_n$, and thus Theorem 3.3.2 holds[5]. That is, there exists $S'_k \subseteq S_k$ that satisfies the desired properties.

It remains to show that T_k is a corruption of S_k. It is easy to see that $|T_k \cap S_k| \geq$

[5]If $(x_1, \ldots, x_n)$ are i.i.d., then choosing the buckets $B_i = \{x_{(i-1)m}, \ldots, x_{im}\}$ for $i \in [k]$ preserves independence. In particular, any partition of k sets of equal cardinality that does not depend on the values of $(x_1, \ldots, x_n)$ suffices. Therefore, Theorem 3.3.1 and Theorem 3.3.2 hold for this bucketing strategy too.

$k - \epsilon n \geq 0.99k$, by choosing ϵ' large enough. That is, for any σ_1 and σ, T_k is at most (0.01)-contamination of the set S_k. □

Proposition 3.3.4. *Let $x_1, \ldots, x_n$ be n i.i.d. points from a distribution P and $\sigma_1(\cdot)$ be a permutation, potentially depending on $x_1, \ldots, x_n$. Let $\sigma(\cdot)$ be a random permutation independent of $x_1, \ldots, x_n$ and $\sigma_1(\cdot)$. Define the composition permutation be $\sigma'(i) := \sigma_1(\sigma(i))$. Then $x_{\sigma'(1)}, \ldots, x_{\sigma'(n)}$ are also i.i.d. from the distribution P.*

Proof. First observe that $\sigma'(\cdot)$ is a uniform random permutation independent of $x_1, \ldots, x_n$. The result follows from the following fact:

Fact 3.3.5. *Let $x_1, \ldots, x_n$ be n i.i.d. points from a distribution P. Let $\sigma(\cdot)$ be a random permutation independent of $x_1, \ldots, x_n$, then $x_{\sigma(1)}, \ldots, x_{\sigma(n)}$ are also i.i.d. from the distribution P.*

□

3.4 Robust Mean Estimation Under Finite Central Moments

In this section, we consider distributions with identity covariance and bounded central moments. Our main result in this section is the proof of Theorem 3.1.8, which obtains a tighter dependence on ϵ. Our proof strategy closely follows the proof structure of the bounded covariance case. We suggest the reader to read Section 3.2 before reading this section. This section has a similar organization to Section 3.2. We start with a simplified stability condition in Lemma 3.4.2. Sections 3.4.1 and 3.4.2 contain the arguments for controlling the second moment matrix from above and below respectively. Section 3.4.3 contains the results regarding the concentration results for controlling the sample mean.

Finally, we combine the results of the previous sections in Section 3.4.4 to complete the proof of Theorem 3.1.8.

In the bounded covariance setting, we considered δ such that $\delta = \Omega(\sqrt{\epsilon})$. As such, we only needed an upper bound on second moment matrix, $\overline{\Sigma}_{S'}$, for a set $S' \subseteq S$ (For $\delta \geq \sqrt{\epsilon}$, the lower bound in the second condition of stability is trivial). For $\delta = o(\sqrt{\epsilon})$, we need a *sharp* lower bound on the minimum eigenvalue of $\overline{\Sigma}_{S_1}$ for *all large subsets* S_1 of a set S'. Such a result is not possible in general, unless we impose both: (i) identity covariance and (ii) tighter control on tails of X.

We will prove the existence of a stable set with high probability using the following claim. This is analogous to Claim 3.2.1 in the bounded covariance setting. In particular, we also need a lower bound on the minimum eigenvalue of $\overline{\Sigma}_{S'}$ for all large subsets S'.

Claim 3.4.1. *Let $0 \leq \epsilon \leq \delta$ and $\epsilon \leq 0.5$. A set S is $(\epsilon, O(\delta))$ stable with respect to μ and $\sigma^2 = 1$, if it satisfies the following for all unit vectors v.*

1. $\|\mu_S - \mu\|_2 \leq \delta$.
2. $v^\top \overline{\Sigma}_S v \leq 1 + \delta^2/\epsilon$.
3. *For all subsets $S' \subseteq S : |S'| \geq (1-\epsilon)|S|$, $v^\top \overline{\Sigma}_{S'} v \geq (1 - \delta^2/\epsilon)$.*

The proof of Claim 3.4.1 is provided in Appendix A.5.1.

3.4.1 Upper Bound on the Second Moment Matrix

For simplicity, we will state our probabilistic results directly in terms of d instead of $\mathrm{tr}(\Sigma)$ and $\mathrm{r}(\Sigma)$. The proof techniques of Section 3.2 can directly be translated to obtain results in terms of Σ. We follow the same strategy as in Section 3.2.1. We first refine the bound on the truncation threshold in the following result, proved in Appendix A.3.2.

Lemma 3.4.2. *Consider the setting in Theorem 3.1.8. Let* $Q_k = \Theta(\sigma_k \epsilon^{-1/k} + (1/\epsilon)\sqrt{\mathrm{tr}(\Sigma)/n})$. *For each* $M \in \mathcal{M}$, *let* S_M *be the set* $\{i : (x_i - \mu)^\top M(x_i - \mu) \geq Q_k^2\}$. *Let* $\mathcal{E}$ *be the event* $\mathcal{E} = \{\sup_{M \in \mathcal{M}} |S_M| \leq \epsilon n\}$. *Then for a* $c > 0$, *with probability at least* $1 - \exp(-c\epsilon n)$, *event* $\mathcal{E}$ *holds.*

We first find a subset such that its covariance matrix is bounded. For technical reasons, we do not assume that the covariance is exactly identity and allow some slack. The argument is similar to Lemma 3.2.4 for the bounded covariance. We also impose some additional constraints to simplify the expression, as those regimes would not hold anyway in the proof.

Lemma 3.4.3. *Let* $x_1, \ldots, x_n$ *be* n *i.i.d. points in* $\mathbb{R}^d$ *from a distribution with mean* μ, *covariance* Σ, *and for a* $k \geq 4$, *the* k*-th central moment is bounded by* σ_k. *Further assume that for* $\epsilon < 0.5$, *covariance matrix* Σ *satisfies that* $(1 - 2\sigma_k^2 \epsilon^{1-\frac{2}{k}}) \preceq \Sigma \preceq I$. *Further assume the following conditions hold:*

1. $\log(1/\tau)/n = O(\epsilon)$.
2. $\|x_i - \mu\|_2 = O(\sigma_k \sqrt{d} \epsilon^{-1/k})$ *almost surely.*
3. $\sigma_k \epsilon^{\frac{1}{2} - \frac{1}{k}} = O(1)$.

Then, for a $c > 0$, *with probability* $1 - \tau - \exp(-cn\epsilon)$: $\min_{w \in \Delta_{n,\epsilon}} \left\|\overline{\Sigma}_w\right\|_{\mathrm{op}} \leq 1 + \delta^2/\epsilon$, *where* $\delta = O(\sqrt{(d \log d)/n} + \sigma_k \epsilon^{1-\frac{1}{k}} + \sigma_4 \sqrt{\log(1/\tau)/n})$.

Proof. We will assume without loss of generality that $\mu = 0$. We will assume that the event $\mathcal{E}$ in Lemma 3.4.2 holds as it only incurs an additional probability of error of $\exp(-cn\epsilon)$. We use the variational characterization of spectral norm and minimax duality to write the following:

$$\min_{w \in \Delta_{n,\epsilon}} \left\| \sum_i w_i x_i x_i^\top \right\| = \min_{w \in \Delta_{n,\epsilon}} \max_{M \in \mathcal{M}} \sum w_i \langle x_i x_i^\top, M \rangle$$

$$= \max_{M \in \mathcal{M}} \min_{w \in \Delta_{n,\epsilon}} \sum w_i x_i^\top M x_i$$
$$\leq \max_{M \in \mathcal{M}} \sum_{i=1}^{n} \frac{1}{(1-\epsilon)n} (x_i^\top M x_i) \mathop{\mathbb{I}}_{x_i^\top M x_i \leq Q_k^2},$$

where the third inequality uses Lemma 3.4.2, where it chooses the uniform distribution on the set $S_M = \{x_i : x_i^\top M x_i \leq Q_k^2\}$. Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be the following function:

$$f(x) := \begin{cases} x, & \text{if } x \leq Q_k^2 \\ Q_k^2, & \text{otherwise.} \end{cases}$$

Define the following random variables R and R':

$$R = \sup_{M \in \mathcal{M}} \sum_{i=1}^{n} f(x_i^\top M x_i), \qquad R' = \sup_{M \in \mathcal{M}} \sum_{i=1}^{n} f(x_i^\top M x_i) - \mathbb{E}\, f(x_i^\top M x_i).$$

By Lemma A.2.4, we get that $|\mathbb{E}\, f(x_i^\top M x) - 1| \leq 2\sigma_k^2 \epsilon^{1-\frac{2}{k}}$, which gives that

$$|R - n - R'| \leq 2n\sigma_k^2 \epsilon^{1-\frac{2}{k}}.$$

We therefore get that

$$\begin{aligned} \min_{w \in \Delta_{n,\epsilon}} \| \sum_i w_i x_i x_i^\top \|_{\text{op}} - 1 &\leq \max_{M \in \mathcal{M}} \sum_{i=1}^{n} \frac{1}{(1-\epsilon)n} (x_i^\top M x_i) \mathop{\mathbb{I}}_{x_i^\top M x_i \leq Q_k^2} - 1 \\ &\leq \max_{M \in \mathcal{M}} \sum_{i=1}^{n} \frac{1}{(1-\epsilon)n} f(x_i^\top M x_i) - 1 \\ &= \frac{1}{(1-\epsilon)n} R - 1 \\ &\leq \frac{2R'}{n} + 4\sigma_k^2 \epsilon^{1-\frac{2}{k}} + 2\epsilon. \end{aligned}$$

Observe that the last two terms in the above expression are small, i.e., $\sigma_k^2 \epsilon^{1-\frac{1}{k}} + \epsilon = O(\delta^2/\epsilon)$. We next use Lemma A.5.2 in Appendix to conclude that R' concentrates well.

Lemma A.5.2 states that with probability $1-\tau$, $R'/n \le (1/\epsilon)(O(\sqrt{d\log d/n} + \sigma_k\epsilon^{1-\frac{1}{k}} + \sigma_4\sqrt{\log(1/\tau)/n}))^2$. Note that both of the remaining terms are small compared to Overall, we get that

$$\min_{w\in\Delta_{n,\epsilon}} \|\overline{\Sigma}_w\|_{\text{op}} \le 1 + \frac{\delta^2}{\epsilon}\,.$$

Taking a union bound on the event $\mathcal{E}$ and concentration of R' concludes the result. □

3.4.2 Minimum Eigenvalue of Large Subsets

In this section, we prove that under bounded central moments, the minimum eigenvalue of $\Sigma_{S'}$, of each large enough subset S', has a lower bound close to 1. Our result is similar in spirit to Koltchinskii and Mendelson [KM15, Theorem 1.3] that only bounds the eigenvalue of $\overline{\Sigma}_S$. The proof of the following lemma is very similar to the proof of Lemma 3.4.3.

Lemma 3.4.4. *Consider the setting in Lemma 3.4.3. Then, for a constant $c > 0$, with probability $1-\tau-\exp(-cn\epsilon)$, the following holds:*

$$\min_{S':|S'|\ge(1-\epsilon)n} v^\top\overline{\Sigma}_{S'}v \ge 1 - \frac{\delta^2}{\epsilon},$$

where $\delta = O\left(\sqrt{\frac{d\log d}{n}} + \sigma_k\epsilon^{1-\frac{1}{k}} + \sigma_4\sqrt{\frac{\log(\frac{1}{\tau})}{n}}\right)$.

Proof. Without loss of generality, assume that $\mu = 0$. We will assume that event $\mathcal{E}$ from Lemma 3.4.2 holds, with an additional probability of error $\exp(-cn\epsilon)$, that is

$$\sup_{v\in\mathcal{S}^{d-1}} \left|\left\{i : x_i^\top v \ge Q_k\right\}\right| \le n\epsilon.$$

Let f be as defined in the proof of Lemma 3.4.3. For a sequence $y_1,\dots,y_n$, let $y_{(1)},\dots,y_{(n)}$

be its rearrangement in non-decreasing order. For any unit vector v, we have that

$$\begin{aligned}\min_{S':|S'|\geq(1-\epsilon)n} v^\top \overline{\Sigma}_{S'} v \geq \min_{w\in\Delta_{n,\epsilon}} v^\top \overline{\Sigma}_w v &= \min_{w\in\Delta_{n,\epsilon}} \sum_{i=1}^n w_i (x_i^\top v)^2 \\ &\geq \sum_{i=1}^{(1-\epsilon)n} (x_i^\top v)^2_{(i)}/((1-\epsilon)n) \\ &\geq \sum_{i=1}^n (f((x_i^\top v)^2) - Q_k^2\epsilon n)/((1-\epsilon)n),\end{aligned}$$

where we use that at most ϵn points have projections larger than Q_k^2. Thus we get that the minimum eigenvalue of any large subset is lower bounded by:

$$\min_{w\in\Delta_{n,\epsilon}} \min_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n w_i (x_i^\top v)^2 \geq \min_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n f((x_i^\top v)^2) - Q_k^2\epsilon n.$$

Let $h(\cdot)$ be the negative of the function $f(\cdot)$. Define the following random variable Z and its counterpart Z':

$$Z := \sup_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n h((x_i^\top v)^2), \qquad Z' := \sup_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n h((x_i^\top v)^2) - \mathbb{E}\, h((x_i^\top v)^2)$$

From Lemma A.2.4, it follows that $|\,\mathbb{E}\, h((x_i^\top v)^2) + 1| = |\,\mathbb{E}\, f((x_i^\top v)^2) - 1| = O(\sigma_k^2 \epsilon^{1-\frac{2}{k}})$. This immediately gives us that

$$|Z' - Z - n| = O(n\sigma_k^2 \epsilon^{1-\frac{2}{k}}).$$

Therefore, the desired quantity satisfies the following inequalities:

$$\begin{aligned}(1-\epsilon)n \min_{w\in\Delta_{n,\epsilon}} \min_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n w_i (x_i^\top v)^2 &\geq \min_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n f((x_i^\top v)^2) - Q_k^2\epsilon n \\ &= -\sup_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n h((x_i^\top v)^2) - Q_k^2\epsilon n \\ &= -Z - Q_k^2\epsilon n\end{aligned}$$

$$\geq -Z' + n - O(n\sigma_k^2 \epsilon^{1-\frac{2}{k}}) - \epsilon Q_k^2 \epsilon n.$$

We thus require a high probability upper bound on Z'. Note that Z' behaves similarly to R', defined in the proof of Lemma 3.4.3. Similar to the proof of Lemma A.5.2, we get that, with probability at least $1-\tau$,

$$\frac{Z'}{n} \leq \frac{1}{\epsilon}\left(O\left(\sqrt{\frac{d\log d}{n}} + \sigma_k \epsilon^{1-\frac{1}{k}} + \sigma_4\sqrt{\frac{\log(1/\tau)}{n}}\right)\right)^2.$$

Note that the remaining terms $\sigma_k^2\epsilon^{1-\frac{2}{k}} = O(\delta^2/\epsilon)$ and $\epsilon Q_k^2 = O\left(\sigma_k^2\epsilon^{-\frac{2}{k}+\frac{d}{n\epsilon}}\right) = O(\delta^2/\epsilon)$. Therefore, we get the minimum eigenvalue of any large subset is at least

$$\min_{w\in\Delta_{n,\epsilon}} \lambda_{\min}(\overline{\Sigma}_w) \geq 1 - \frac{\delta^2}{\epsilon},$$

where $\delta = O\left(\sqrt{\frac{d\log d}{n}} + \sigma_k\epsilon^{1-\frac{1}{k}} + \sigma_4\sqrt{\frac{\log(\frac{1}{\tau})}{n}}\right)$.

□

3.4.3 Controlling the Mean

Lemmas 3.4.3 and 3.4.4 give a control on the second moment matrix. We will now further remove $O(\epsilon)$ fraction of points to obtain w such that $\|\mu_w - \mu\|_2$ is small.

Lemma 3.4.5. *Let $x_1, \ldots, x_n$ be n i.i.d. random variables from a distribution with mean μ and covariance $\Sigma \preceq I$. Further, assume that the x_i's are drawn from a distribution with k-th bounded central moment σ_k for a $k \geq 4$. Let $u \in \Delta_{n,\epsilon}$. Assume that $\log(1/\tau)/n = O(\epsilon)$. Then, for a constant $c > 0$, the following holds with probability $1 - \tau - \exp(-cn\epsilon)$:*

$$\min_{w\in\Delta_{n,4\epsilon,u}} \|\sum_{i=1}^n w_i x_i - \mu\|_2 = O(\sqrt{d/n} + \sigma_k\epsilon^{1-\frac{1}{k}} + \sqrt{\log(1/\tau)/n}).$$

Proof. Without loss of generality, let us assume that $\mu = 0$. Also, assume that the event $\mathcal{E}$ from Lemma 3.4.2 holds, with the additional error of $\exp(-cn\epsilon)$. Let $g(\cdot)$ be the following function:

$$g(x) = \begin{cases} x, & \text{if } x \in [-Q_k, Q_k] \\ Q_k, & \text{if } x > Q_k \\ -Q_k, & \text{if } x < -Q_k. \end{cases}$$

Let N be the following random variable:

$$N = \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^{n} g(v^\top x_i) = \sup_{v \in \mathcal{S}^{d-1}} \left| \sum_{i=1}^{n} g(v^\top x_i) \right|,$$

where we use that $g(\cdot)$ is an odd function. We also define the following empirical process, where each term is centered:

$$N' = \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^{n} g(v^\top x_i) - \mathbb{E}[g(v^\top x_i)] = \left| \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^{n} g(v^\top x_i) - \mathbb{E}[g(v^\top x_i)] \right|.$$

As $Q_k = \Omega(\sigma_k \epsilon^{-1/k})$, Lemma A.2.4 states that $\sup_v \mathbb{E}\, g(v^\top x) = O(\sigma_k \epsilon^{1-\frac{1}{k}})$, and this gives that

$$|N - N'| = O(n\sigma_k \epsilon^{1-\frac{1}{k}}).$$

We now use duality to write the following:

$$\begin{aligned} \min_{w \in \Delta_{n,\epsilon,u}} \left\| \sum_{i=1}^{n} w_i x_i \right\|_2 &= \min_{w \in \Delta_{n,\epsilon,u}} \max_{v \in \mathcal{S}^{d-1}} \left\langle \sum_{i=1}^{n} w_i x_i, v \right\rangle \\ &= \max_{v \in \mathcal{S}^{d-1}} \min_{w \in \Delta_{n,\epsilon,u}} \left\langle \sum_{i=1}^{n} w_i x_i, v \right\rangle \end{aligned}$$

$$\leq 5\epsilon Q_k + \left|\frac{1}{(1-\epsilon)n}N\right| \leq O(\epsilon Q_k) + O(\sigma_k \epsilon^{1-1/k}) + 2N',$$

where the last step uses Lemma 3.2.6. We now use Lemma A.5.3 to conclude that N' concentrates. Recall that $\epsilon Q_k = O(\sigma_k \epsilon^{1-\frac{1}{k}} + \sqrt{d/n})$. Overall, we get that, with probability $1 - \tau - \exp(-n\epsilon)$, there exists a $w \in \Delta_{n,\epsilon,u}$, such that $\|\sum w_i x_i\|_2 = O(\sqrt{d/n} + \sigma_k \epsilon^{1-\frac{1}{k}} + \sqrt{\log(1/\tau)/n})$. □

3.4.4 Proof of Theorem 3.1.8

We now combine the results in the previous lemmas to obtain the stability of a subset with high probability. Although we prove the following result showing the existence of $(2\epsilon', \delta)$ stable subset, this can generalized to existence of $(C\epsilon, O(\delta))$ stable subset for a large constant C.

Theorem 3.4.6 (Theorem 3.1.8)**.** *Let $S = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be n i.i.d. points from a distribution with mean μ and covariance Σ such that $(1 - 2\sigma_k^2 \gamma^{1-\frac{1}{k}})I \preceq \Sigma \preceq I$. Further assume that for a $k \geq 4$, the k^{th} central moment is bounded by σ_k. Let $\epsilon' = \Theta(\epsilon + \frac{\log(1/\tau)}{n}) \leq c$ for a sufficiently small constant c.*

Then, with probability at least $1 - \tau$, there exists a subset $S' \subseteq S$ s.t. $|S'| \geq (1-\epsilon')n$ and $|S'|$ is $(2\epsilon', \delta)$-stable with $\delta = O(\sigma_k \epsilon^{1-\frac{1}{k}} + \sqrt{\frac{d \log d}{n}} + \sigma_4 \sqrt{\frac{\log(1/\tau)}{n}})$.

Proof. First note that, for the bounded covariance condition, Theorem 3.1.4 already gives a guarantee that, with probability at least $1 - \tau$,

$$\|\widehat{\mu} - \mu\|_2 = O\left(\sqrt{(d \log d)/n} + \sqrt{\epsilon} + \sqrt{\log(1/\tau)/n}\right). \tag{3.17}$$

Therefore, the guarantee of this theorem statement is tighter only in the following

regimes:

$$\log(1/\tau)/n = O(\epsilon), \qquad O(\sigma_k \epsilon^{\frac{1}{2}-\frac{1}{k}}) = O(1), \qquad d \log d/n = O(\epsilon). \tag{3.18}$$

For the rest of the proof, we will assume that all three of these conditions hold. Similar to the proof of Theorem 3.1.4, we will first prove the statement when the samples are bounded. Without loss of generality, we will assume $\mu = 0$.

Base case: Bounded support In this case, we will assume that $\|x_i\|_2 = O(\sigma_k \epsilon^{-1/k}\sqrt{d})$ almost surely. We will use Lemma A.5.1 to show that the set is stable. Set $\tilde{\epsilon} = \epsilon'/C'$ for a large enough constant C' to be determined later.

Note that $x_1, \ldots, x_n$ satisfy the conditions of Lemmas 3.4.4, 3.4.3, and 3.4.5. In particular, we will use Lemma 3.4.4 with $C\tilde{\epsilon}$, where C is large enough. By choosing $\epsilon' = \Omega(\log(1/\tau)/n)$, we get that, with probability $1 - \tau/3$, for any $S' : |S'| \geq (1 - C\tilde{\epsilon})n$ and unit vector v,

$$\frac{\sum_{i \in S'} (v^\top x_i)^2}{|S'|} \geq 1 - \frac{\delta^2}{C\tilde{\epsilon}}. \tag{3.19}$$

We first look at the variance using the guarantee in Lemma 3.4.3: Let $u \in \Delta_{n,\tilde{\epsilon}}$ be the distribution achieving the minimum in Lemma 3.4.3. By choosing $\epsilon' = \Omega(\log(1/\tau)/n)$, we get that with probability $1 - \tau/3$,

$$\sum_{i=1}^{n} u_i (x_i^\top v)^2 \leq 1 + \frac{\delta^2}{\tilde{\epsilon}}. \tag{3.20}$$

We now obtain a guarantee on the mean using Lemma 3.4.5. For this u, let $w \in \Delta_{n,4\tilde{\epsilon},u}$ be the distribution achieving the minimum in Lemma 3.4.5. Then with probability $1 - \tau/3$,

$$\left\| \sum_{i=1}^{n} w_i x_i \right\|_2 \leq \delta. \tag{3.21}$$

Since $u \in \Delta_{n,4\tilde{\epsilon},w}$ and $w \in \Delta_{n,\tilde{\epsilon},u}$, we have that $u \in \Delta_{n,5\tilde{\epsilon}}$. Moreover,

$$\sum_{i=1}^{n} w_i (x_i^\top v)^2 \leq \sum_{i=1}^{n} \frac{u_i}{1-\tilde{\epsilon}} (x_i^\top v) = \frac{1}{1-\tilde{\epsilon}}(1 + \frac{\delta^2}{\tilde{\epsilon}}) \leq 1 + \frac{1}{1-\tilde{\epsilon}}(\tilde{\epsilon} + \frac{\delta^2}{\tilde{\epsilon}}) \leq 1 + \frac{4\delta^2}{\tilde{\epsilon}}. \quad (3.22)$$

Therefore, we have that $u \in \Delta_{n,5\tilde{\epsilon}}$ and satisfies the requirements of Lemma A.5.1, where we note that $r_1 = O(1)$ and $r_2 = O(1)$ to get the desired statement. By a union bound, the failure probability is τ. Finally, we choose C and C' large enough such that the cardinality of the stable set is at least $(1-\epsilon')n$ and it is $(2\epsilon', \delta)$ stable.

General case: Unbounded support We first do a simple truncation. Let E be the following event:

$$E = \{X : \|X\|_2 \leq C\sigma_k \epsilon^{-\frac{1}{k}} \sqrt{d}\}. \quad (3.23)$$

Let Q be the distribution of X conditioned on E. Note that P can be written as convex combination of two distributions: Q and some distribution R,

$$P = (1 - \mathbb{P}(E))Q + \mathbb{P}(E^c)R. \quad (3.24)$$

Let $Z \sim Q$. Using Lemma A.2.5, we get that $\|\mathbb{E}\, Z\|_2 \leq 2\sigma_k \epsilon^{1-\frac{1}{k}}/C^k$ and $(1 - 3\sigma_k^2 \epsilon^{1-\frac{2}{k}}/C^k) \preceq \mathrm{Cov}(Z) \preceq I$. Thus the distribution Q satisfies the assumptions of the base case for $C \geq 2$.

Let S_E be the set $\{X_i : X_i \in E\}$. A Chernoff bound gives that given n samples from P, with probability at least $1 - \exp(-n\epsilon')$,

$$E_1 = \{|S_E| \geq (1 - \epsilon'/2)n\}. \quad (3.25)$$

For a fixed $m \geq (1-\epsilon'/2)n$, let $z_1, \ldots, z_m$ be m i.i.d. draws from the distribution Q.

Applying the theorem statement for Q, as it satisfies the base case above, we get that, with probability at least $1 - \exp(-cm\epsilon')$, there $\exists\ S' \subset [m] : |S'| \geq (1-\epsilon'/2)m \geq (1-\epsilon'/2)^2 n \geq (1-\epsilon')n$, such that S' is $(2\epsilon', \delta')$-stable. This gives us a set S' which is stable with respect to $\mathbb{E}\, Z$. Using triangle inequality, we get that the set S' is (ϵ, δ') stable with respect to μ as well, where $\delta' = \delta + \|\mu - \mathbb{E}\, Z\|_2 = \delta + O(\sigma_k \epsilon^{1-\frac{1}{k}})$.

We can now marginalize over m to get that with probability except $1 - 2\exp(-cn\epsilon')$, the desired claim holds. Choosing $\epsilon' = \Omega(\log(1/\tau)n)$, we can make probability of failure less than τ. □

3.5 Conclusions and Open Problems

In this paper, we showed that a standard stability condition from the recent high-dimensional robust statistics literature suffices to obtain near-subgaussian rates for robust mean estimation in the strong contamination model. With a simple pre-processing (bucketing), this leads to efficient outlier-robust estimators with subgaussian rates under only a bounded covariance assumption. An interesting technical question is whether the extra $\log d$ factor in Theorem 3.1.4 is actually needed. (Our results imply that it is not needed when $\epsilon = \Omega(1)$.) If not, this would imply that stability-based algorithms achieve subgaussian rates without the pre-processing.

4 ROBUST SPARSE MEAN ESTIMATION

थोड़ा है थोड़े की ज़रुरत है
ज़िन्दगी फिर भी यहाँ खूबसूरत है

— गुलज़ार

We study the fundamental task of outlier-robust mean estimation for heavy-tailed distributions in the presence of sparsity. Specifically, given a small number of corrupted samples from a high-dimensional heavy-tailed distribution whose mean μ is guaranteed to be sparse, the goal is to efficiently compute a hypothesis that accurately approximates μ with high probability. Prior work had obtained efficient algorithms for robust sparse mean estimation of light-tailed distributions. In this work, we give the first sample-efficient and polynomial-time robust sparse mean estimator for heavy-tailed distributions under mild moment assumptions. Our algorithm achieves the optimal asymptotic error using a number of samples scaling logarithmically with the ambient dimension. Importantly, the sample complexity of our method is optimal as a function of the failure probability τ, having an *additive* $\log(1/\tau)$ dependence. Our algorithm leverages the stability-based approach from the algorithmic robust statistics literature, with crucial (and necessary) adaptations required in our setting. Our analysis may be of independent interest, involving the delicate design of a (non-spectral) decomposition for positive semi-definite matrices satisfying certain sparsity properties.

4.1 Introduction

4.1.1 Background

One of the most fundamental problem setups in statistics is as follows: given n i.i.d. samples drawn from an unknown distribution P chosen arbitrarily from some known distri-

bution family $\mathcal{P}$, infer some particular property of P from the data. This generic model captures a range of statistical problems of interest, for example, parameter estimation (such as the mean and (co)variance of P), as well as hypothesis testing. While long lines of work have given us a deep understanding of the statistical and computational possibilities and limitations on these problems, these results are not always applicable in real-world settings due to (i) modeling issues, that the underlying distribution P might not actually be in the known family $\mathcal{P}$ but only being close to it, and (ii) the fact that the n samples supplied might be corrupted, for example by nefarious actors in high-stakes applications [ABHHRT72].

The field of *robust statistics* aims to design estimators and testers that can tolerate up to a *constant* fraction of corrupted samples, independent of the potentially high dimensionality of the data [Tuk60; HR09]. Classical works in the field have identified and resolved the statistical limits of problems in this setup, both in terms of constructing estimators and proving impossibility results [Yat85; DL88; DG92; HR09]. However, the proposed estimators were not computationally efficient, often requiring exponential time to compute either in the number of samples or the number of dimensions [HR09].

A recent line of work, originating in the computer science community, has developed the subfield of *algorithmic* robust statistics, aiming to design estimators that not only attain tight statistical guarantees, but are also computable in polynomial time. This line of research has provided computationally and statistically efficient estimators in a variety of problem settings (e.g., mean estimation, covariance estimation, and linear regression) under different assumptions (e.g., the distribution might be assumed to be (sub-)Gaussian, or can be heavy tailed); see [DK19] for a recent survey of results.

The focus of this paper is the robust mean estimation problem under sparsity constraints on the mean vector. Sparsity is an important structural constraint that is both relevant in practice, especially in the face of increasing dimensionality of modern data,

and extensively studied for statistical estimation (see, e.g., the books [HTW15; EK12; van16]). In the specific context of robust sparse mean estimation, prior works have studied the case where the underlying distribution has light-tails, e.g., sub-exponential tails [BDLS17; DKKPS19; CDKGGS22; DKKPP22b]. In particular, the case of a spherical Gaussian distribution is now rather well-understood both in terms of the optimal information-theoretic estimation error, as well as the conjectured *computational-statistical tradeoff* — namely, that there is a gap between the statistical performance of computationally efficient and inefficient estimators [DKS17; BB20]. In this work, we initiate the investigation of outlier-robust sparse mean estimation for *heavy-tailed* distributions, under only mild moment assumptions. Our main result is the first computationally efficient robust mean estimator in the heavy-tailed setting which leverages sparsity to reduce the sample complexity from depending polynomially on the dimensionality to a logarithmic dependence. Importantly, our algorithm also achieves the optimal dependence on the failure probability τ as it tends to 0; see the next two subsections for further discussion.

4.1.2 Problem Setup

We first define the input contamination model before formally stating the statistical problem.

Definition 4.1.1 (Strong Contamination Model)**.** *Given a* corruption *parameter* $\epsilon \in (0, 1/2)$ *and a distribution* P *on uncorrupted samples, an algorithm takes samples from* P *with* ϵ-contamination *as follows:* (i) *The algorithm specifies the number* n *of samples it requires.* (ii) n *i.i.d. samples from* P *are drawn but not yet shown to the algorithm.* (iii) *An arbitrarily powerful adversary then inspects the entirety of the* n *i.i.d. samples, before deciding to replace any subset of* $\lceil \epsilon n \rceil$ *samples with arbitrarily corrupted points, and returning the modified set of* n *samples to the algorithm.*

Define the $\ell_{2,k}$-norm of a vector v, denoted by $\|v\|_{2,k}$, as the ℓ_2-norm of the largest k entries of a vector v in magnitude. The goal is to estimate the mean vector in this sparse norm.

Problem 4.1.2. *Fix a corruption parameter $\epsilon \in (0, 1/2)$, error parameter $\delta > 0$, failure probability $\tau \in (0, 1)$, and distribution family $\mathcal{D}$ over $\mathbb{R}^d$. Suppose we have access to ϵ-contaminated samples drawn from an unknown distribution $P \in \mathcal{D}$ with mean μ. The task is to compute an estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_{2,k}$ is bounded above by error δ with probability at least $1 - \tau$ over n samples. The goal is then to give an estimator with the minimal* sample complexity $n(k, \epsilon, \delta, \tau)$.

The above problem is slightly more general than sparse mean estimation in the following sense. To estimate a k-sparse mean vector μ to error δ, it suffices (see, e.g., [CDKGGS22, Lemma 3.2]) to: 1) obtain an estimate $\tilde{\mu}$ with $\|\tilde{\mu} - \mu\|_{2,k} \leq \delta/3$, and 2) round $\tilde{\mu}$ to the k entries with the largest magnitude, and zero out all the other entries. The main result of this paper solves the problem of robust mean estimation in the $\ell_{2,k}$ norm.

A key aspect of robust statistics is that, depending on the distribution family $\mathcal{D}$ we consider, the above problem is generally not solvable for all error parameters $\delta > 0$. This work focuses on sparse mean estimation for *heavy-tailed* distributions, where a commonly used model for heavy-tailedness is imposing only the mild assumption that the covariance of the clean distribution is bounded by the identity I, without any further tail assumptions (see Section 4.1.4 for more discussion). Even when $d = 1$ and even when there are infinitely many samples [DK19], it is known that in the heavy-tailed setting the minimum δ achievable is in the order of $\sqrt{\epsilon}$. This immediately implies the same lower bound of $\Omega(\sqrt{\epsilon})$ for the minimum achievable δ in Problem 4.1.2.

Before discussing the algorithmic results in this paper, we state known information-theoretic bounds on the sample complexity that applies to all estimators, efficient or not, for Problem 4.1.2 on distributions with covariance bounded by I, and for $\delta = \Theta(\sqrt{\epsilon})$.

Fact 4.1.3 (Information-theoretic sample complexity: computationally-inefficient)**.** *In Problem 4.1.2, for the distribution family $\mathcal{D}_2$ defined as the set of distributions with covariance $\Sigma \preceq I$, and for $\delta = \Theta(\sqrt{\epsilon})$, we have that $n(k, \epsilon, \delta, \tau) \asymp (k \log(d/k) + \log(1/\tau))/\epsilon$. That is, any estimator requires at least these many samples and there exists a (computationally-inefficient) estimator with this sample complexity. The upper bound is from [Dep20b; PSBR20] and the lower bound follows from [LM19b], even in the absence of outliers (see also Footnote 2 in [DL21]) and even when we restrict to the distribution family $\mathcal{D}_{Gaussian}$ which is the set of the Gaussian distributions with identity covariance.*

An interesting aspect of robust sparse mean estimation is that there is a conjectured statistical-computational tradeoff, namely that efficient algorithms require a qualitatively larger sample complexity than inefficient ones. Specifically, there is evidence (in the form of SQ lower bounds and reduction-based hardness) that all efficient algorithms have a quadratically worse dependence on k; that is, even for constant ϵ, δ, τ, and $\mathcal{D}_{\text{Gaussian}}$ being identity-covariance Gaussians in Problem 4.1.2, the sample complexity of all efficient algorithms is at least $\tilde{\Omega}(k^2)$, as opposed to $\tilde{O}(k)$ in Fact 4.1.3. See [DKS17; BB20] for a detailed discussion.

Both the information-theoretic bound and the conjectured information-computation tradeoff serve as benchmarks for our algorithm to match.

The main result of this paper is the following.

Theorem 4.1.4 (Computationally Efficient Heavy-Tailed Robust Sparse Mean Estimation)**.** *Let $\epsilon \in (0, \epsilon_0)$ for some sufficiently small universal constant $\epsilon_0 > 0$. Let P be a distribution over $\mathbb{R}^d$, where the mean and covariance of P are μ and Σ respectively. Suppose $\Sigma \preceq I$ and further suppose that for all $j \in [d]$, $\mathbb{E}[(X_j - \mu_j)^4] = O(1)$. Then there is an algorithm such that on input (i) the corruption parameter ϵ, (ii) the failure probability τ, (iii) the sparsity parameter k, and (iv) T, an ϵ-corrupted set of $n \gg (k^2 \log d + \log(1/\tau))/\epsilon)$ i.i.d. samples from P, the algorithm outputs $\widehat{\mu}$ satisfying $\|\widehat{\mu} - \mu\|_{2,k} = O(\sqrt{\epsilon})$ with probability $1 - \tau$ in $\mathrm{poly}(n, d)$ time.*

Phrased in a slightly different language, when our estimator is given a sufficiently large number n of ϵ-corrupted samples, it outputs an estimate $\hat{\mu}$ satisfying $\|\hat{\mu} - \mu\|_{2,k} = O\left(\sqrt{\frac{k^2 \log d}{n}} + \sqrt{\epsilon} + \sqrt{\frac{\log(1/\tau)}{n}}\right)$ with probability $1 - \tau$.

We note that the guarantees of our algorithm remain the same under a weaker assumption on Σ: we need only that $\|\Sigma\|_{\mathcal{X}_k} \leq 1$ instead of the spectral norm being bounded (the norm $\|\cdot\|_{\mathcal{X}_k}$ is formally defined in Definition 4.1.5). Informally, the $\mathcal{X}_k$ norm of a square matrix A is a convex relaxation of finding the maximum of $v^\top A v$ over k-sparse vectors v. See Theorem 4.4.2 in Section 4.4 for the stronger version of the main result, which assumes only that $\|\Sigma\|_{\mathcal{X}_k} \leq 1$.

As outlined above, the dependence of our sample complexity result on k is tight with respect to the conjectured lower bound for efficient algorithms, and its dependence on τ and ϵ are also tight with respect to the information-theoretic lower bounds, even in the Gaussian case. In terms of the smallest achievable asymptotic error (even in infinite sample regime), we show in Lemma 4.7.1 that, even after adding the mild axis-wise 4^{th} moment assumption in Theorem 4.1.4, the asymptotic error remains bounded below by $\Omega(\sqrt{\epsilon})$ when k is sufficiently large. The restriction on k is fairly mild, covering most parameter regimes of interest.

The sample complexity of our algorithm has a dependence on the failure probability that is $\log 1/\tau$, and — importantly — this is an *additive* term in the complexity instead of multiplicative. To be precise, in the i.i.d. setting with no outliers, we can artificially set $\epsilon = C \max(k^2 \log d, \log(1/\tau))/n$ for a large constant C. In this setting, when the number of samples n is such that $n \gg k^2 \log d + \log(1/\tau)$, then with probability $1 - \tau$, our algorithm outputs an estimate $\hat{\mu}$ satisfying

$$\|\hat{\mu} - \mu\|_{2,k} = O\left(\sqrt{\frac{k^2 \log d}{n}} + \sqrt{\frac{\log(1/\tau)}{n}}\right) . \tag{4.1}$$

This additive dependence is non-trivial to achieve even in the optimal rates for heavy-

tailed mean estimation in the non-robust (and non-sparse) setting. See the [LM19a] survey for a more detailed discussion. Our work provides the first *computationally efficient* estimator for heavy-tailed sparse mean estimation with such additive dependence, even in the non-robust setting.

4.1.3 Our Approach

Our algorithm fits into the stability-based filtering approach; see [DKKLMS16] and the survey [DK19]. The filtering framework is a by-now-standard algorithmic technique in robust statistics. The approach for robust mean estimation can be summarized as follows: 1) with high probability over the sampling of the n uncorrupted samples, there exists a large subset of uncorrupted samples (say, a $1 - O(\epsilon)$ fraction) satisfying a "stability" condition with respect to the mean of the uncorrupted distribution, and 2) a filtering algorithm taking as input an ϵ-*corrupted* version of the stable set of samples will remove some of the samples, such that the sample mean of the remaining points is guaranteed to be close to the true mean (which can then be returned as the final mean estimate). The notion of "stability" depends crucially on the task at hand, and is defined below for the sparse mean estimation problem.

Stability-Based Algorithms under Sparsity Informally speaking, in the context of robust mean estimation, we say that a set S is stable when the mean and the covariance of S do not deviate too much when we remove a small fraction of elements from S. For the task of *sparse* mean estimation, we would like to measure the deviation only along the k-sparse directions. However, it is computationally hard to calculate the maximum of $v^\top Av$ over k-sparse unit vectors for an arbitrary matrix A (this is known as the sparse PCA problem [TP14]). Following [BDLS17], our definition of stability involves a convex relaxation of the above optimization problem, using the following definition of the set $\mathcal{X}_k$ and the associated matrix norm $\|\cdot\|_{\mathcal{X}_k}$.

Definition 4.1.5 (The set $\mathcal{X}_k$ and the norm $\|\cdot\|_{\mathcal{X}_k}$)**.** *The set $\mathcal{X}_k$ is defined as the set of positive semidefinite matrices that have trace 1 and ℓ_1-norm at most k when flattened as a vector. The matrix norm $\|A\|_{\mathcal{X}_k}$ is then defined as $\sup_{M\in\mathcal{X}_k} |A \bullet M|$, where $A \bullet M$ denotes the trace product $\mathrm{tr}(A^\top M)$.*

Note that for any square matrix A, $\|A\|_{\mathcal{X}_k}$ is always bounded above by its spectral norm. Furthermore, observe that for any square matrix A, the maximum of $v^\top Av$ over k-sparse unit vectors is bounded above by $\|A\|_{\mathcal{X}_k}$, and the latter can be calculated efficiently using a convex program. We are now ready to define the stability condition for our sparse mean estimation task.

Definition 4.1.6 (Stability Condition for Robust Sparse Mean Estimation)**.** *For $0 < \epsilon < 1/2$ and $\epsilon \le \delta$, a set S is (ϵ, δ, k)-stable with respect to $\mu \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}_+$ if it satisfies the following condition: for all subsets $S' \subset S$ with $|S'| \ge (1-\epsilon)|S|$, the following holds: (i) $\|\mu_{S'} - \mu\|_{2,k} \le \sigma\delta$, and (ii) $\|\overline{\Sigma}_{S'} - \sigma^2 I\|_{\mathcal{X}_k} \le \sigma^2\delta^2/\epsilon$, where $\mu_{S'} = (1/|S'|)\sum_{x\in S'} x$ is the sample mean of S' and $\overline{\Sigma}_{S'} = (1/|S'|)\sum_{x\in S'}(x-\mu)(x-\mu)^\top$ is the second moment of S'.*

Definition 4.1.6 is intended for distributions with covariance matrices at most σ^2 times the identity. We will omit μ and σ above when they are clear from the context.

Focusing on the class of identity covariance Gaussian distributions, [BDLS17] gave a computationally-efficient algorithm for robust sparse mean estimation using roughly $k^2 \log d$ samples.[6] As we explain below, their algorithm succeeds under the stability condition of Definition 4.1.6.

By using the standard median-of-means pre-processing described in Section 4.2, we can reduce the robust sparse mean estimation task to the case when the corruption parameter ϵ is constant, say 0.01, and aim to achieve only a constant estimator error in

[6]The additional factor of k in their sample complexity (cf. Fact 4.1.3) is because the convex relaxation involving $\mathcal{X}_k$ norm can be loose. However, [DKS17; BB20] suggest that k^2 samples are needed for efficient algorithms.

the $\ell_{2,k}$ norm. For this regime, we state the guarantees of robust sparse mean estimation algorithm of [BDLS17] (developed for the Gaussian setting) as follows[7]:

Fact 4.1.7. *Let S be a set in $\mathbb{R}^d$ such that there exists a set $S' \subseteq S$ such that (i) $|S'| \geq 0.99|S|$, and (ii) S' is an $(0.01, O(1), k)$-stable with respect to (unknown) μ and (unknown) σ. There is a $\mathrm{poly}(|S|, d)$-time algorithm that takes as input T, an 0.01-corruption of S, and returns a mean estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_{2,k} \leq O(\sigma)$.*

Given this prior algorithmic result, the key challenge is to show that, even in the setting of *heavy-tailed* data, a large subset of the uncorrupted samples satisfies the stability condition with high probability. Without sparsity constraints, [DKP20; HLZ20] showed that $O(d)$ samples suffice for stability (under a different definition appropriate for the dense setting), which is too large for our purposes.

Truncation is Necessary for Stability Recall that our goal is to show that if we draw roughly $k^2 \log d$ samples from a heavy-tailed distribution, then it contains a large stable subset. For the light-tailed data (Gaussian), this was shown in [BDLS17]. However, this desired statement is not true for general heavy-tailed distributions. Consider the standard setting for modeling heavy-tailed data, namely that the covariance Σ of the uncorrupted distribution is upper bounded by the identity. For simplicity, also assume that the sparsity parameter k, corruption parameter ϵ and failure probability τ are all constants. Thus, our goal is to show that, with high probability, there is a large stable subset among $\log d$ samples. Yet, as we show in Example 4.3.1 in Section 4.3, there exists a distribution where deterministically for *any* set of up to $o(d)$ many uncorrupted samples, *no* large subset can be stable. This distribution is the one returning a vector of length $\sqrt{d}$ from a randomly chosen axis direction, which has unit covariance. Essentially,

[7] See also [ZJS22b] for a related algorithm.

the long length of $\sqrt{d}$ along directions as sparse as the axis directions causes stability to fail to hold.

In order to circumvent this obstacle, we propose to "truncate" all the samples in ℓ_∞ norm before using a stability-based filtering robust mean estimation algorithm. Specifically, we start by computing an initial mean estimate, and then clip each sample coordinate-wise to within a radius of $\Theta(\sqrt{k})$ of the initial mean estimate. This radius is chosen carefully to ensure that the mean of the original distribution and the clipped distribution is close in $\ell_{2,k}$-norm. Ensuring that the clipped distribution also has small variance turns out to be non-trivial, as we detail below.

Necessity of Bounded Higher Moments After truncation, we have the guarantee that no point is too far from the true mean. Unfortunately, truncation can potentially also *rotate* a point about the true mean, in the sense that for a sample, the direction of its difference from the true mean may change after such truncation. In general, this rotation effect can cause much of the mass of the distribution to rotate and concentrate towards certain directions, and significantly *increase* the variance in those directions. (See Appendix B.2.1 for more details.) In this work, we identify the mild condition that the 4^{th} moment is bounded along each *axis* direction by some constant, on top of our assumption that $\Sigma \preceq I$, to be sufficient to show that truncation can only increase variance in directions that are non-sparse — in the sense that the resulting covariance will still have bounded $\mathcal{X}_k$ norm (see Lemma 4.3.2). Thus, under these mild conditions, we can safely truncate our samples (which is necessary for stability to hold, as outlined above), and modify our goal to show this truncated distribution contains a large stable set with high probability.

Stability of Truncated Samples with High Probability Even after truncation and after imposing an axis-wise 4^{th} moment bound, it remains challenging to show that,

with high probability, there exists a large subset of samples that are stable with respect to the true mean.

As we see in Section 4.5, the analysis reduces to showing that with high probability over the uncorrupted samples, for every matrix $M \in \mathcal{X}_k$, there exists a large subset of samples S whose empirical covariance $\overline{\Sigma}_S$ has a small inner product with M, namely that $M \bullet \overline{\Sigma}_S$ is bounded. In the non-sparse setting, the strategy used in [DL22a] and [DKP20] is to first show a high probability event for all $M = vv^\top$ for unit vectors v, and then to show that the event for all $M = vv^\top$ deterministically implies that the event holds also for all $M \succeq 0$ with $\mathrm{tr}(M) = 1$. This strategy is important because although the cover of PSD matrices would roughly be exponential in d^2, the cover of $vv^\top$ is only exponential in d. Thus, the first step holds with roughly d samples, and the second step crucially uses the spectral decomposition (SVD) of positive semidefinite (PSD) matrices. On the other hand, in our sparse setting, if we applied the usual SVD to the PSD matrices $M \in \mathcal{X}_k$, the resulting decomposition will generally not yield sparse components, and thus not allowing us to leverage sparsity. Instead, inspired by certain matrix norm results derived by Li [Li18], we carefully design a (non-spectral) decomposition that does yield k^2-sparse components and can be covered with $k^2 \log d$ samples; as well as a more delicate argument to complete the second step, namely that the event holding for all components M in the decomposition implies the event holding for all $M \in \mathcal{X}_k$. The intricacies of these arguments also allow us to get a sample complexity that ultimately yields an *additive* (as opposed to multiplicative) dependence on $\log 1/\tau$, which as described in the previous section is a crucial feature of our result, and in line with the non-robust non-sparse sub-Gaussian mean estimation setting.

4.1.4 Related Work

Algorithmic Robust Statistics The goal of algorithmic robust statistics is to obtain dimension-independent asymptotic error even in the presence of constant fraction of outliers in high dimensions in a computationally efficient way. Since the dissemination of [DKKLMS16; LRV16], which focused on high-dimensional robust mean estimation, the body of work in the field has grown rapidly. For example, prior work has obtained dimension-independent guarantees for various problems such as linear regression [KKM18; DKS19] and convex optimization [PSBR20; DKKLSS19]. See the survey [DK19] for a more detailed description. Most relevant to us are the works on robust mean estimation that leverage the sparsity constraints and obtain improved sample complexity. The algorithms developed in [BDLS17; DKKPS19; CDKGGS22; DKKPP22b] obtain optimal asymptotic error for light-tailed distributions, such as Gaussians. However, these algorithms (and their analyses) crucially rely on the light-tails and, as outlined in Section 4.1.3, provably do not work for heavy-tailed distributions.

Heavy-Tailed Statistical Estimation The recent decades also saw a growing interest in studying statistics in heavy-tailed settings. Even for the basic question of univariate mean estimation without sample corruption, the statistical limits are only recently resolved by a line of work started by Catoni [Cat12] and ending with Lee and Valiant [LV20] (see also [Min22] for an alternative estimator).

The high-dimensional heavy-tailed setting turns out to be much more challenging and has been extensively studied in recent years, e.g., for mean estimation in the ℓ_2 norm [LM19d] and in other norms [LM19b; DL21], covariance estimation [MZ20], and stochastic convex optimization [BM22]. In absence of contamination, the goal is to obtain sample complexity as if the distribution were Gaussian. Roughly speaking, this corresponds to an additive dependence on the logarithm of failure probability in various estimation tasks (as we achieve also in this work). We refer the reader to the

survey for more details [LM19a]. This line of work focuses on the statistical limits, and the estimators developed are generally computationally inefficient.

A closely-related body of research aims to obtain *efficient* algorithms for heavy-tailed distributions with optimal statistical performance, ideally matching the above guarantees. These works include high-dimensional (dense) mean estimation [Hop20; CFB19; DL22a; LLVZ20; DKP20; HLZ20; CTBJ22; LV20], linear regression [CHKRT20; PJL20b; Dep20a], and covariance estimation [CHKRT20]. We note that many of these works are inspired by the algorithmic robust statistics literature and can also tolerate a constant fraction of contaminated data.

To the best of our knowledge, none of these works studies sparse estimation under heavy-tailed distributions (even in absence of outliers), and our work is the first result with sample complexity that is additive in the logarithm of the failure probability.

4.1.5 Organization

The structure of this paper is as follows: After the necessary technical preliminaries in Section 4.2, in Section 4.3 we describe and analyze our simple pre-processing truncation scheme. In Section 4.4, we provide a detailed description of our algorithm and an outline of its analysis, assuming the necessary stability conditions are satisfied. Sections 4.5 and 4.6 establish that the stability condition will be satisfied with the appropriate sample complexity, and are the main technical contributions of this work. Finally, Section 4.7 shows that the error guarantee of our algorithm is information-theoretically optimal under a mild assumption on the sparsity. For the sake of the presentation, some technical proofs have been deferred to an appendix.

4.2 Preliminaries

Notations Here we define the notations we use in the rest of the paper. For a (multi-)set $S \subset \mathbb{R}^d$, we denote $\mu_S = (1/|S|)\sum_{x\in S} x$ and $\Sigma_S = (1/|S|)(\sum_{x\in S}(x-\mu_S)(x-\mu_S)^\top)$. When the vector μ notation is clear from context, we use $\overline{\Sigma}_S$ to denote $(1/|S|)\sum_{x\in S}(x-\mu)(x-\mu)^\top$.

Let $\mathcal{U}_k$ denote the set of k-sparse unit vectors in $\mathbb{R}^d$. For two vectors x and y, $\langle x, y\rangle$ denotes the dot product $x^\top y$. For a vector $x \in \mathbb{R}^d$, we use $\|x\|_{2,k} := \sup_{v\in\mathcal{U}_k}\langle x, v\rangle$ and $\|x\|_\infty$ to denote $\max_j |x_j|$. For a matrix M, we use $\|M\|_1$ to denote $\sum_{i,j}|M_{i,j}|$ and $\|M\|_0$ to denote the number of non-zero entries of M. For two matrices A and B, we use $A \bullet B$ to denote the trace inner product $\mathrm{tr}(A^\top B)$. Define $\mathcal{X}_k := \{M : M \succcurlyeq 0, \mathrm{tr}(M) = 1, \|M\|_1 \le k\}$. For a matrix A, we define $\|A\|_{\mathcal{X}_k} := \sup_{M\in\mathcal{X}_k}|A \bullet M|$. For an $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, \ldots, n\}$. For a set $S \subseteq \mathbb{R}^d$ and a function f, we also define the set function notation $f(S)$ as $\{f(x) \mid x \in S\}$.

Coordinate-wise Median-of-Means We use the coordinate-wise median-of-means algorithm to robustly obtain a preliminary mean estimate, with guarantees captured by the following fact.

Fact 4.2.1. *The coordinate-wise median-of-means algorithm satisfies the following guarantee: given the corruption parameter ϵ, failure probability τ, and a set T of n many ϵ-corrupted samples from a distribution D with mean μ and axis-wise variance $\mathbb{E}_{X\sim D}[(X_j-\mu_j)^2] \le \sigma^2$ for all $j \in [d]$, then with probability at least $1-\tau$ over the sample set T, the output of the algorithm $\hat{\mu}$ is such that $\|\hat{\mu}-\mu\|_\infty \le \sigma O(\sqrt{\epsilon} + \sqrt{(\log(d/\tau))/n})$.*

Median-of-Means Pre-Processing Another standard technique we use in this paper is the median-of-means pre-processing, which is a distinct technique from the coordinate-wise median-of-means algorithm mentioned right above. Recall that in Theorem 4.1.4,

the asymptotic error term is $\sqrt{\epsilon}$, which tends to 0 as the corruption parameter $\epsilon \to 0$. The following pre-processing step allows us to reduce the problem from the $\epsilon \to 0$ case to a constant ϵ case: Let T be the input ϵ-corrupted set of samples. Split the samples T randomly into g equally-sized groups of size $m = n/g$ where $g = 100\epsilon n$, and replace each group by the sample mean of the group. Let T_{grouped} be this new set of points. It is easy to check that at most 0.01-fraction of T_{grouped} can be corrupted by outliers. The effects of this pre-processing is captured by the following Fact 4.2.2, which we prove for completeness in Appendix B.1.2.

Fact 4.2.2 (Median-of-Means Pre-Processing)**.** *Suppose there is an efficient algorithm such that, on input $\sigma \in \mathbb{R}_+$ and a 0.01-corrupted set of $n \gg k^2 \log d + \log(1/\tau)$ samples from a distribution D with mean μ and covariance Σ with $\|\Sigma\|_{\mathcal{X}_k} \le \sigma^2$ and $\mathbb{E}_{X\sim D}[(X_j - \mu_j)^4] = O(\sigma^4)$ for each coordinate $j \in [d]$, returns $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_{2,k} \le O(\sigma)$ with probability at least $1 - \tau$.*

Then, there is an efficient algorithm such that, on input $\epsilon \in (0, 0.01]$ and an ϵ-corrupted set of $n \gg (k^2 \log d + \log(1/\tau))/\epsilon$ samples from a distribution with mean μ and covariance Σ, satisfying (i) $\|\Sigma\|_{\mathcal{X}_k} \le 1$ and (ii) $\mathbb{E}_{X\sim D}[(X_j - \mu_j)^4] = O(1)$ for every coordinate $j \in [d]$, returns a mean estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_{2,k} \le O(\sqrt{\epsilon})$ with probability at least $1 - \tau$.

4.3 Truncation Pre-Processing

The general approach of using a stability-based filtering algorithm for robust mean estimation is to show that, given sufficiently many samples, there exists a large (say $1 - O(\epsilon)$ fraction) subset of the samples that are stable with respect to the true mean μ. However, the following simple example shows that it is not possible for i.i.d. samples drawn from a heavy-tailed distribution to satisfy the sparse stability condition using sample size of $\text{poly}(\log d)$.

Example 4.3.1. *For any number of moments $t \geq 2$, there is a distribution X satisfying the following conditions: (i) The mean of X is 0, and for every unit vector v, the t^{th} moment in direction v is upper bounded by 1, that is, $\mathbb{E}[|\langle v, x\rangle|^t] \leq 1$ for $t \geq 2$, (ii) If S is an arbitrary set of $n \leq o(d^{2/t})$ points from the support of X, then the set S cannot be $(\epsilon, O(\sqrt{\epsilon}), k)$-stable, for any $\epsilon > 0$, with respect to the mean of the distribution. As a corollary, no subset of S can be stable either.*

Proof. For $j \in [d]$, let e_j be the vector that is 1 on the j-th coordinate and 0 otherwise. For a fixed r, consider the distribution P, supported uniformly on the $2d$ points $S = \{\pm re_1, \pm re_2, \ldots, \pm re_d\}$.

It follows that P is a zero mean distribution. The covariance of the distribution P is $\sum_j (1/d) r^2 e_i e_i^\top = (r^2/d) I$. Furthermore, for any unit vector v and $t \geq 2$, we have that the t-th moment in the direction v is bounded as follows:

$$\mathbb{E}[|v \cdot X|^t] = \sum_{j=1}^{d} \frac{1}{d} |v_j|^t r^t = \frac{r^t}{d} \|v\|_t^t \leq \frac{r^t}{d} \|v\|_2^t \leq \frac{r^t}{d} \ ,$$

where we use that $t \geq 2$ and $\|v\|_t \leq \|v\|_2$ for any vector v. Thus, we choose $r = d^{1/t}$ for the distribution.

Now we show the second claim, that *any* set of at most $\Omega(d^{2/t})$ samples from this distribution cannot be stable. Let S be any (multi-)set of n points from the support of X. Let $x_1 \in S$. Since x_1 is 1-sparse and has ℓ_2 norm r, we have that $x_1 x_1^\top / r^2$ belongs to $\mathcal{X}_k$. Thus, we have the following:

$$\left\| \frac{1}{n} \sum_{i \in S'} x_i x_i^\top \right\|_{\mathcal{X}_k} \geq \left\langle \frac{1}{n} \sum_{i \in S'} x_i x_i^\top, \frac{1}{r^2} x_1 x_1^\top \right\rangle \geq \frac{\|x_1\|^4}{r^2 n} = \frac{r^2}{n} \ .$$

Therefore, for r^2/n to be upper bounded by a constant, n has to be $\Omega(d^{2/t})$. □

Consequently, if we want to perform robust sparse mean estimation using

$\mathrm{poly}(k, \log d)$ samples, we need to modify the algorithm. Our approach is to perform an initial truncation of the samples before using a stability-based robust mean estimator. A balance needs to be struck in order to truncate sufficiently for stability to hold (with high probability over the samples), but also to truncate mildly enough such that the mean (and covariance) of the truncated distribution does not shift too much.

For a scalar $a \in \mathbb{R}_+$ and a vector $b \in \mathbb{R}^d$, let $h_{a,b} : \mathbb{R}^d \to \mathbb{R}^d$ be the following thresholding function:

$$\forall i \in [d], \quad h_{a,b}(x)_i = \begin{cases} x_i, & \text{if } |x_i - b_i| \le a \\ b_i + a & \text{if } x_i - b_i \ge a \\ b_i - a & \text{if } x_i - b_i \le -a \end{cases} . \tag{4.2}$$

Note that $h_{a,b}(x)$ projects the point x to the ℓ_∞ ball of radius a around b.

As explained in the Introduction, truncation in general rotates a point about the true mean, and thus can in fact cause the covariance of the distribution to grow in certain directions. The following lemma captures the fact that, if we make the further mild assumption that the distribution has bounded 4^{th} moment along all the axis directions, then we will at least be able to preserve the $\mathcal{X}_k$ norm of the covariance matrix. The proof of Lemma 4.3.2 is in Appendix B.2.2.

Lemma 4.3.2 (Truncation in ℓ_∞)**.** *Let P be a distribution over $\mathbb{R}^d$ with mean μ_P and covariance Σ_P, with $\|\Sigma\|_{\mathcal{X}_k} \le \sigma^2$ for some $\sigma^2 > 0$. Let $X \sim P$ and assume that for all $j \in [d]$, $\mathbb{E}[(X - \mu_P)_j^4] \le \sigma^4\nu^4$ for some $\nu \ge 1$. Let $b \in \mathbb{R}^d$ be such that $\|b - \mu\|_\infty \le a/2$ and $a := 2\sigma\sqrt{k/\epsilon}$ for some $\epsilon \in (0, 1)$. Define Q to be the distribution of $Y := h_{a,b}(X)$. Let the mean and covariance of Q be μ_Q and Σ_Q respectively. Then the following hold:*

(1) $\|\mu_P - \mu_Q\|_\infty \le \sigma\sqrt{\epsilon/k}$

(2) $\|\mu_P - \mu_Q\|_{2,k} \le \sigma\sqrt{\epsilon}$

(3) $\|\Sigma_P - \Sigma_Q\|_{\mathcal{X}_k} \leq 3\sigma^2\epsilon\nu^4$

(4) *For all* $i \in [d]$, $\mathbb{E}[(Y - \mu_Q)_i^4] \leq 8\nu^4\sigma^4$

(5) $\|Y - \mu_Q\|_\infty \leq 2a = 4\sigma\sqrt{k/\epsilon}$ *almost surely.*

In Lemma 4.3.2 above, b represents the initial mean estimate, and $\tilde{\mu}$ will be obtained by Fact 4.2.1.

4.4 Algorithm and Analysis

The high-level algorithm we propose is stated as follows.

Algorithm 1 Robust Sparse Mean Estimation with High Probability

1. Input: An ϵ-corrupted sample set $T \subseteq \mathbb{R}^d$ of size n
2. Median-of-Means pre-processing: Group points in T into g groups, each of size $m = n/g$, where $g = 100\epsilon n$, and take the sample mean of a group to be a new point. Call these new points T_{grouped}.
3. Define $\sigma = 1/\sqrt{m}$.
4. Compute coordinate-wise median-of-means estimate $\tilde{\mu}$ from Fact 4.2.1 with corruption parameter 0.01 and failure probability $\tau/2$, using the set of points T_{grouped}.
5. Truncate all points in T_{grouped} to within $B_\infty(\tilde{\mu}, 4\sigma\sqrt{k})$, namely, given a point x, we replace it with the point $h_{4\sigma\sqrt{k},\tilde{\mu}}(x)$, where $h_{a,b}$ is defined in Equation (4.2).
6. Run the stability-based robust *sparse* mean estimator from Fact 4.1.7 on the truncated samples, i.e. $\left\{h_{4\sigma\sqrt{k},\tilde{\mu}}(x) \mid x \in T_{\text{grouped}}\right\}$.

We note that this algorithm is shift and scale invariant, based on the same invariance of the median-of-means pre-processing as well as the invariance of the robust sparse mean estimator from Fact 4.1.7.

We will now prove that Algorithm 1 satisfies the guarantees of Theorem 4.1.4 and its stronger version, Theorem 4.4.2 stated below. Our analysis crucially uses Theorem 4.4.1,

which states that, with high probability, there exists a set consisting of most of the truncated samples that is stable with respect to some point close to the true mean in ℓ_∞ norm. This is the main structural result of our paper.

Theorem 4.4.1. *Let S be a set of n i.i.d. data points from a distribution P over $\mathbb{R}^d$, and let T be a 0.01-corruption of S. Let $\tilde{\mu}$ be the coordinate-wise median-of-means estimate computed from set T. Let the mean of P be μ and covariance Σ such that $\|\Sigma\|_{\mathcal{X}_k} \leq \sigma^2$, and for all $i \in [d]$, $\mathbb{E}[X_i^4] \leq O(\sigma^4)$. Suppose that $n = \Omega(k^2 \log d + \log(1/\tau))$. Let $a = \sigma\sqrt{k}$. With probability $1 - \tau$ over S, for all T we have that there exists a subset $S' \subseteq T$ with $|S'| \geq 0.95n$ such that $h_{a,\tilde{\mu}}(S')$ is $(0.01, O(1), k)$-stable with respect to some μ' and σ with $\|\mu' - \mu\|_\infty \leq O(\sigma/\sqrt{k})$.*

We show Theorem 4.4.1 in two steps. First, we show a simpler, analogous stability result assuming that we truncate with respect to the true mean vector μ instead of the coordinate-wise median-of-means estimate $\tilde{\mu}$ as in Algorithm 1 and Theorem 4.4.1. This is stated as Theorem 4.5.1 and proved in Section 4.5. Then, in Section 4.6, we show a "Lipschitzness" argument that lets us conclude Theorem 4.4.1. The final proof of Theorem 4.4.1 is given in Section 4.6.3.

Theorem 4.4.2 (Main Result, Strong Version)**.** *Let $\epsilon \in (0, \epsilon_0)$ for small constant $\epsilon_0 > 0$. Let P be a multivariate distribution over $\mathbb{R}^d$, where the mean and covariance of P are μ and Σ respectively. Suppose $\|\Sigma\|_{\mathcal{X}_k} \leq 1$ and further suppose that for all $j \in [d]$, $\mathbb{E}[(X_j - \mu_j)^4] = O(1)$. Then, there is an algorithm such that, on input (i) the corruption parameter ϵ, (ii) the failure probability τ, (iii) the sparsity parameter k, and (iv) T, an ϵ-corrupted set of $n \gg (k^2 \log d + \log(1/\tau))/\epsilon)$ i.i.d. samples from P, it outputs $\widehat{\mu}$ satisfying $\|\widehat{\mu} - \mu\|_{2,k} = O(\sqrt{\epsilon})$ with probability $1 - \tau$ in $\mathrm{poly}(n, d)$ time.*

We note that, since the $\mathcal{X}_k$ norm of a covariance matrix is upper bounded by its maximum eigenvalue, Theorem 4.1.4 is an immediate corollary of Theorem 4.4.2.

Proof of Theorem 4.4.2. Step 2 of Algorithm 1, the median-of-means pre-processing, is exactly the same as the reduction in Fact 4.2.2. Thus, by Fact 4.2.2, it suffices to show that, for every $\sigma > 0$, Steps 4–6 in Algorithm 1 yield an $O(\sigma)$ estimation error in $\ell_{2,k}$ norm when given 0.01-corrupted samples from a distribution D with covariance bounded by σ^2 in $\mathcal{X}_k$-norm and axis-wise 4th moment bounded by $O(\sigma^4)$.

Theorem 4.4.1 states that, with probability at least $1-\tau$, the samples after the processing of Step 5 are such that there exists a 95% of the samples that form a $(0.1, O(1), k)$-stable subset with respect to some vector μ' and σ with $\|\mu' - \mu\|_\infty \leq O(\sigma/\sqrt{k})$. Fact 4.1.7 then guarantees that, on input such a set of samples, the routine we invoke in Step 6 of Algorithm 1 will return a mean estimate $\hat{\mu}$ such that $\|\hat{\mu} - \mu'\|_{2,k} \leq O(\sigma)$. Further, since $\|\mu' - \mu\|_\infty \leq O(\sigma/\sqrt{k})$, we have that $\|\mu' - \mu\|_{2,k} \leq O(\sigma)$, and therefore we can conclude via the triangle inequality that the mean estimate $\hat{\mu}$ satisfies $\|\hat{\mu} - \mu\|_{2,k} \leq O(\sigma)$. □

4.5 Stability After Removing Points: Additive dependence on $\log(1/\tau)$

In this section, we give the core part of the argument (Theorem 4.5.1) for the main stability result (Theorem 4.4.1) in this paper. Recall, via the median-of-means pre-processing, that we only need to consider the constant contamination case ($\epsilon = \Theta(1)$). Thus, the goal is to show (Theorem 4.4.1) that with high probability, after truncation according to the coordinate-wise median-of-means preliminary estimate, there exists a large subset of uncontaminated samples that is $(\Theta(1), O(1), k)$-stable with respect to (a vector close in $\ell_{2,k}$ norm of) the *true mean* of the distribution as well as σ where $\sigma^2 = \|\Sigma\|_{\mathcal{X}_k}$.

The key difference between Theorems 4.5.1 and 4.4.1 is that the former is a stability result that applies only to uncontaminated i.i.d. samples truncated according to some

fixed vector close to the true mean. On the other hand, the final stability result we require concerns samples truncated according to the coordinate-wise median-of-means estimate, which itself depends on the samples and is not fixed. Section 4.6 shows the delicate argument going from Theorem 4.5.1 to Theorem 4.4.1.

Theorem 4.5.1. *Let S be a set of n i.i.d. data points from a distribution P over $\mathbb{R}^d$. Let the mean of P be μ and covariance Σ such that $\|\Sigma\|_{\mathcal{X}_k} \leq \sigma^2$, and for all $j \in [d]$, $\mathbb{E}[(X_j - \mu)^4] \leq \nu^4$. Suppose P is supported over the set $\{x : \|x-\mu\|_\infty \leq \sigma \times r \times \sqrt{k}\}$. If $n = \Omega(k^2 \log d + \log(1/\tau))$, then with probability $1-\tau$ there exists a set $S' \subseteq S$ such that:*

1. $|S'| \geq 0.98n$
2. *S' is $(0.01, \delta, k)$-stable with respect to μ and σ where $\delta = O(\max(1, r^2, \nu^2/\sigma^2))$.*

Before giving the proof, we point out that the specific application of Theorem 4.5.1 will be on the samples $h_{a,\mu}(x_i)$, where x_i are the original uncontaminated i.i.d. samples, μ is the underlying mean vector we are trying to estimate, and for a appropriately chosen to match Algorithm 1.

Proof. In the following proof, we will use notations q, s_1, s_2, s_3, V_Z and B, all of which are either constants or functions of σ, r and ν in the theorem statement. The functions are explicitly chosen in Appendix B.3.

We will assume $\mu = 0$ without loss of generality. Instead of directly showing the existence of subset $S' \subseteq S$ (with high probability over the samples S) that is stable, Proposition B.1.1 in Appendix B.1 lets us show the following simpler condition: let $\Delta_{n,\epsilon}$ be the set of weights/distributions w such that $w_i \leq 1/(1-\epsilon)$, then there exists a weighting $w \in \Delta_{n,0.01}$ such that $\|\Sigma_w\|_{\mathcal{X}_k} \leq B$ for the function B chosen in Appendix B.3, which satisfies $B = O(\sigma^2 \max(1, r^2, \nu^2/\sigma^2))$. That is, for the following proof, we just need to prove that $\min_{w \in \Delta_{n,0.01}} \|\Sigma_w\|_{\mathcal{X}_k} \leq B$.

We proceed as follows:

$$\min_{w\in\Delta_{n,0.01}} \|\Sigma_w\|_{\mathcal{X}_k} = \min_{w\in\Delta_{n,0.01}} \max_{M\in\mathcal{X}_k} \langle M, \Sigma_w\rangle = \max_{M\in\mathcal{X}_k} \min_{w_M\in\Delta_{n,0.01}} \langle M, \Sigma_w\rangle \ ,$$

where the last equality is a straightforward application of the minimax theorem for a minimax optimization problem with independent convex domains and a bilinear objective. It thus suffices to show the following: with probability $1-\tau$,

$$\forall M \in \mathcal{X}_k : \ |\{x \in S : x_i^\top M x_i > B\}| \le 0.01|S| \tag{4.3}$$

from which we can construct the weighting w_M as uniform distribution over the elements outside the above set.

Define the following sets of sparse matrices:

$$\begin{aligned}\mathcal{A}_k &:= \left\{A \in \mathbb{R}^{d\times d} : \|A\|_0 \le k^2, \|A\|_F \le 1\right\} \\ \mathcal{A}_{k,P} &:= \left\{A \in \mathcal{A}_k : \mathbb{P}\left\{x^\top A x \ge s_1\right\} \le q\right\} \ ,\end{aligned} \tag{4.4}$$

where q and s_1 are chosen in Appendix B.3. If $n \gtrsim (k^2 \log d + \log(1/\tau))/(q^2)$, then a standard covering/VC-dimension bound (see Lemma B.2.2 for details) implies that the following event holds with probability $1-\tau$:

$$\forall A \in \mathcal{A}_{k,P} : \ |\{x \in S : x_i^\top A x_i > s_1\}| \le 2 \times q \cdot |S| \ . \tag{4.5}$$

Our choice of q is a constant (cf. Appendix B.3) and thus the required sample complexity for Equation (4.5) to hold is $\Omega(k^2 \log d + \log(1/\tau))$. We will now show that the event in Equation (4.5) implies that the event in Equation (4.3) holds.

Suppose, for the sake of contradiction, that the event in Equation (4.3) does not hold. Then there exists an $M \in \mathcal{X}_k$ such that $|\{x \in S : x_i^\top M x_i > B\}| > 0.01|S|$. We will show

the existence of a matrix Q violating event Equation (4.5), via the probabilistic method, to reach the desired contradiction.

Fixing an M that violates Equation (4.3), consider the random matrix Q where each entry $Q_{i,j}$ is sampled independently from the following distribution, defined using the constant s_2 chosen in Appendix B.3:

$$Q_{i,j} := \begin{cases} M_{i,j}, & \text{with prob. 1 if } |M_{i,j}| \geq s_2/k, \\ \frac{s_2}{k}\text{sign}(M_{i,j}), & \text{with prob. } |kM_{i,j}|/s_2 \text{ if } |M_{i,j}| \leq s_2/k, \\ 0, & \text{with remaining prob. if } |M_{i,j}| \leq s_2/k \end{cases} . \tag{4.6}$$

Defining $p_{i,j}$ to be $\min(1, k|M_{i,j}|/s_2)$, then $Q_{i,j}$ is equivalently $M_{i,j}/p_{i,j}$ with probability $p_{i,j}$ and 0 otherwise.

We will show that the following events hold simultaneously with non-zero probability, leading to a contradiction to event Equation (4.5):

(I) $Q \in s_3\mathcal{A}_{k,P}$,

(II) $|\{x \in S : x_i^\top Q x_i > s_3 \times s_1\}| > 2 \times q \cdot |S|$,

where s_3 is also a constant, larger than 2, and explicitly chosen in Appendix B.3. Using different techniques, we will show that the first condition holds with probability at least $1 - 2 \times 10^{-6}$ and the second condition holds with probability at least 4×10^{-6}, thus implying that the events hold simultaneously with non-zero probability.

Condition (I) Showing that Q belongs to $s_3\mathcal{A}_k$ with high constant probability is straightforward: by the construction of Q, it has small expected sparsity as well as small expected Frobenius norm. An application of Markov's inequality shows that $Q \in s_3\mathcal{A}_k$ with high constant probability (Lemma 4.5.2).

The trickier part is to show that Q is also in $s_3\mathcal{A}_{k,P}$, namely that $\Pr_{x\sim P}(x^\top Qx > s_3 \times s_1)$ is upper bounded by the small constant. We consider the distribution of $x^\top Qx$ over the probability of independently drawing $x \sim P$ and a random Q, and show that $x^\top Qx$ is small with high probability over this joint distribution (Lemma 4.5.4), which requires using the axis-wise 4^{th} moment bounds on P as well as the fact that $M \in \mathcal{X}_k$. Lemma 4.5.3 implies that with high probability, we will draw a Q satisfying $\Pr_{x\sim P}(x^\top Qx > s_3 \times s_1)$ being bounded by a small constant.

We now show Condition (I) as sketched above, beginning with the following lemma showing that Q lies in $s_3\mathcal{A}_k$ with high probability.

Lemma 4.5.2 (Q lies in $s_3\mathcal{A}_k$ with high probability)**.** *Let Q be generated as described in Equation (4.6), for an $M \in \mathcal{X}_k$. Then with probability except $(1/s_2) + (s_2/s_3^2)$, we have that $Q \in s_3\mathcal{A}_k$.*

Proof. The expected sparsity of Q is at most $\sum_{i,j} \frac{k}{s_2}|M_{i,j}| \le \frac{k^2}{s_2}$ since $|M|_1 \le k$. Thus, by Markov's inequality, except with $1/s_2$ probability, Q and hence Q/s_3 is k^2-sparse. We also have to show that with probability at least $1 - 10^8$, $\|Q\|_F \le s_3$.

$$\mathbb{E}\,\|Q\|_F^2 \le \sum_{i,j}\left(\frac{s_2}{k}\right)^2\left(\frac{k|M_{i,j}|}{s_2}\right) = \frac{s_2|M_{i,j}|}{k} = s_2. \tag{4.7}$$

Again, by Markov's inequality, we get that with probability except s_2/s_3^2, the Frobenius norm of Q is at most s_3. The lemma statement follows from the union bound. □

Our choice of constants in Appendix B.3 would ensure that the failure probability in Lemma 4.5.2 is at most 10^{-6}. That is,

$$\frac{1}{s_2} + \frac{s_2}{s_3^2} \le 10^{-6}. \tag{4.8}$$

It remains to show that Q belongs to $s_3\mathcal{A}_{k,P}$ with high (constant) probability, i.e., with probability 10^{-6} over sampling of Q, we have that $\mathbb{P}_{x\sim P}(x^\top Qx > s_3 \times s_1|Q) \leq q$. Let $R := x^\top Qx$, where both x and Q are sampled independently from P and Equation (4.6) respectively.

To show this, we use the following the lemma for a sufficient condition involving sampling both x and Q.

Lemma 4.5.3. *Consider a probability space over the randomness of independent variables X and Y. Suppose the event E (over pairs (X, Y)) happens with probability at least $1 - \alpha\beta$ for some $\alpha, \beta \in [0, 1]$. Then, it must be the case that, with probability at least $1 - \alpha$ over the sampling of X, the conditional probability of E given X is at least $1 - \beta$.*

Proof. For the sake of contradiction, suppose the lemma conclusion is false. Then

$$\Pr_{X,Y}(E) = \int \Pr_{Y}(E|X)\,\mathrm{d}\Pr(X) < (1-\alpha) + \alpha(1-\beta) = 1 - \alpha\beta,$$

which contradicts the premise. □

To conclude that $Q \in s_3\mathcal{A}_{k,P}$ with high probability, it thus suffices to show that with probability $1 - 10^{-6} \times q$ over both x and Q, $R \leq s_3 \times s_1$.

Lemma 4.5.4. *Let $R = x^\top Qx$, where Q is independently drawn from the distribution in Equation (4.6) and x is drawn independently from P. Under the assumptions of Theorem 4.5.1,*

$$\mathbb{P}\{R > s_3 \times s_1\} \leq \frac{\sigma^2}{s_1} + \frac{4}{s_3} + \frac{s_2 \times \nu^4}{s_3 \times s_1^2}\,. \tag{4.9}$$

Proof. We consider three exhaustive events, over x and Q, of $\mathcal{E} := \{R > s_3 \times s_1\}$, and bound the probability of each of them:

1. $\mathcal{E}_1 := \{(x, Q) : \mathbb{E}[R|x] > s_1\}$. Since $\mathbb{E}[R|x] = x^\top M x$, the event corresponds to $\{x : x^\top M x > s_1\}$. We have that $\mathbb{E}[x^\top M x] = \langle \Sigma, M \rangle \leq \|\Sigma\|_{\mathcal{X}_k} = \sigma^2$. By Markov's inequality, $\mathbb{P}(\mathcal{E} \cap \mathcal{E}_1) \leq \mathbb{P}(\mathcal{E}_1) \leq \sigma^2/(s_1)$.

2. $\mathcal{E}_2 := \{(x, Q) : x \in \mathcal{F}\}$, where $\mathcal{F}$ is the following event over x: $\mathcal{F} = \{x : \mathbb{E}[R|x] \leq s_1, \mathbf{Var}(R|x) \leq s_3 \times s_1^2\}$. Observe that conditioned on $x \in \mathcal{F}$, we have that $R|x$ is a random variable with mean at most s_1 and variance at most $s_3 \times s_1^2$. Thus for each such $x \in \mathcal{F}$, the conditional probability that $R > s_3 \times s_1$ is at most $s_3 s_1^2/((s_3 - 1)^2 s_1^2)$ by Chebyshev's inequality. We thus get that $\mathbb{P}(\mathcal{E}_2 \cap \mathcal{E}) \leq \mathbb{P}(\mathcal{E}|\mathcal{E}_2) = \mathbb{P}(\mathcal{E}|x \in \mathcal{F}) \leq 4/(s_3)$, where we use that $s_3 \geq 2$.

3. $\mathcal{E}_3 := \{(x, Q) : \mathbf{Var}(R|x) \geq s_3 \times s_1^2\}$. We will upper bound $\mathbb{P}(\mathcal{E}_3)$. We first calculate the $\mathbf{Var}(R|x)$ using the independence of entries of Q as follows:

$$\mathbf{Var}(R|x) = \sum_{i,j} x_i^2 x_j^2 \, \mathbf{Var}(Q_{i,j}) = \sum_{i,j:|M_{i,j}| \leq s_2/k} x_i^2 x_j^2 |M_{i,j}| \left(\frac{s_2}{k} - |M_i, j| \right).$$

To show that $\mathbf{Var}(R|x)$ is small with high probability, we will upper bound $\mathbb{E}[\mathbf{Var}(R|x)]$.

$$\begin{aligned}
\mathbb{E}[\mathbf{Var}(R|x)] &= \sum_{i,j:|M_{i,j}| \leq s_2/k} |M_{i,j}| \left(\frac{s_2}{k} - |M_i, j| \right) \mathbb{E}[x_i^2 x_j^2] \\
&\leq \sum_{i,j} \frac{s_2}{k} |M_{i,j}| \, \mathbb{E}[x_i^2 x_j^2] \\
&\leq \frac{s_2 \times \|M\|_1 \times \nu^4}{k} && \text{(using } \mathbb{E}[x_i^2 x_j^2] \leq \sqrt{\mathbb{E}[x_i^4] \, \mathbb{E}[x_j^4]} = \nu^4\text{)} \\
&\leq s_2 \times \nu^4. && \text{(using } \|M\|_1 \leq k\text{)}
\end{aligned}$$

Thus Markov's inequality implies that $\mathbb{P}(\mathcal{E} \cap \mathcal{E}_3) \leq \mathbb{P}(\mathcal{E}_3) \leq (s_2 \times \nu^4)/(s_3 \times s_1^2)$.

Taking the union bound, we get the desired result. □

As reasoned above, we want the failure probability in Equation (4.9) to be less than $10^{-6} \times q$. That is,

$$\frac{\sigma^2}{s_1} + \frac{4}{s_3} + \frac{s_2 \times \nu^4}{s_3 \times s_1^2} \leq 10^{-6} \times q. \tag{4.10}$$

In Appendix B.3, we choose s_1, s_2, s_3 and q such that the bound holds. This, by the reasoning after Lemma 4.5.3, guarantees that Q satisfies the extra condition for $s_3\mathcal{A}_{k,P}$ (on top of being in $s_3\mathcal{A}_k$) with probability at least $1 - 10^{-6}$.

Taking a union bound, with failure probabilities 10^{-6} (for Q being in $s_3 \in \mathcal{A}_k$, Lemma 4.5.2) and 10^{-6} for satisfying the additional criterion for being in $s_3\mathcal{A}_{k,P}$, we conclude that Condition 1 happens with probability $1 - 2 \cdot 10^{-6}$.

Condition (II) Define the random variable Z to be

$$Z = \sum_i \mathbb{I}_{(x_i x_i^\top) \bullet Q > s_3 \times s_1}. \tag{4.11}$$

The second condition is equivalent to saying that $Z > 2 \times q \times |S|$, which we show to happen with probability at least 4×10^{-6}.

The strategy is to lower bound $\mathbb{E}[Z]$, and then use Paley-Zygmund to show that Z is large with constant probability. To lower bound the expectation, for any i such that $(x_i x_i^\top) \bullet M > B$, we want to lower bound $\Pr_Q((x_i x_i^\top) \bullet Q > s_3 \times s_1)$, using either Chebyshev's inequality or the Berry-Esseen theorem (see Fact 4.5.5). First, note that for these i, $\mathbb{E}[(x_i x_i^\top) \bullet Q] = (x_i x_i^\top) \bullet M > B$ by our assumption. If $\mathbf{Var}[(x_i x_i^\top) \bullet Q] \leq V_Z$, where V_Z is a fixed function of r and σ chosen in Appendix B.3, then by Chebyshev's inequality, we have

$$\Pr\left((x_i x_i^\top) \bullet Q \geq s_3 \times s_1\right) \geq \Pr\left((x_i x_i^\top) \bullet Q \geq B - 10 \times \sqrt{V_Z}\right) \geq 0.99\,, \tag{4.12}$$

where the first inequality is true by our choice of s_1, s_3, V_Z and B in Appendix B.3. Otherwise, we have the case where $\mathbf{Var}[(x_i x_i^\top) \bullet Q] > V_Z$. In this case, we treat $(x_i x_i^\top) \bullet Q$ as a sum of independent variables

$$(x_i x_i^\top) \bullet Q = \sum_{s,t} (x_i)_s (x_i)_t Q_{s,t}$$

and use the Berry-Esseen theorem, which requires bounding the sum of the third central absolute moment of the summands.

Fact 4.5.5 (Berry-Esseen Theorem for Sums of Independent Variables)**.** *Consider a random variable $\xi = \sum_i \xi_i$, where the variables ξ_i are independent (but not necessarily identical) and each of them has finite third moment. Denote μ_i as $\mathbb{E}[\xi_i]$, σ_i^2 as $\mathbf{Var}(\xi_i)$ and ρ_i as the third central absolute moment, namely $\rho_i = \mathbb{E}[|\xi_i - \mu_i|^3]$. Then,*

$$d_{\mathrm{K}}(\xi, \mathcal{N}(\sum_i \mu_i, \sigma_i^2)) \leq 0.57 \frac{\sum_i \rho_i}{(\sum_i \sigma_i^2)^{1.5}} = 0.57 \frac{\sum_i \rho_i}{(\mathbf{Var}(\xi))^{1.5}} \ ,$$

where d_{K} is the Kolmogorov distance between two distributions (namely, the ℓ_∞ distance between the cumulative density functions).

Let $\rho_{s,t}$ be the third central absolute moment of $(x_i)_s (x_i)_t Q_{s,t}$. For any (s,t) such that $0 < |M_{s,t}| \leq s_2/k$, we can calculate its third moment as follows:

$$\begin{aligned}
\rho_{s,t} &= \mathbb{E}[|(x_i)_s (x_i)_t Q_{s,t} - \mathbb{E}[(x_i)_s (x_i)_t Q_{s,t}]|^3] \\
&= |(x_i)_s|^3 |(x_i)_t|^3 \frac{|M_{s,t}|^3}{p_{s,t}^3} \mathbb{E}[|\mathrm{Ber}(p_{s,t}) - p_{s,t}|^3] \\
&= |(x_i)_s|^3 |(x_i)_t|^3 \frac{|M_{s,t}|^3}{p_{s,t}^3} p_{s,t}(1-p_{s,t})(1 - 2p_{s,t} + 2p_{s,t}^2) \\
&\leq |(x_i)_s|^3 |(x_i)_t|^3 \frac{|M_{s,t}|^3}{p_{s,t}^3} p_{s,t}(1-p_{s,t}) \qquad \text{for all } p_{s,t} \in [0,1]
\end{aligned}$$

$$\leq (\sigma^2 \times r^2 \times s_2)(x_i)_s^2(x_i)_t^2 \frac{M_{s,t}^2}{p_{s,t}^2} p_{s,t}(1-p_{s,t})$$

(since $|x_i|_\infty \leq \sigma \times r \times \sqrt{k}$ and $|M_{s,t}|/p_{s,t} = s_2/k$)

The same inequality holds trivially for (s,t), where $|M_{s,t}| \geq s_2/k$ or $M_{s,t} = 0$ since $\rho_{s,t} = 0$ in both of these edge cases. Thus, the sum of the third central absolute moment of the summands we need for Berry-Esseen is

$$\begin{aligned}\sum_{s,t} \rho_{s,t} &\leq (\sigma^2 \times r^2 \times s_2) \sum_{s,t} (x_i)_s^2 (x_i)_t^2 \frac{M_{s,t}^2}{p_{s,t}^2} p_{s,t}(1-p_{s,t}) \\ &= (\sigma^2 \times r^2 \times s_2)\, \mathbf{Var}\left((x_i x_i^\top) \bullet Q\right) \ ,\end{aligned}$$

where the last equality is a simple calculation to calculate the term-by-term variance for $(x_i x_i^\top) \bullet Q$. Thus, Fact 4.5.5 implies that the Kolmogorov distance between the distribution of $(x_i x_i^\top) \bullet Q$ and the Gaussian with the same mean and variance is at most

$$0.57 \frac{\sum_{s,t} \rho_{s,t}}{(\mathbf{Var}((x_i x_i^\top) \bullet Q))^{1.5}} \leq 0.57(\sigma^2 \times r^2 \times s_2) \frac{\mathbf{Var}\left((x_i x_i^\top) \bullet Q\right)}{\mathbf{Var}^{1.5}\left((x_i x_i^\top) \bullet Q\right)} \leq \frac{0.57(\sigma^2 \times r^2 \times s_2)}{\sqrt{V_Z}} \ , \tag{4.13}$$

where the inequality comes from the assumption that the variance is at least V_Z. Therefore, $(x_i x_i^\top) \bullet Q$ has at least probability $0.5 - \frac{0.57(\sigma^2 \times r^2 \times s_2)}{\sqrt{V_Z}}$ of exceeding its expectation. By our choice of quantities in Appendix B.3, this probability is at least 0.4. Furthermore, $\mathbb{E}((x_i x_i^\top) \bullet Q) = x_i x_i^\top \bullet M$ is bigger than B and in turn bigger than $s_3 \times s_1$ (by our choice for these quantities). Thus, with probability at least 0.4, $(x_i x_i^\top) \bullet Q$ exceeds $s_3 \times s_1$.

Combined with the guarantee that $\Pr_Q((x_i x_i^\top) \bullet Q > s_3 \times s_1) > 0.99$ in the case where $\mathbf{Var}((x_i x_i^\top) \bullet Q) \leq V_Z$ (cf. Equation (4.12)), we have shown that in all cases, $\Pr_Q((x_i x_i^\top) \bullet Q > s_3 \times s_1) > 0.4$ whenever $x_i x_i^\top \bullet M > B$.

Thus, we have shown that $\mathbb{E}[Z] = \sum_i \Pr_Q((x_i x_i^\top) \bullet Q > s_3 \times s_1) > 0.4 \times 0.01n =$

$0.004n$, since at least 0.01 fraction of points satisfy $x_i x_i^\top \bullet M > B$ and thus also satisfy $\Pr_Q((x_i x_i^\top) \bullet Q > s_3 \times s_1) > 0.4$. Note also that $Z \in [0, n]$ always, meaning that $\mathbb{E}[Z^2] \leq n^2$. Since $0.004 \geq 4 \times q$ by our choice of q, it then follows from the Paley-Zygmund inequality that

$$\Pr(Z > 2 \times q \times |S|) \geq 0.25 \frac{(\mathbb{E}[Z])^2}{n^2} > \frac{0.25 \times 0.004^2 n^2}{n^2} = 4 \cdot 10^{-6}$$

showing the second claim above, and completing the proof of this lemma. □

4.6 Smoothness of Stability Under Truncation

The goal of this section is to prove Theorem 4.4.1 (restated below), the stability result we use in the proof of our main result, Theorem 4.4.2 and hence Theorem 4.1.4. Recall the function $h_{a,b}$ as defined in Equation (4.2).

Theorem 4.4.1. *Let S be a set of n i.i.d. data points from a distribution P over $\mathbb{R}^d$, and let T be a 0.01-corruption of S. Let $\tilde{\mu}$ be the coordinate-wise median-of-means estimate computed from set T. Let the mean of P be μ and covariance Σ such that $\|\Sigma\|_{\mathcal{X}_k} \leq \sigma^2$, and for all $i \in [d]$, $\mathbb{E}[X_i^4] \leq O(\sigma^4)$. Suppose that $n = \Omega(k^2 \log d + \log(1/\tau))$. Let $a = \sigma\sqrt{k}$. With probability $1 - \tau$ over S, for all T we have that there exists a subset $S' \subseteq T$ with $|S'| \geq 0.95n$ such that $h_{a,\tilde{\mu}}(S')$ is $(0.01, O(1), k)$-stable with respect to some μ' and σ with $\|\mu' - \mu\|_\infty \leq O(\sigma/\sqrt{k})$.*

In Section 4.5, we proved Theorem 4.5.1. While it is tempting to directly use Theorem 4.5.1 to prove the main result of Theorem 4.4.2, it does not apply as-is for analyzing Algorithm 1. If we tried to use Theorem 4.5.1 to prove Theorem 4.4.2, the intuitive way is to apply Theorem 4.5.1 to the distribution $h_{a,\mu}(X)$ for $X \sim P$ — $h_{a,\mu}(X)$ is by construction bounded in ℓ_∞ norm, and the means and covariances of P and $h_{a,b}(X)$ are close in $\ell_{2,k}$ norm and $\mathcal{X}_k$ norm, respectively, by Lemma 4.3.2. However, in Algorithm 1, we

do not truncate samples by centering at μ, but instead use the coordinate-wise median-of-means estimate as the truncation center, which is data-dependent and not any fixed vector. Thus, we have to show that the stability result in Theorem 4.5.1 is insensitive to which point we center the truncation at, and that as a corollary, an analogous result holds even if we truncate using the coordinate-wise median-of-means estimate as the center. The final statement of this section is captured by Theorem 4.4.1, and much of this section is dedicated to showing the "Lipschitzness" of the stability of samples, as we truncate using different preliminary mean estimates.

To show this "Lipschitzness" property, we make the following observation. Suppose we start with the set of n i.i.d. samples S from P, which we know by Theorem 4.5.1 contains a large subset S_1 such that $h_{a,\mu}(S_1)$ is stable with respect to the mean vector of the truncated distribution $\mu' = \mathbb{E}_{X\sim P}[h_{a,\mu}(X)]$. Further suppose we are able to show that S contains another large subset S_2 that is "coordinate-wise regular", meaning that for each coordinate $j \in [d]$, most samples in S_2 are close to μ_j in coordinate j. Then the intersection $S_3 = S_1 \cap S_2$ enjoys both stability and coordinate-wise regularity, and furthermore the stability of $h_{a,b}(S_3)$ holds for any vector b that is close to both μ and μ'. This observation is shown as Lemma 4.6.1 in Section 4.6.1, and we show the existence of S_2 in Lemma 4.6.3 in Section 4.6.2. The proof of Theorem 4.4.1 combines the above two lemmas and Theorem 4.5.1, and is presented in Section 4.6.3.

4.6.1 Lipschitzness of Truncation Under Coordinate-wise Regularity

As explained before this subsection, Lemma 4.6.1 shows that if our set of uncontaminated samples S is such that 1) there exists a large subset S_1 with $h_{a,\mu}(S_1)$ being stable, and 2) there is another large "coordinate-wise regular" subset S_2, then $S_3 = S_1 \cap S_2$ is a large subset of S that is *both* "coordinate-wise regular" and $h_{a,b}(S_3)$ is stable for any b sufficiently close to μ. In the following statement, instead of saying that $h_{a,\mu}(S_1)$ and

$h_{a,b}(S_3)$ are stable, we use the essentially equivalent condition (by Proposition B.1.1) that the $\mathcal{X}_k$-norms of the empirical covariance matrices are small.

Lemma 4.6.1 (Lipschitzness of Truncation under Coordinate-wise Regularity)**.** *Let μ, μ' be vectors in $\mathbb{R}^d$ and let $a \in \mathbb{R}_+$ be greater than 2. Let $S = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ be the set of n points. Suppose there exist a set $S_1 \subset [n]$ satisfying the following for some $r \in \mathbb{R}_+$:*

$$|S_1| \geq 0.98n \;\; \textit{and} \;\; \left\| \frac{1}{|S_1|} \sum_{i \in S_1} (h_{a,\mu}(x_i) - \mu')(h_{a,\mu}(x_i) - \mu')^\top \right\|_{\mathcal{X}_k} \leq r \; . \tag{4.14}$$

Suppose also that, for some $\gamma \in (0,1)$, there exist a set $S_2 \subset [n]$ satisfying the following:

$$|S_2| \geq 0.99n \;\; \textit{and} \;\; \forall j \in [d] : \sum_{i \in S_2} \mathbb{I}_{|x_{i,j} - \mu_j| \geq a/2} \leq \gamma n \; . \tag{4.15}$$

Then, we have the following: there exists a set $S_3 \subset [n]$ such that for all $b \in \mathbb{R}^d$ satisfying $\|b - \mu\|_\infty \leq a/2$ and $\|b - \mu'\|_\infty \leq a$, we have that

$$|S_3| \geq 0.97n \;\; \textit{and} \;\; \left\| \frac{1}{|S_3|} \sum_{i \in S_3} (h_{a,b}(x_i) - \mu')(h_{a,b}(x_i) - \mu')^\top \right\|_{\mathcal{X}_k} \leq 1.1r + 5a\gamma k \|b - \mu\|_\infty \; . \tag{4.16}$$

Proof. We will take $S_3 = S_1 \cap S_2$, which directly implies that $|S_3| \geq 0.97n$. For any $M \in \mathcal{X}_k$, since $xx^\top \bullet M \geq 0$, we have the following:

$$\begin{aligned} &\left\langle M, \frac{1}{|S_3|} \sum_{i \in S_3} (h_{a,\mu}(x_i) - \mu')(h_{a,\mu}(x_i) - \mu')^\top \right\rangle \\ &\qquad \leq \left\langle M, \frac{1}{|S_3|} \sum_{i \in S_1} (h_{a,\mu}(x_i) - \mu')(h_{a,\mu}(x_i) - \mu')^\top \right\rangle \\ &\qquad \leq \frac{1}{0.97} r \; . \end{aligned}$$

Let $F(b)$ be the following matrix:

$$F(b) = \frac{1}{|S_3|} \sum_{i \in S_3} (h_{a,b}(x_i) - \mu')(h_{a,b}(x_i) - \mu')^\top .$$

We will establish that $\|F(b) - F(\mu)\|_{\chi_k} \leq 5a\gamma k \|b - \mu\|_\infty$, which establishes the lemma statement by the triangle inequality. In order to do that, we will show that $\|F(b) - F(\mu)\|_\infty \leq 5a\gamma \|b - \mu\|_\infty$ and then use Lemma 4.6.2 below (proved in Appendix B.1.3).

Lemma 4.6.2. *Let $A \in \mathbb{R}^{d \times d}$ be a symmetric matrix such that $|A_{i,i}| \leq \eta_1$ for each $i \in [d]$, and $|A_{i,j}| \leq \eta_2$ for each $i \neq j \in [d] \times [d]$. Then $\|A\|_{\chi_k} \leq \eta_1 + k\eta_2$.*

Consider an arbitrary (j, ℓ)-entry of these matrices. By abusing notation, when x and y are scalar, we use $h_{a,y}(x)$ to be the function from $\mathbb{R} \to \mathbb{R}$ defined analogously to Equation (4.2). Let $g(\cdot, \cdot)$ be the following function that is equal to the (j, ℓ) entry of the matrix $F(b)$, which is explicitly

$$g(b_j, b_\ell) = \frac{1}{|S_3|} \sum_{i \in S_3} (h_{a,b_j}(x_j) - \mu'_j)(h_{a,b_\ell}(x_\ell) - \mu'_\ell) .$$

We will show that $g(\cdot, \cdot)$ is locally Lipschitz in its arguments. Consider a particular $i \in S_3$ and define the following:

$$g_i(b_j, b_\ell) = (h_{a,b_j}(x_{i,j}) - \mu'_j)(h_{a,b_\ell}(x_{i,\ell}) - \mu'_\ell).$$

Then, we can upper bound the difference for each sample by

$$\begin{aligned}
&|g_i(b_j,b_\ell) - g_i(\mu_j, \mu_\ell)| \\
&\quad = |(h_{a,b_j}(x_{i,j}) - \mu'_j)(h_{a,b_\ell}(x_{i,\ell}) - \mu'_\ell) - (h_{a,\mu_j}(x_{i,j}) - \mu'_j)(h_{a,\mu_\ell}(x_{i,\ell}) - \mu'_\ell)| \\
&\quad \leq |(h_{a,b_j}(x_{i,j}) - h_{a,\mu_j}(x_{i,j}))(h_{a,b_\ell}(x_{i,\ell}) - \mu'_\ell)| \\
&\qquad + |(h_{a,\mu_j}(x_{i,j}) - \mu'_j)(h_{a,b_\ell}(x_{i,\ell}) - h_{a,\mu_\ell}(x_{i,\ell}))|
\end{aligned}$$

$$\leq (a + \|b - \mu'\|_\infty) \cdot \|b - \mu\|_\infty \left(\mathop{\mathbb{I}}_{|x_{i,j} - \mu_j| \geq a - \|b-\mu\|_\infty} + \mathop{\mathbb{I}}_{|x_{i,\ell} - \mu_\ell| \geq \|b-\mu\|_\infty} \right)$$
$$\leq (a + \|b - \mu'\|_\infty) \cdot \|b - \mu\|_\infty \left(\mathop{\mathbb{I}}_{|x_{i,j} - \mu_j| \geq a/2} + \mathop{\mathbb{I}}_{|x_{i,\ell} - \mu_\ell| \geq a/2} \right)$$
$$\leq 2a \cdot \|b - \mu\|_\infty \left(\mathop{\mathbb{I}}_{|x_{i,j} - \mu_j| \geq a/2} + \mathop{\mathbb{I}}_{|x_{i,\ell} - \mu_\ell| \geq a/2} \right),$$

where we use that $|h_{a,y}(x) - h_{a,z}(x)| \leq |y - z|$, $|h_{a,y}(x) - z| \leq |a + y - z|$, and $|h_{a,y}(x) - h_{a,z}(x)|$ is non-zero only if $|x - y| \geq a - |y - z|$.

Combined with assumption Equation (4.15), this implies that

$$|g(b_j, b_\ell) - g(\mu_j, \mu_\ell)| \leq 2a \cdot \|b - \mu\|_\infty \cdot \frac{2\gamma}{0.97} \leq 5a\gamma \|b - \mu\|_\infty.$$

By Lemma 4.6.2, we have the following:

$$\|F(b) - F(\mu)\|_{\mathcal{X}_k} \leq k \|F(b) - F(\mu)\|_\infty \leq 5a\gamma k \cdot \|b - \mu\|_\infty.$$

□

4.6.2 Existence of Large Subset of "Coordinate-wise Regular" Samples

This subsection shows Lemma 4.6.3, which states that with high probability there exists a large subset of "coordinate-wise regular" samples where in each dimension at most a negligible fraction of the points have large magnitude. As explained earlier, we will combine Lemmata 4.6.1 and 4.6.3 to show Theorem 4.4.1.

For a vector $X_i \in \mathbb{R}^d$, we will use $X_{i,j}$ to refer to the j-th coordinate of X_i.

Lemma 4.6.3. *Let P be a distribution over $\mathbb{R}^d$, and $k \in [d]$. For $X \sim P$, suppose for all $j \in [d]$, $\mathbb{E}[X_j^4] \leq \nu^4$. Then, there exists a positive constant c_1 such that, with probability at least $1 - \tau$*

over the set S of $n \geq c_1(k^{1.5} + \log(1/\tau))$ i.i.d. samples from P, S contains a (large) subset S' such that the following hold simultaneously:

1. $|S'| \geq 0.99|S|$, *and*
2. *For each $j \in [d]$, the number of points in S' with their j-th coordinate at least $2\nu\sqrt{k}$ in magnitude is at most $n/k^{1.5}$. Equivalently,*

$$\forall j \in [d] : \quad \sum_{i \in S} \mathbb{I}_{|X_{i,j}| \geq 2\nu\sqrt{k}} \leq \frac{n}{k^{1.5}} \ . \tag{4.17}$$

Before providing the proof of Lemma 4.6.3, we highlight why the result is not obvious. The first approach that one may try is to show that the original set S directly satisfies the claim, that is, (with high probability) in each coordinate, the fraction of points with large magnitude in that coordinate is at most $k^{-1.5}$. At the population level, this is indeed true by the fourth moment assumption: for any fixed $i \in [n]$ and $j \in [d]$, the probability that $|X_{i,j}|$ is large is at most $O(1/k^2)$. However, for this to hold with probability $1 - \tau$, one requires roughly $k^{1.5}\log(1/\tau)$ samples *even in 1 dimension*[8], which would give a multiplicative dependence on $\log(1/\tau)$ instead of additive dependence.

The second approach that one may try would be the following: define S' to be the set of all "good" samples, where we say a sample is "good" if all of its coordinates are smaller than $c\nu\sqrt{k}$. However, for any fixed coordinate $j \in [d]$, the probability that the j-th coordinate of a sample being larger than $c\nu\sqrt{k}$ can be as large as $1/k^2$. Thus, when $d \gg k^2$, the probability that a particular sample is "bad" may be arbitrarily close to 1 — for example, when coordinates are independent — and the resulting set S' will be too small with high probability.

We now give the proof of Lemma 4.6.3, which phrases the existence of the set S' as an integer program feasibility problem. The proof considers the LP relaxation and uses

[8]The upper bound follows from a Chernoff bound, and the lower bound follows from the fact that Chernoff bounds are essentially tight for sums of Bernoulli coins.

LP duality techniques to show that the integer program has to be feasible.

Proof. We will assume that $k \geq C$ for a large enough constant. If k is smaller than the constant, then the result follows by applying Bernstein inequality and taking $S' = S$.

Let $S = \{Y_1, \ldots, Y_n\}$. For $i \in [n]$ and $j \in [d]$, we use $Z_{i,j}$ to denote $\mathbb{I}_{|Y_{i,j}| \geq c_2 \nu \sqrt{k}}$. For simplicity, we set $\alpha = k^{-1.5}/3$. Our goal is to show that the following integer program is feasible:

$$\begin{aligned} \text{variables} \quad & p_1, \ldots, p_n \\ \text{subject to} \quad & \forall j \in [d] : \sum_{i=1}^{n} p_i Z_{i,j} \leq 3\alpha n \\ & \sum_{i=1}^{n} p_i \geq 0.99n \\ & \forall i \in [n] : p_i \in \{0, 1\}. \end{aligned} \tag{F1}$$

As argued above in the prose after the statement, one needs to argue about all the samples, and their coordinates, simultaneously to prove the statement. Since directly handling the feasibility program (F1) seems difficult, our argument will go in the following steps: (i) first consider the LP relaxation of (F1), (ii) using duality theory, the LP relaxation is feasible if and only if the dual LP is infeasible, (iii) simplify the dual LP and show that, with high probability, the resulting program is infeasible.

We begin by considering the LP relaxation.

$$\begin{aligned} \text{variables} \quad & p_1, \ldots, p_n \\ \text{subject to} \quad & \forall j \in [d] : \sum_{i=1}^{n} p_i Z_{i,j} \leq \alpha n \\ & \sum_{i=1}^{n} p_i \geq 0.999n \\ & \forall i \in [n] : p_i \in [0, 1]. \end{aligned} \tag{F2}$$

We first show that if the above LP relaxation, (F2), is feasible, then (F1) is also feasible.

Claim 4.6.4 (Feasibility of (F2) implies feasibility of (F1))**.** *Suppose* $n > 10^6$ *and* $\alpha \geq (4 \log n)/n$*. If (F2) is feasible, then (F1) is also feasible.*

Proof. Let $p_1, \ldots, p_n$ be the feasible solution to (F2). Consider the following random assignment, for $i \in [n]$, $P_i \sim \mathrm{Ber}(p_i)$ independently. We will show that, with non-zero probability, P_i's satisfy (F1). We will use the following inequality:

Fact 4.6.5 (Chernoff Inequality)**.** *Let* $a_1, \ldots, a_n$ *such that* $a_i \in \{0, 1\}$*. Let* $W_1, \ldots, W_n$ *be independent Bernoulli random variables and consider the random variable* $Z = \sum_{i=1}^{n} W_i$*. Then, with probability* $1 - \tau$*,* $Z \leq 2(\mathbb{E}\, Z + \log(1/\tau))$*.*

By Fact 4.6.5, we get that each of the inequalities in (F1) holds with probability $1 - 1/(2n)$ as long as $n\alpha \geq 2\log(2n)$ and $n > 1000 \log(2n)$. The latter holds when $n \geq 10^6$. □

Since $n > 10^6$ in our setting (as k is large and choosing c_1 to be large enough) and $\alpha = 1/(3k^{1.5})$, we have that $\alpha \geq 4(\log n)/n$ is equivalent to $n \geq 12k^{1.5} \log n$, which is satisfied when $n \geq 100k^{1.5} \log k$. The latter holds when $n \geq ck^{1.5} \log d$ for a large enough constant c. Thus, in the remainder of this section, we will show that, with high probability, this LP program is indeed feasible. We begin by considering the following dual program:

$$\begin{aligned} \text{variables} \quad & w_1, \ldots, w_d, y_1, \ldots, y_n, x \\ \text{subject to} \quad & \sum_{i=1}^{n} y_i + \alpha n \sum_{j=1}^{d} w_j < 0.999nx \\ & \forall i \in [n] : y_i + \sum_{j=1}^{d} Z_{i,j} w_j \geq x \\ & z \geq 0, \quad \forall i \in [n] : y_i \geq 0, \quad \forall j \in [d] : w_j \geq 0. \end{aligned} \tag{F3}$$

Suppose for the sake of contradiction that (F2) is infeasible. By Farkas' lemma [GKT51], it means that the (dual) program in (F3) is feasible. Formally, we have the following claim:

Claim 4.6.6 (LP Duality for (F2))**.** *(F3) is infeasible if and only if (F2) is feasible.*

Claim 4.6.6 follows from Farkas' lemma. We will argue that (F3) is infeasible by showing that the following program, which is feasible whenever (F3) is feasible, is infeasible.

$$\begin{aligned} \text{variables} \quad & w_1, \ldots, w_d, A \\ \text{subject to} \quad & \forall i \in A : \sum_{j=1}^{d} Z_{i,j} w_j \geq \alpha(\sum_{j=1}^{d} w_j) \\ & \forall j \in [d] : w_j \geq 0, \\ & A \subset [n], |A| \geq 10^{-3} n \end{aligned} \tag{F4}$$

(F4) states that for at least 10^{-3} fraction of i's in n, the following inequality holds: $\sum_{j=1}^{d} Z_{i,j} w_j \geq \alpha \|w\|_1$. The following claim relates the two programs above.

Claim 4.6.7. *If (F3) is feasible, then (F4) is feasible.*

Proof. Let $y_1, \ldots, Y_n, w_1, \ldots, w_d, x$ be any feasible solution to (F3). Then the first constraint in (F3) that the average of y_i's is less than $0.999x - \alpha(\sum_{j=1}^{d} w_j)$. By Markov's inequality, the fraction of the y_i's such $y_i \geq (x - \alpha(\sum_{j=1}^{d} w_j))$ is at most $\frac{0.999x - \alpha(\sum_{j=1}^{d} w_j)}{(x - \alpha(\sum_{j=1}^{d} w_j))} \leq 0.999$. Thus the fraction of y_i's such that $y_i < (x - \alpha(\sum_{j=1}^{d} w_j))$ is at least 0.001.

Let $A \subset [n]$ be the set of such indices. For any $i \in A$, the second constraint in (F3) implies that $\sum_{j=1}^{d} Z_{i,j} w_j \geq x - y_i \geq \alpha(\sum_{j=1}^{d} w_j)$. This implies that (F4) is feasible. □

In order to argue that (F4) is infeasible, we first consider a particular w. Using calculations provided below, it can be seen that the probability that a particular w satisfies (F4) is exponentially small in n. However, a direct approach at covering w

seems difficult since w is a dense vector in $\mathbb{R}^d$ and $n = o(d)$. Using a randomized rounding mechanism, we show that it suffices to consider only sparse w as follows:

$$\begin{aligned} \text{variables} \quad & w_1, \ldots, w_d, A \\ \text{subject to} \quad & \forall i \in A : \sum_{j=1}^{d} Z_{i,j} w_j \geq 1 \\ & \forall j \in [d] : w_j \in \{0, 1\}, \\ & \sum_{j=1}^{d} w_j \leq \frac{2 \times 10^7}{\alpha} \\ & A \subset [n], |A| \geq 10^{-4} n \end{aligned} \tag{F5}$$

The following claim shows that if (F4) is feasible then (F5) is also feasible.

Claim 4.6.8. *If (F4) is feasible, then (F5) is also feasible.*

Proof. Let $w_1, \ldots, w_d$ and A be the feasible solution to (F5). Set $q_j = \min(1, w_j/(\alpha \|w\|_1))$ for $j \in [d]$. Consider the following random assignment: set $W_j \sim \mathrm{Ber}(q_j)$ independently for $j \in [d]$. We will show that with non-zero probability W_j's satisfy (F5). Consider the following events:

$$\mathcal{E}_1 := \left\{ \sum_{j=1}^{d} W_j \leq 2 \times 10^{-7} (1/\alpha) \right\}, \quad \text{and} \quad \mathcal{E}_2 := \left\{ |\{i : \sum_{j=1}^{d} Z_{i,j} W_j \geq 1\}| \geq 10^{-4} n \right\} \tag{4.18}$$

We will show that $\mathbb{P}\{cE_1\} \geq 1 - 5 \times 10^{-8}$ and $\mathcal{P}\{cE_2\} \geq 10^{-7}$. By a union bound, we will have that $\mathcal{E}_1 \cap \mathcal{E}_2$ has non-zero probability and thus (F5) is feasible.

Let $F := \{j \in [d] : W_j = 1\}$ be the set of coordinates where W_j is non-zero. Then $\mathbb{E}[|F|] = \mathbb{E}[\sum_{j=1}^{d} W_j] = \sum_{j=1}^{d} q_j \leq 1/\alpha$. Thus with probability at least $1 - 5 \times 10^{-8}$, we have that the number of non-zero W_j's is at most $\frac{2 \times 10^7}{\alpha}$. Equivalently, $\mathbb{P}\{\mathcal{E}_1\} \geq 1 - 5 \times 10^{-8}$.

We now focus on the second event $\mathcal{E}_2$. Let $S_1, \ldots, S_n$ be the subsets of $[d]$ such that $S_i = \{j \in [d] : Z_{i,j} = 1\}$, i.e., for each sample i, S_i is the set of indices where the

coordinates are large. Consider the random variables $R_1, \ldots, R_n$, where for $i \in [n]$, $R_i := \sum_{j=1}^{d} Z_{i,j} W_j = \sum_{j \in S_i} Z_{i,j} W_j$. (F5) requires that for at least 10^{-4} fraction of i's, $R_i \geq 1$. Since $Z_{i,j}$'s are binary and fixed, we have that R_i is distributed as Binomial random variable and is thus anti-concentrated.

Fact 4.6.9 (Anti-concentration of Binomial)**.** *Let $X \sim$ Binomial(n, p) for some $n \in \mathbb{N}$ and $p \in [0, 1]$. Suppose $\mathbb{E}[X] \geq 1$. Then $\mathbb{P}\{X \geq 1\} \geq (1 - 1/e)$.*

Proof. Using the fact that $1 + x \leq e^x$ for all $x \in \mathbb{R}$, we get the following:

$$\mathbb{P}\{X \geq 1\} = 1 - \mathbb{P}\{X = 0\} = 1 - (1-p)^n \geq 1 - (e^{-p})^n = 1 - e^{-np} \geq (1 - 1/e). \qquad \square$$

Consider a fixed $i \in A$. Then either there exists a $j \in S_i$ such that $q_j = 1$, or for all $j \in S_i$, $q_j < 1$. In the former case, we have that R_i is at least one since $W_j = 1$.

In the latter setting, we have that $q_j = w_j/(\alpha \|w\|_1)$ for all $j \in S_i$, and thus $\mathbb{E}[R_i] = \sum_{j \in S_i}^{d} Z_{i,j} q_j = \sum_{j=1}^{d} Z_{i,j} w_j/(\alpha \|w\|_1) \geq 1$. Applying Fact 4.6.9 to any such $i \in A$, we get that the probability of R_i being positive is at least $1 - 1/e$. Let A' be set of i's such that $R_i \geq 1$, i.e., $A' = \{i : R_i \geq 1\}$. Thus combining the two cases above, we have the following:

$$\forall i \in A : \mathbb{P}\{i \in A'\} \geq 0.5. \tag{4.19}$$

Thus $\mathbb{E}[|A'|] \geq 0.5|A| \geq 5 \times 10^{-4}$. Since $|A'|$ lies in $[0, n]$, applying Paley-Zygmund inequality to the random variable $|A'|$, we get the following:

$$\mathbb{P}\{|A'| \geq 10^{-4} n\} \geq \mathbb{P}\{|A'| \geq 0.2\,\mathbb{E}[|A'|]\} \geq 0.64 \frac{(\mathbb{E}[|A'|])^2}{n^2} \geq 0.64 \times 25 \times 10^{-8} > 10^{-7}. \tag{4.20}$$

Equivalently, $\mathbb{P}\{\mathcal{E}_2\} \geq 10^{-7}$. This completes the proof. $\square$

Thus, it suffices to show that, with high probability, (F5) is infeasible.

Lemma 4.6.10 (Infeasibility of (F5))**.** *Under the setting of Lemma 4.6.3 and when $k > 10^{26}$, there exists a constant $c_1 > 0$ such that if $n \geq c_1(k^{1.5}\log d + \log(1/\tau))$, then with probability $1-\tau$, (F5) is infeasible.*

Proof. First consider any fixed $w = (w_1, \ldots, w_d)$ such that $w_i \in \{0, 1\}$ and $\sum_{j=1}^{d} w_j \leq 2 \times 10^7 \cdot (1/\alpha)$.

Consider the integer-valued random variables $R_1, \ldots, R_n$ such that $R_i = \sum_{j=1}^{d} Z_{i,j} w_j$, and observe that R_i's are i.i.d. random variables (since X_i's are i.i.d. random variables). Thus, (F5) requires that at least $10^{-4}\%$ of R_i's are non-zero.

By the fourth moment bound on each coordinate, we have that $\mathbb{E}[Z_{i,j}] = \mathbb{P}\{X_{i,j} \geq 2\nu\sqrt{k}\} \leq 1/k^2$ for each i and j. Therefore, the expectation of each R_i is at most

$$\sum_{j=1}^{d} w_j \,\mathbb{E}[Z_{i,j}] \leq \sum_{j=1}^{d} w_j (1/k^2) \leq (2 \times 10^7)/(k^2\alpha) = (2 \times 10^7)/(k^2\alpha) = (6 \times 10^7)/\sqrt{k}$$

, which is less than 10^{-5} for k large enough. By Markov's inequality, the probability that $\mathbb{P}\{R_1 \geq 1\} \leq 10^{-5}$.

Hence, by the Chernoff bound (since R_i's are independent), with probability at least $1 - \mathbb{E}(-c'n)$, the fraction of R_i's that are non-zero is at most 5×10^{-5}. Hence, with the same probability, this particular choice of w does not satisfy (F5). Since there are at most $d^{(2\times 10^7)/\alpha)}$ such choices of w, applying a union bound, we get that (F5) is infeasible with probability at least $1 - \mathbb{E}((2 \times 10^7)/\alpha) \cdot \log d - c'n)$. The failure probability is at most τ when $n \gtrsim \log(1/\tau) + k^{1.5}\log d$. This concludes the proof. □

Since we assumed k is large enough, Lemma 4.6.10 is applicable. Lemma 4.6.10 implies that, with high probability, the program (F5) is infeasible. Hence, with the same high probability, the programs (F4) and (F3) are also infeasible, and the programs (F1) and (F2) are feasible. This completes the proof. □

4.6.3 Proof of Theorem 4.4.1

We now combine Lemmata 4.6.1 and 4.6.3 to show Lemma 4.6.11, stating that with high probability over the uncontaminated and untruncated samples S, there is a large subset S' such that for any truncation center b that is close to the true mean, $h_{a,b}(S')$ is stable for $a = \Theta(\sigma\sqrt{k})$ as chosen in Algorithm 1. Theorem 4.4.1 follows a corollary, by instantiating b to be the coordinate-wise median-of-means estimate.

Lemma 4.6.11. *Let S be a set of n i.i.d. data points from a distribution P over $\mathbb{R}^d$. Let the mean of P be μ, and covariance Σ such that $\|\Sigma\|_{\mathcal{X}_k} \le \sigma^2$, and for all $i \in [d]$, $\mathbb{E}[X_i^4] \le O(\sigma^4)$. Suppose $n = \Omega(k^2 \log d + \log(1/\tau))$. Let $a = 4\sigma\sqrt{k}$. With probability $1-\tau$ over S, there exists a subset $S' \subset S$ with $|S'| \ge 0.95n$ such that for any b satisfying $\|b - \mu\|_\infty = O(\sigma)$, we have $h_{a,b}(S')$ is $(0.01, O(1), k)$-stable with respect to some μ' and σ with $\|\mu' - \mu\|_\infty \le O(\sigma/\sqrt{k})$.*

Proof. Let P' be the distribution of $h_{a,\mu}(P)$ and let μ' and Σ' be the mean and covariance of P'. This will be the μ' in the lemma statement. By Lemma 4.3.2, we get that (i) $\|\mu' - \mu\|_\infty \le \sigma/\sqrt{k}$, (ii) $\|\Sigma - \Sigma'\|_{\mathcal{X}_k} \le O(\sigma^2)$, (iii) P' is supported on the set $\{x : \|x - \mu'\|_\infty \le 2a\}$, and (iv) the axis-wise fourth moment of P' is upper bounded by a constant multiple of that of P. Thus Theorem 4.5.1 can be applied to P'.

Applying Theorem 4.5.1 to P' gives that, with probability at least $1-\tau$, there exists a subset $S_1 \subset S$ (which are samples from P, before truncation) with $|S_1| \ge 0.98n$ such that $h_{a,\mu}(S_1)$ is $(0.01, O(1), k)$-stable with respect to μ'. In particular, we have

$$\left\| \frac{1}{|S_1|} \sum_{i \in S_1} (h_{a,\mu}(x_i) - \mu')(h_{a,\mu}(x_i) - \mu')^\top \right\|_{\mathcal{X}_k} \le O(\sigma^2). \tag{4.21}$$

Let $r := \sigma / (\max_{j \in [d]} \mathbb{E}_{X \sim P}[X_j^4])^{1/4}$. By our assumption on P in the lemma, we have $r = \Theta(1)$. By applying Lemma 4.6.3 to $P - \mu$, and using kr^2 in place of k in Lemma 4.6.3, with probability at least $1-\tau$, there exists a subset $S_2 \subset S$ with $|S_2| \ge 0.99n$ such that

for all $j \in [d]$,

$$\sum_{i \in S_2} \mathbb{I}_{|x_{i,j} - \mu_j| \geq a/2} = \sum_{i \in S_2} \mathbb{I}_{|x_{i,j} - \mu_j| \geq 2\nu\sqrt{kr^2}} \leq O(k^{-1.5} r^{-3}) n \leq O(k^{-1.5}) n. \tag{4.22}$$

We can then apply Lemma 4.6.1 to show that, conditioned on the above two existence events, there exists a third subset $S_3 \subset S$ with $|S_3| \geq 0.97n$ such that for all b satisfying $\|b - \mu\|_\infty \leq O(\sigma)$ and $\|b - \mu'\|_\infty \leq O(\sigma)$ (the latter holds by the triangle inequality for all b with $\|b - \mu\|_\infty \leq O(\sigma)$), we have that

$$\left\| \frac{1}{|S_3|} \sum_{i \in S_3} (h_{a,b}(x_i) - \mu')(h_{a,b}(x_i) - \mu')^\top \right\|_{\mathcal{X}_k} \leq O(\sigma^2) + O(ak^{-1.5} k \|b - \mu\|_\infty) \tag{4.23}$$

$$\leq O(\sigma^2). \tag{4.24}$$

By Proposition B.1.1, this implies S_3 contains a set S' satisfying the following: (i) $|S'| \geq 0.95n$ and (ii) $h_{a,b}(S')$ is $(0.1, O(1), k)$-stable with respect to μ' and σ, for any b satisfying $\|b - \mu\|_\infty \leq O(\sigma)$. Thus, we choose S' in the lemma statement to be this set.

Taking a union bound, all the above events fail with probability at most $O(\tau)$. Reparameterizing yields the lemma statement.

□

Theorem 4.4.1. *Let S be a set of n i.i.d. data points from a distribution P over $\mathbb{R}^d$, and let T be a 0.01-corruption of S. Let $\tilde{\mu}$ be the coordinate-wise median-of-means estimate computed from set T. Let the mean of P be μ and covariance Σ such that $\|\Sigma\|_{\mathcal{X}_k} \leq \sigma^2$, and for all $i \in [d]$, $\mathbb{E}[X_i^4] \leq O(\sigma^4)$. Suppose that $n = \Omega(k^2 \log d + \log(1/\tau))$. Let $a = \sigma\sqrt{k}$. With probability $1 - \tau$ over S, for all T we have that there exists a subset $S' \subseteq T$ with $|S'| \geq 0.95n$ such that $h_{a,\tilde{\mu}}(S')$ is $(0.01, O(1), k)$-stable with respect to some μ' and σ with $\|\mu' - \mu\|_\infty \leq O(\sigma/\sqrt{k})$.*

Proof. By Fact 4.2.1, we know that with probability at least $1 - \tau$, we have $\|\tilde{\mu} - \mu\|_\infty \leq O(\sigma)O(1 + (\log(d/\tau))/n) = O(\sigma)$ by the assumption that n is sufficiently large.

Thus, we use $\tilde{\mu}$ as "b" in Lemma 4.6.11 to yield the stability guarantee in the theorem statement.

The total failure probability is at most 2τ, and reparameterizing yields the theorem statement. □

4.7 Information-Theoretic Lower Bound

In this section, we show that the asymptotic error of Theorem 4.1.4 is optimal under a mild assumption on k. Let $\mathcal{D}_k$ be the family of all distributions over $\mathbb{R}^d$ that satisfy the following:

1. For every $D \in \mathcal{D}_k$, the mean of D is k-sparse,
2. For every D in $\mathcal{D}_k$ the covariance of $\mathcal{D}$ is upper bounded by I in spectral norm, and
3. For every $D \in \mathcal{D}_k$ we have that $\mathbb{E}[(X_i - \mathbb{E}[X_i])^4] = O(1)$, where $X = (X_1, \ldots, X_d) \sim D$.

Lemma 4.7.1. *Let $k \geq 1/\sqrt{\epsilon}$. Then there exist two distributions in D_1 and D_2 in $\mathcal{D}_k$ such that the following hold: (i) $d_{TV}(D_1, D_2) = \epsilon$, and (ii) The means of D_1 and D_2 are separated by $\Omega(\sqrt{\epsilon})$ in $\ell_{2,k}$-norm.*

Before giving the proof of Lemma 4.7.1, we remark that the assumption of $k \geq 1/\sqrt{\epsilon}$ is mild. First, the assumption is independent of the ambient dimensionality d—the most challenging parameter regime in algorithmic robust statistics is when we fix a small ϵ and then take the dimensionality d to ∞. Second, the typical interesting sparsity regime is when k is super-constant but grows very slowly in d, say, logarithmically. The assumption that $k \geq 1/\sqrt{\epsilon}$ applies readily to the above regime.

Proof. Let D_1 be the distribution that places all of its mass at origin, i.e., $(0, \ldots, 0)$. Let D_2 be the distribution that places $(1 - \epsilon)$ probability mass at origin and places ϵ probability

mass at y, where the first k-coordinates of y are α for some α to be decided later and the remaining $d-k$ coordinates are 0.

It is easy to see that the total variation distance between D_1 and D_2 is ϵ, and that $D_1 \in \mathcal{D}_k$. We will now show that $D_2 \in \mathcal{D}_k$ for a suitable value of α.

1. First the mean of D_2 is ϵy, which is k-sparse by construction.

2. We have that the covariance of D_2 is $\epsilon yy^\top - \epsilon^2 yy^\top = \epsilon(1-\epsilon)yy^\top \preceq \epsilon yy^\top$, which is upper bounded by 1 in spectral norm if $\|y\|_2 \leq 1/\sqrt{\epsilon}$. Since $\|y\|_2 = \sqrt{k}\alpha$, we want that $\alpha \leq 1/\sqrt{k\epsilon}$.

3. Finally, let $X \sim D_2$. For every $i > k$, we have that $\mathbb{E}[(X_i - \mathbb{E}[X_i])^4] = 0$. For $i \in [k]$, $\mathbb{E}[(X_i - \mathbb{E}[X_i])^4] = \mathbb{E}[(X_i - \epsilon\alpha)^4] \leq 8(\mathbb{E}[X_i^4 + \epsilon^4\alpha^4]) = 8(\epsilon\alpha^4 + \epsilon^4\alpha^4) \leq 16\epsilon\alpha^4$, which is less than 16, if $\alpha \leq \epsilon^{-1/4}$.

Thus, the above construction goes through as long as $\alpha \leq \min(1/\sqrt{k\epsilon}, \epsilon^{-1/4})$. When $k \geq 1/\sqrt{\epsilon}$, it suffices that $\alpha = 1/\sqrt{k\epsilon}$. Finally, we note that the difference in means of D_1 and D_2 is $\epsilon\|y\|_2 = \epsilon\sqrt{k}\alpha = \sqrt{\epsilon}$ for the chosen value of α. $\square$

5 ROBUST LINEAR REGRESSION

किसी को घर से निकलते ही मिल गई मंज़िल
कोई हमारी तरह उम्र भर सफ़र में रहा

— अहमद फ़राज़

We study the problem of linear regression where both covariates and responses are potentially (i) heavy-tailed and (ii) adversarially contaminated. Several computationally efficient estimators have been proposed for the simpler setting where the covariates are sub-Gaussian and uncontaminated; however, these estimators may fail when the covariates are either heavy-tailed or contain outliers. In this work, we show how to modify the Huber regression, least trimmed squares, and least absolute deviation estimators to obtain estimators which are simultaneously computationally and statistically efficient in the stronger contamination model. Our approach is quite simple, and consists of applying a filtering algorithm to the covariates, and then applying the classical robust regression estimators to the remaining data. We show that the Huber regression estimator achieves near-optimal error rates in this setting, whereas the least trimmed squares and least absolute deviation estimators can be made to achieve near-optimal error after applying a postprocessing step.

5.1 Introduction

Robust linear regression is a well-studied topic in statistics, both from the viewpoint of theory and practice [HR09; HRRS11; MMYS19]. It has long been observed that the introduction of even a handful of outliers can massively affect the quality of a regression estimator; furthermore, high-leverage points, which are outlying in terms of their covariate values, have the potential for even more drastic consequences. Various methods have been proposed to alleviate the effect of outliers in the data, including

diagnostic tests which focus on identifying and removing outliers [CW82]. On the other hand, such methods are mostly heuristic and few theoretical results exist in this area.

Much classical work in robust linear regression focuses on developing and analyzing estimators that are applied aggregately to an entire data set and are relatively insensitive to certain types of perturbations in the data. These estimators include different families of M-estimators [Hub73], GM-estimators [Mal75], S-estimators [RY84], and MM-estimators [Yoh87]. Notably, most of the corresponding statistical theory has focused on analyzing i.i.d. data, often assumed to be drawn from a mixture distribution involving the parametric model and a (possibly heavy-tailed) contaminating distribution. Recent years have seen a flurry of activity on the somewhat different topic of adversarial contamination—spurred by advances in the theoretical computer science community and motivated by modern machine learning applications—and several approaches have subsequently been proposed for estimating the mean of a multivariate distribution [DK19]. An interesting question which has remained largely unaddressed is whether simpler and seemingly more straightforward approaches such as M-estimation can be proven to achieve similar error guarantees as the more complicated proposals which have emerged from this line of work.

On the topic of M-estimation, Sasai and Fujisawa [SF20] recently derived bounds for linear regression with a Huber loss when adversarial contamination may be present in the response variables. Slightly earlier analysis from [BJK15; BJKK17] provided guarantees for the popular least trimmed squares estimator [Rou84] with adversarially contaminated responses. In contrast, no analogous error bounds have been furnished for the behavior of these or other estimators when the covariates are adversarially contaminated. Rather, a series of classical results on the low breakdown point of regression estimators [Dav93] established the rather pessimistic message that adversarially contaminating even a single data point in both covariates and responses may have an unbounded

effect on the accuracy of a convex M-estimators such as the Huber or least absolute deviation regression estimators (see, e.g., the book [MMYS19] and the references cited therein). Of course, the difficulty in using nonconvex loss functions is that nontrivial challenges arise in optimization.

We note, however, that the failure of simple M-estimation assumes that all the points are included in the estimation procedure, whereas a grossly outlying point might easily be flagged before fitting a moderately robust estimator on the remaining data. In Huber's textbook [HR09, p. 152], we find the following comment: "Undoubtedly, a typical cause for breakdown in regression are gross outliers in the carrier X. In the robustness literature, the problem of leverage points and groups has therefore been tackled by so-called high breakdown point regression.... I doubt that this is the proper approach.... In my opinion, if there are sizable minority components, the task of the statistician is not to suppress them, but to disentangle them." However, the literature on how to perform outlier removal in a theoretically rigorous manner is fairly sparse.

Regarding heavy-tailed distributions, the ordinary least squares estimator may be shown to be highly suboptimal when the additive errors are allowed to be heavy-tailed (cf. Proposition C.5.2 in the appendix). Concretely, in a setting with p parameters, n data points, and noise variance σ^2, the ℓ_2-error of the ordinary least squares estimator may increase as $\Theta\left(\sigma\sqrt{\frac{p}{n\tau}}\right)$ with probability τ—in contrast to the error bound $O\left(\sigma\sqrt{\frac{p}{n}} + \sigma\sqrt{\frac{\log(1/\tau)}{n}}\right)$, which may be achieved under sub-Gaussian distributional assumptions. Starting from the seminal work of Catoni [Cat12], the topic of heavy-tailed estimation has been an active area of research in theoretical statistics in recent years [Men15; MZ20; LL20; Hop20; LM19a; DL22b; HS16], and for regression, Lugosi and Mendelson [LM19c; LM19a] introduced an estimator based on a median-of-means algorithm which achieves the sub-Gaussian error rate even in heavy-tailed scenarios, provided $n = \Omega(p)$. On the other hand, the proposed estimator has runtime exponential in the

dimension, hence is not computationally feasible for large p. More recently, Cherapanamjeri et al. [CHKRT20] proposed a polynomial-time estimator with the desired error rate when $n = \tilde{\Omega}\left(p\sqrt{\log(1/\tau)}\right)$. However, the estimator requires the covariates to satisfy a stronger condition: a sum-of-squares (SOS) certifiable proof of degree 8. The proposed algorithm involves solving a large semidefinite program which, although achievable in polynomial time, is not very practical.

In this paper, we take a cue from the literature on robust mean estimation under adversarial contamination, in which the proposed algorithms implicitly involve a filtration or screening step to identify and remove outlying data points, after which a (weighted) empirical mean is computed on the remaining data [LRV16; DKKLMS16] (cf. Appendix C.1.1). The success of such algorithms stems from a useful lemma which states that when the distribution of the uncontaminated data is isotropic, the empirical mean of a set of data points which have an approximately isotropic empirical covariance matrix will be close to the true mean. The filtering mechanism consequently operates by iteratively removing data points until the remaining set is approximately isotropic—theoretically, one can show that the proposed filters do not remove too many uncontaminated data points, while removing any adversarially introduced outliers that move the sample mean sufficiently far from the true mean. A key insight of this paper is that the condition of approximate isotropy of the empirical covariance (also known as stability) is in fact a sufficient condition for the success of classical robust regression estimators such as the Huber M-estimator, least trimmed squares (LTS), and least absolute deviation (LAD) estimator. Thus, an adversarially contaminated data set may first be preprocessed by applying a filter to the covariates, and then the classical estimator may be applied to the remaining data to obtain an overall estimate close to the true regression vector. A careful analysis shows that this method can be applied to data sets which possess adversarial contamination in *both* the covariates and responses. Furthermore,

the same method can be used to obtain error guarantees for heavy-tailed covariates and/or responses. Perhaps it is unsurprising that both adversarial contamination and heavy-tailed distributions may be treated using similar estimators, since in the latter case, "outlying" points may be seen as occurring due to randomness naturally present in the sample rather than having been introduced adversarially.

In concurrent work, Zhu et al. [ZJS22b] and Bakshi and Prasad [BP21] studied computationally-efficient algorithms for heavy-tailed robust regression in a more general setting, where the covariance Σ of the covariates is unknown (but bounded) and the noise may not be independent, with the goal of minimal dependence on the level of adversarial contamination ϵ. Initiated by Klivans et al. [KKM18], their algorithms are based in a sum-of-squares framework, and impose a *certifiable* hypercontractivity assumption on covariates, which is a somewhat more restrictive than our assumption of hypercontractivity [KSS18]. As the goal in these works is slightly different, the resulting estimators have suboptimal dependence on sample complexity and probability of error in comparison to ours.

Recently, Cherapanamjeri et al. [CATJFB20] and Depersin [Dep20a] considered covariates with bounded fourth moments, and proposed an iterative gradient based procedure for robust regression. When Σ is unknown (but bounded) and the noise is independent, Cherapanamjeri et al. [CATJFB20] obtained a near-linear time estimator (when ϵ is constant) with near-optimal sample complexity, but with a constant error probability. Depersin [Dep20a] studied the case of known Σ and possibly dependent noise, and proposed a computationally efficient estimator with a sub-Gaussian error rate and $O(\sqrt{\epsilon})$ dependence. However, the error guarantee for the estimator does not improve when higher-order moments are bounded.

We emphasize that the focus of our work is slightly different, in that we seek to show that several classical estimators *which are known to be robust to corruptions in the responses*

can also be made robust to corruptions in the covariates after a simple outlier filtration step. For each of the Huber, LAD, and LTS estimators, our guarantees for heavy-tailed covariates (nearly) match their corresponding known results for sub-Gaussian covariates. In addition, we highlight the fact that our filtered Huber estimator (cf. Theorem 5.3.6) obtains the tightest rates and is near-optimal in multiple parameters simultaneously, among all known polynomial-time estimators, for the case of isotropic covariates and independent noise. For a more extensive discussion of related work, see Appendix C.3.

The rest of the paper is organized as follows: In Section 5.2, we explain the problem setup and connection with robust mean estimation. In Section 5.3, we analyze the Huber regression estimator. We prove our results regarding the LTS and LAD estimators in Sections 5.4 and 5.5, respectively. Section 5.6 contains the details regarding a postprocessing step which can be used to improve the accuracy of the LTS and LAD estimators. Finally, Section 5.7 contains simulation results reporting the effect of the proposed filtering step. Section 5.8 concludes the paper with a short discussion of open questions. Pseudocode for the algorithms mentioned in the paper is contained in Appendix C.1.

5.2 Background and Problem Setup

For a list of notation and some basic definitions, see Appendix C.2.

5.2.1 Linear Model

Suppose we have observations drawn from the linear model

$$y_i = x_i^\top \beta^* + z_i, \qquad 1 \le i \le n, \tag{5.1}$$

where $\beta^* \in \mathbb{R}^p$, the x_i's are sampled i.i.d. from a distribution over $\mathbb{R}^p$, and the z_i's are i.i.d. noise. We also write equation (5.1) as $y = X\beta^* + z$, where $y, z \in \mathbb{R}^n$, $\beta^* \in \mathbb{R}^p$, and

$X \in \mathbb{R}^{n \times p}$. Our goal is to estimate β^* from the data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. We make the following assumption about the distribution of the covariates:

Assumption 5.2.1. *The covariates satisfy $\mathbb{E}\, x_i = 0$ and $\mathbb{E}\, x_i x_i^\top = I$. Moreover, the covariates satisfy $(4, 2)$-hypercontractivity with parameter $\sigma_{x,4} \leq C$, for a known constant C.*

Note that the case of a known, non-identity covariance matrix can be reduced to the setting of identity covariance via a linear transformation. We relax the condition of an identity covariance matrix to an unknown but *bounded* covariance matrix in Section 5.3.4.

We assume an identity covariance structure in Assumption 5.2.1 because of the statistical query (SQ) lower bound from Diakonikolas et al. [DKS19], stating that in the case of an unknown (but bounded) covariance matrix, any computationally efficient SQ algorithm requires approximately $\Omega(p^2)$ samples to achieve an error rate of $o(\sqrt{\epsilon})$ in the strong contamination model (cf. Theorem 5.3.6). This suggests that the knowledge of the covariance matrix is needed to obtain improved statistical rates using computationally-efficient algorithms. We show that the filtered Huber estimator achieves the rate $O(\sqrt{\epsilon})$ in the unknown covariance setting in Section 5.3.4. However, even with an identity covariance matrix, the covariates could have a degenerate distribution such that, with high probability, all the sampled points have norm 0 and all information about β^* would be lost. As a result, we also include the hypercontractivity condition in Assumption 5.2.1, which is a standard assumption in this field. Note that under the identity covariance assumption, the hypercontractivity condition can simply be written as $(\mathbb{E}(v^\top x_i)^4)^{1/4} \leq C$, i.e., bounded fourth moments.

Remark 5.2.2. *Note that the assumption that an upper bound C on the hypercontractivity constant $\sigma_{x,4}$ is known is necessary for running the algorithms in this paper in practice (e.g., Algorithms 11, 13, and 14 below), since our theory requires the filtering parameter ϵ' to be smaller than some value which depends on C.*

We also make the following assumption about the additive noise distribution:

Assumption 5.2.3. *The noise $\{z_i\}$ is independent of the covariates $\{x_i\}$, and $\mathbb{E}\, z_i = 0$.*

The independence assumption on the z_i's and x_i's is somewhat restrictive. We will relax this assumption on noise for a subset of our results: (i) Theorems 5.3.1 and 5.3.4 hold even if the first moment of the z_i's is infinite, and (ii) Theorem 5.5.1 holds even if the z_i's are dependent on x_i's and have nonzero mean.

In the sequel, we also study the robustness of our estimators when a fraction of data points are adversarially contaminated.

Definition 5.2.4. (*Strong Contamination Model*) *We say that a set T is an ϵ-*corrupted version of a set S *if $|T| = |S|$ and $|T \cap S| \geq (1-\epsilon)|S|$.*

This contamination model is called the *strong contamination model* in the literature, since no computational or statistical restrictions are imposed on T. In contrast, Huber's ϵ-contamination model requires the contamination mechanism to be oblivious and additive, i.e., it can only add outliers to the uncontaminated i.i.d. data without looking at the inliers.

5.2.2 Stability Conditions

Our technical results will rely on appropriately defined notions of stability. Recall the following stability condition from the robust mean estimation literature [DKKLMS16; DKKLMS17; SCV18; DHL19; DK19; CDG19; CDGS20]:

Definition 5.2.5. (*Strong stability*) *For $\epsilon < 1/2$, we say that a multiset $S = \{x_1, \ldots, x_n\}$* satisfies (ϵ, δ)-stability *for $\epsilon \leq \delta$ with respect to μ and σ^2 if for all $S' \subseteq S$ such that $|S'| \geq (1-\epsilon)n$, we have*

$$\left\| \frac{1}{|S'|} \sum_{i \in S'} x_i - \mu \right\|_2 \leq \sigma\delta, \quad \text{and} \quad \left\| \frac{1}{|S'|} \sum_{i \in S'} (x_i - \mu)(x_i - \mu)^\top - \sigma^2 I \right\|_2 \leq \frac{\sigma^2\delta^2}{\epsilon}.$$

Definition 5.2.5 is designed for samples from a distribution with mean μ and covariance $\Sigma \preceq \sigma^2 I$. Note that a set which is (ϵ, δ)-stable is also (ϵ', δ')-stable for any $\epsilon' \leq \epsilon$ and $\delta' \geq \delta$. The (ϵ, δ)-stability condition states that for every large enough subset, (i) the ℓ_2-distance between the empirical mean and μ is at most $\sigma\delta$, and (ii) the spectral distance between the (centered) second moment matrix and $\sigma^2 I$ is at most $\frac{\sigma^2\delta^2}{\epsilon}$. Since our primary focus will be on distributions with $\mu = 0$ and $\sigma^2 = 1$, we will not explicitly state these parameters when they are clear from context. In Appendix C.1.1, we review the iterative filtering algorithm, guaranteed to succeed under strong stability, which is a key building block for our work.

Next, we mention a deterministic condition on the covariates that appeared in the analysis of least trimmed squares regression in Bhatia et al. [BJK15]:

Definition 5.2.6. (*Weak stability*) *Let $\epsilon \in (0, 1)$. The set $\{x_1, \ldots, x_n\}$* satisfies (ϵ, L, U)-weak stability *if for every subset $S \subseteq [n]$ such that $|S| \geq (1 - \epsilon)n$, the second moment matrix of S is approximately isotropic, i.e., $L \leq \lambda_{\min}\left(\frac{1}{n}\sum_{i\in S} x_i x_i^\top\right) \leq \lambda_{\max}\left(\frac{1}{n}\sum_{i\in S} x_i x_i^\top\right) \leq U$.*

Bhatia et al. [BJK15] established the convergence of an alternating minimization algorithm under weak stability for a fixed ϵ, provided (i) $L = \Theta(1)$ and (ii) $U = \Theta(1)$. We will show in Section 5.3 that under the same conditions, Huber regression also succeeds with high probability. This leads to the question of whether weak stability holds with high probability for heavy-tailed covariates; following arguments in Koltchinskii and Mendelson [KM15], it can be shown that condition (i) holds with high probability [DKP20]. However, known concentration results suggest that condition (ii) does *not* hold with high probability for heavy-tailed covariates when $S = [n]$: The usual matrix Chernoff bounds [Tro15] yield $U = O(1)$ with probability $1 - \tau$ if $n = \Omega(p \log(1/\tau))$, which may be much larger than the ideal sub-Gaussian sample complexity which is *additive* rather than multiplicative in p and $\log(1/\tau)$. Bhatia et al. [BJK15] also defined the following notions in their analysis of LTS:

Definition 5.2.7. (*SSC and SSS*) *Let $x_1, \ldots, x_n$ be n points in $\mathbb{R}^p$. For $m \in [n]$, we say that the x_i's satisfy the* Subset Strong Convexity (SSC) property at level m with parameter λ_m *if* $\lambda_m \leq \min_{S \subseteq [n]:|S|=m} \lambda_{\min}\left(\sum_{i \in S} x_i x_i^\top\right)$. *The x_i's satisfy the* Subset Strong Smoothness (SSS) property at level m with parameter Λ_m *if* $\max_{S \subseteq [n]:|S|=m} \lambda_{\max}\left(\sum_{i \in S} x_i x_i^\top\right) \leq \Lambda_m$.

Note that if a set satisfies (ϵ, L, U)-weak stability, then it satisfies the SSC and SSS properties at level $(1-\epsilon)n$ with parameters nL and nU, respectively. However, the results of Bhatia et al. (cf. Lemma C.8.1 below) require finer control of the minimum and maximum eigenvalues at different levels, in addition to the assumption of weak stability. Proposition C.6.1 in Appendix C.6 shows that strong stability implies weak stability.

Our final notion of stability comes from Karmalkar and Price [KP19]:

Definition 5.2.8. (ℓ_1-*stability*) *We say a set of data points $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^p$ satisfies* (m, M, ϵ, ℓ_1)-stability *if for all subsets $S \subseteq [n]$ with $|S| \geq (1-\epsilon)n$ and all unit vectors $v \in \mathbb{R}^p$, we have $\frac{1}{n}\sum_{i \in S} |x_i^\top v| \geq M$ and $\frac{1}{n}\sum_{i \in [n] \setminus S} |x_i^\top v| \leq m$.*

Note that this definition of stability controls the ℓ_1-norm of projections, whereas weak stability (or strong stability) is a statement about ℓ_2-norms. This notion of stability was used by Karmalkar and Price [KP19] in their analysis of the LAD estimator and will also be used in our analysis of the LAD estimator to follow. As shown later (cf. Lemma C.6.5), the upper bound in the definition of ℓ_1-stability can be derived directly from strong stability.

5.3 Huber Regression

Recall that the Huber loss with parameter γ is defined as $\ell_\gamma(x) = \frac{x^2}{2}$, if $|x| \leq \gamma$, and $\ell_\gamma(x) = \gamma|x| - \frac{\gamma^2}{2}$ otherwise [Hub64; HR09]. Let $\psi_\gamma(x) = \nabla \ell_\gamma(x)$. We now define $\mathcal{L}_\gamma(\beta) := \frac{1}{n}\sum_{i \in [n]} \ell_\gamma(y_i - x_i^\top \beta)$ and let Huber's M-estimator be defined as $\widehat{\beta}_{H,\gamma} = \operatorname{argmin}_\beta \mathcal{L}_\gamma(\beta)$.

Note that the Huber objective function is convex, so it is possible to (approximately) obtain the minimizer $\widehat{\beta}_{H,\gamma}$ in a computationally feasible manner. Thus, we will begin by analyzing statistical properties of the Huber regression estimator and then comment only briefly on optimization (cf. Section 5.3.5). We present our statistical analysis in increasing levels of complexity: fixed design covariates satisfying weak stability and i.i.d. symmetric noise (Section 5.3.1), random i.i.d. covariates and asymmetric noise (Section 5.3.2), and adversarially contaminated data (Section 5.3.3).

5.3.1 Fixed Design and Symmetric Noise

Our main result in this subsection is the following:

Theorem 5.3.1. *Suppose we have n i.i.d. samples from the following (fixed design) model: $y_i = x_i^\top \beta^* + z_i$, where the covariates $\{x_i\}$ satisfy weak stability with some ϵ, L, and U. Suppose the errors $\{z_i\}$ are sampled independently from a symmetric distribution. Let $\widehat{\beta}_{H,\gamma} \in \arg\min \mathcal{L}_\gamma(\beta)$. Let τ be such that $\frac{\log(1/\tau)}{n} = O(\epsilon)$. Then setting γ such that $\mathbb{P}(|z_i| \geq \gamma/2) = O(\epsilon)$, we have, with probability at least $1 - \tau$,*

$$\|\widehat{\beta}_{H,\gamma} - \beta^*\|_2 \lesssim \frac{\gamma\sqrt{U}}{L}\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}}\right), \quad \text{as long as } n = \Omega\left(\frac{U^2(p + \log(1/\tau))}{L^2\epsilon^2}\right).$$

Furthermore, $\mathcal{L}_\gamma(\beta)$ is L-strongly convex in a ball of radius $\Omega(\epsilon\gamma/\sqrt{U})$ around $\widehat{\beta}_{H,\gamma}$.

Theorem 5.3.1 provides an error bound on the Huber regression estimator under a deterministic condition on the covariates; the probabilistic nature of the theorem comes from the randomness in the additive errors. In Theorems 5.3.4 and 5.3.6 below, we will show that the weak stability condition holds with high probability when the covariates are drawn from possibly heavy-tailed, possibly contaminated distributions and then passed through a filtering algorithm. We will also show how to relax the assumption that the distribution of z_i is symmetric via an appropriate preprocessing step.

Remark 5.3.2. *When $\Omega(1) = L \leq U = O(1)$ and $\epsilon = \Omega(1)$, the sample complexity reduces to $n = \Omega(p)$ (by assumption, $n = \Omega(\log(1/\tau))$). Also, the radius of strong convexity is $\Omega(\gamma)$.*

Remark 5.3.3. *Note that Theorem 5.3.1 does not require the additive noise to have finite moments. If the noise distribution has a finite k^{th} moment, however, Markov's inequality implies that we can always set $\gamma = \Omega(\epsilon^{-1/k}(\mathbb{E}\,|z_i|^k)^{1/k})$. In particular, if the z_i's have a finite variance σ^2, we can take $\gamma = \Omega(\sigma/\sqrt{\epsilon})$.*

The assumption that $\mathbb{P}(|z_i| \geq \gamma/2) = O(\epsilon)$ implies that the parameter γ used to define the Huber loss needs to be sufficiently large in order for our theory to succeed, in a sense being calibrated to the tail behavior of the error distribution. Indeed, the heavier the tails of the z_i's, the larger γ would need to be, leading to a worse error bound. Since it is generally unreasonable to assume that the scale of the additive noise distribution is known in practice, we will discuss methods for adaptively choosing γ from the data in our results below. The proof of Theorem 5.3.1 is provided in Appendix C.7.2.

5.3.2 Generalization to Random Design and Asymmetric Noise

We now generalize the result of the previous section to the random design model with asymmetric noise. We proceed by reducing the case of asymmetric noise to symmetric noise: we will randomly subtract two points so that the additive noise in the new linear model has symmetric noise. Next, we will show that the iterative filtering algorithm from Diakonikolas et al. [DK19; DKKLMS16] (Theorem C.1.1) can be used to obtain a large subset of data points for which the covariates satisfy weak stability. We will then invoke Theorem 5.3.1.

Theorem 5.3.4. *Suppose we have $2n$ i.i.d. samples $\{(x_i, y_i)\}_{i=1}^{2n}$ from the following (random-design) model: $y_i = x_i^\top \beta^* + z_i$, where the covariates satisfy Assumption 5.2.1 and the noise distribution satisfies Assumption 5.2.3. Let τ be such that $\frac{\log(1/\tau)}{n} = O(1)$. Suppose γ is such*

that $\mathbb{P}\left(|z_1 - z_2| \geq \frac{\gamma}{\sqrt{2}}\right) \leq c^*$ *for a small enough constant* $c^* > 0$*, and suppose* ϵ' *is equal to a sufficiently small constant. Then running Algorithm 11 with parameters* γ *and* ϵ' *produces an estimator that, with probability at least* $1 - 2\tau$*, satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \gamma\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}}\right), \quad \text{as long as } n = \Omega(p \log p).$$

On the same event, the loss function is $\Omega(1)$*-strongly convex in a radius of* $\Omega(\gamma)$ *around* $\widehat{\beta}$*.*

The proof of Theorem 5.3.4 is contained in Appendix C.7.3. In Appendix C.7.1.1, we provide a rigorous method for estimating an appropriate tuning parameter γ from the data.

Remark 5.3.5. *Similar to Remark 5.3.3, if the* k^{th} *moment of the noise distribution is finite, we can set* $\gamma = \Omega((\mathbb{E}\,|z_1 - z_2|^k)^{1/k})$*, for any positive* k*.*

5.3.3 Adversarial Corruption

We will now consider adversarial corruption in both covariates and responses. Let S be the set of n i.i.d. samples and let T be an ϵ-corrupted version of S in the sense of Definition 5.2.4. One might expect Algorithm 11 to be robust to adversarial contamination, as Huber regression has been shown to be robust against corruption in responses [SF20] and the filtering step can handle corruptions in covariates. In this section, we will crucially use the strong stability condition, and not just weak stability, to obtain tighter control on deviations. In fact, the following result shows that Huber regression also achieves near-optimal statistical guarantees in the adversarial setting with a slightly different choice of parameters:

Theorem 5.3.6. *Let* $S = \{(x_i, y_i)\}_{i=1}^{2n}$ *be a set of i.i.d. samples drawn according to the same distributional assumptions as in Theorem 5.3.4. Further suppose that the covariates satisfy*

$(k, 2)$-hypercontractivity with parameter $\sigma_{x,k} = O(1)$, for some $k \geq 4$. Let T be an ϵ-corrupted version of S. Suppose γ is such that $\mathbb{P}\left(|z_1 - z_2| \geq \frac{\gamma}{\sqrt{2}}\right) \leq c^*$ for a small enough constant $c^* > 0$. Then running Algorithm 11 on the set T with parameters $\epsilon' = \Theta\left(\epsilon + \frac{\log(1/\tau)}{n}\right)$ produces an estimator that, with probability at least $1 - \tau$, satisfies

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \gamma \left(\sqrt{\frac{p \log p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \epsilon^{1-1/k} \right),$$

provided $n = \Omega(p \log p + \log(1/\tau))$ and ϵ is less than a sufficiently small constant. Moreover, on the same event, the loss function is $\Omega(1)$-strongly convex in a radius of $\Omega(\gamma)$ around $\widehat{\beta}$.

The proof of Theorem 5.3.6 is provided in Appendix C.7.4. For a discussion of how to tune the Huber parameter, see Appendix C.7.1.2.

Remark 5.3.7. *In order to run Algorithm 11 with the theoretical choice of ϵ' in Theorem 5.3.6, we assume knowledge of the level of adversarial contamination. On the other hand, note that if T is an ϵ_1-corrupted version of S, then T is also an ϵ_2-corrupted version of S, for any $\epsilon_1 \leq \epsilon_2$. Thus, knowledge of an upper bound on the contamination level suffices. (The same remark applies to Theorems 5.4.2 and 5.5.1, and Theorems 5.6.1 and 5.6.2 below.)*

Briefly, our proof strategy is similar to the proof of Theorem 5.3.1: Although the covariates and noise are not necessarily independent on the filtered set, we can establish modified versions of the structural Lemmas C.7.2 and C.7.3. In particular, we crucially use the stability property of the filtered set, which is stronger than the assumption of weak stability.

Remark 5.3.8. *We also note that Algorithm 11 has another favorable property when only the covariates are corrupted: Suppose $\{x_i\}_{i=1}^n$ and $\{z_i\}_{i=1}^n$ are generated from distributions satisfying Assumptions 5.2.1 and 5.2.3, respectively. Instead of observing $(X, X\beta^* + z)$, the statistician observes $(\tilde{X}, \tilde{y})$, where $\tilde{y} = \tilde{X}\beta^* + z$, and $\tilde{X}$ matches X in all but ϵn rows and is independent of*

z. Then as long as ϵ is smaller than a fixed constant, the error guarantee of Theorem 5.3.6 would be of the form $O\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}}\right)$ and is independent of ϵ. Since $\tilde{X}$ and $\tilde{y}$ still follow a linear relationship and independence is maintained between the errors and covariates, the setting is essentially reduced to that of Theorem 5.3.1.

Remark 5.3.9. *Finally, we mention a slightly stronger guarantee for Algorithm 11 for Gaussian covariates, i.e., $X \sim \mathcal{N}(0, I)$. As can be seen in Appendix C.7.5 in the proof of Theorem 5.3.6, we could instead obtain the error bound $O\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \epsilon\sqrt{\log(1/\epsilon)}\right)$. This is because a set of n i.i.d. samples from $\mathcal{N}(0, I)$ is (ϵ, δ)-stable with probability $1 - \tau$, where $\delta \lesssim \sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \epsilon\sqrt{\log(1/\epsilon)}$ [DKKLMS16; Li18; DKKLMS17]. Sub-Gaussian distributions with identity covariance and sub-Gaussian norm $O(1)$ also achieve this rate.*

5.3.4 Generalization to Unknown Covariance

We now discuss the case where the covariates have an unknown but bounded covariance matrix. We replace Assumption 5.2.1 with the following assumption:

Assumption 5.3.10. *The covariates satisfy $\mathbb{E}\, x_i = 0$ and $\kappa_l I \preceq \mathbb{E}\, x_i x_i^\top \preceq \kappa_u I$ for some $\kappa_l \in (0, 1)$ and $\kappa_u \geq 1$. (For simplicity, we will assume that $\kappa_l = 1/2$ and $\kappa_u = 2$ in our arguments, but similar results hold as long as $\kappa_u = \Theta(\kappa_l)$.) Moreover, the covariates satisfy $(4, 2)$-hypercontractivity with parameter $\sigma_{x,4} \leq C$, for a known constant C.*

We are able to generalize our result from Theorem 5.3.6 to the setting under Assumption 5.3.10.

Theorem 5.3.11. *Suppose we have $2n$ i.i.d. samples $\{(x_i, y_i)\}_{i=1}^{2n}$ from the following (random-design) model: $y_i = x_i^\top \beta^* + z_i$, where the covariates satisfy Assumption 5.3.10 and the noise distribution satisfies Assumption 5.2.3. Let τ be such that $\frac{\log(1/\tau)}{n} = O(1)$. Suppose γ is such that $\mathbb{P}\left(|z_1 - z_2| \geq \frac{\gamma}{\sqrt{2}}\right) \leq c^*$ for a small enough constant $c^* > 0$, and suppose ϵ' is equal to a sufficiently small constant. Let T be an ϵ-corrupted version of S. Then running Algorithm 11 on*

the set T with parameters $\epsilon' = \Theta\left(\epsilon + \frac{\log(1/\tau)}{n}\right)$ and $\gamma = \Omega(\sigma)$ produces an estimator that, with probability at least $1 - \tau$, satisfies

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \gamma \left(\sqrt{\frac{p \log p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \sqrt{\epsilon}\right),$$

provided $n = \Omega(p \log p + \log(1/\tau))$ and ϵ is less than a sufficiently small constant. Moreover, on the same event, the loss function is $\Omega(1)$-strongly convex in a radius of $\Omega(\gamma)$ around $\widehat{\beta}$.

The proof of Theorem 5.3.11 is given in Appendix C.7.6, and follows the same strategy as Theorem 5.3.6, by noting that Huber regression primarily relies on (ϵ, L, U)-weak stability, where $\epsilon = \Omega(1), L = \Omega(1)$, and $U = O(1)$. The first two conditions hold by the small ball property, and the guarantee of the filter algorithm in the unknown covariance case is strong enough to ensure the third condition [DKP20]. However, these algorithms do not adapt to higher moments of the data in the unknown covariance setting. This drawback is reflected in the worse dependence on ϵ, i.e., $O(\sqrt{\epsilon})$ instead of $O(\epsilon^{3/4})$ under $(4, 2)$-hypercontractivity. Note that the SQ lower bound of Diakonikolas et al. [DKS19] suggests that this $O(\sqrt{\epsilon})$ dependence is essentially optimal for computationally-efficient algorithms when $n = o(p^2)$ even when the covariates are Gaussian (with an unknown covariance).

Remark 5.3.12. *In the absence of adversarial contamination, we can follow the same strategy as in Theorem 5.3.4: Under Assumption 5.3.10, run the filter algorithm with ϵ' equal to a small enough constant (independent of τ) to obtain a sub-Gaussian tail in the error guarantee.*

5.3.5 Optimization

As noted above, the Huber objective function $\mathcal{L}_\gamma(\beta)$ is convex in β, so optimization should in principle be easy. Taking a closer look, we see that as established in Theorems 5.3.4 and 5.3.6, the loss function is *strongly* convex in a ball of sufficiently large enough radius

$\Omega(\gamma)$ around $\widehat{\beta}$. Therefore, running gradient descent yields linear convergence if the initialization is inside that ball [Bub15]. Considering the case when we set the Huber parameter to be $\gamma = \Theta(\sigma)$, our theory shows that we can guarantee such an initialization using the LAD estimator (cf. Theorem 5.5.1) or LTS estimator (cf. Theorem 5.4.2).

Instead of using a different robust regression estimator for a warm start, we can directly apply the ellipsoid algorithm to the Huber loss. However, running the ellipsoid algorithm might be undesirable, as its running time, although polynomial, is practically slow [Bub15].

5.4 Least Trimmed Squares Estimator

In this section, we study the least trimmed squares (LTS) estimator [Rou84]:

$$\widehat{\beta}_{LS,m} = \operatorname*{argmin}_{\beta} \min_{S \subseteq n : |S| = n-m} \sum_{i \in S} (y_i - x_i^\top \beta)^2, \tag{5.2}$$

where m is the trimming parameter. We will establish conditions under which $\|\widehat{\beta}_{LS,m} - \beta^*\|_2$ is small, with very high probability.

Unlike the Huber regression estimator, a significant drawback of the LTS estimator is that the objective function (5.2) is nonconvex. Nonetheless, various methods have been developed to efficiently obtain a local optimum of the LTS objective function, which have been shown to perform well empirically [RV06]. In recent work, Bhatia et al. [BJK15; BJKK17] proved that under sufficiently nice assumptions on the covariates, the alternating minimization algorithm (Algorithm 12) succeeds in finding a good candidate solution. Here, $P_X = X(X^\top X)^{-1}X^\top$ denotes the hat matrix, and the function HT_m is defined as follows:

Definition 5.4.1. *For any $v \in \mathbb{R}^n$ and $m \in [n]$, let $S_{m,v} \subseteq [n]$ be the set of cardinality of m such that for any $i \in S_{m,v}$ and $j \in [n] \setminus S_{m,v}$, we have $|v_i| \geq |v_j|$. To ensure uniqueness, we*

choose the smaller indices if ties occur. The m-hard thresholding operator *is the function* $\mathrm{HT}_m : \mathbb{R}^n \to \mathbb{R}^n$ *defined as follows: For any* $v \in \mathbb{R}^n$*, we have* $(\mathrm{HT}_m(v))_i = v_i$*, if* $i \in S_{m,v}$*, and* $(\mathrm{HT}_m(v))_i = 0$ *otherwise.*

In other words, the set $S_{m,v}$ identifies the indices of the largest m coordinates of v in magnitude, and the HT_m function returns a vector that preserves only these top m components. Note that Algorithm 12 is derived by recasting the optimization problem (5.2) as $\min_{\beta \in \mathbb{R}^p, \|b\|_0 \leq m} \|X\beta - (y - b)\|_2^2$ and alternately minimizing over β and b, where we explicitly solve for β on each iteration (see Bhatia et al. [BJKK17] for more details).

The proof of the following main result is contained in Appendix C.8.1:

Theorem 5.4.2. *Let* $S = \{(x_i, y_i)\}_{i=1}^n$ *be a set of i.i.d. samples drawn according to the same distributional assumptions as in Theorem 5.3.6. Let* $T = \{(x_i', y_i')\}_{i=1}^n$ *be an* ϵ*-corrupted version of* S*, where* ϵ *is less than a sufficiently small constant. Further suppose the errors satisfy* $(k', 2)$*-hypercontractivity with parameter* $\sigma_{z,k'} = O(1)$*, for some* $k' \geq 2$*. Let* τ *be such that* $\frac{\log(1/\tau)}{n} = O(1)$*. With probability at least* $1 - O(\tau)$*, running Algorithm 13 on the set* T *with parameters* $m = \Theta\left(p \log p + \epsilon n + \log\left(\frac{1}{\tau}\right)\right)$ *and* $\epsilon' = \Theta\left(\frac{m}{n}\right)$ *yields an estimator* $\widehat{\beta}$ *satisfying*

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \sigma \left(\sigma_{z,k'} \left(\frac{p \log p}{n} + \epsilon + \frac{\log(1/\tau)}{n} \right)^{1/2 - 1/k'} \right),$$

provided $n = \Omega(p \log p)$ *and* $J \gtrsim \log_2\left(\frac{\|y'\|_2 + \|X'\|_2 \|\beta^*\|_2}{\alpha}\right)$*, where* α *is defined to be the error bound given above. If we further suppose that the errors satisfy* $(4, 2)$*-hypercontractivity with* $\sigma_{z,4} = O(1)$*, then* $J \gtrsim \log_2\left(\frac{\|y'\|_2 (1 + \|X'\|_2)}{\alpha}\right)$ *iterations suffice.*

Remark 5.4.3. *The error guarantee of the LTS estimator in Theorem 5.4.2 is weaker than that of the Huber regression estimator in Theorem 5.3.6. It is not clear whether the suboptimality of the LTS error bound is intrinsic to the LTS estimator or an artifact of our analysis; we leave this question for future work. In the case of sub-Gaussian noise, it can be shown that the guarantee of*

Theorem 5.4.2 matches the guarantee of Bhatia et al. [BJK15; BJKK17] (up to log factors) who assume, in addition, that the covariates are sub-Gaussian.

Remark 5.4.4. *The two statements in Theorem 5.4.2 differ in the number of iterations we require to guarantee that the output of the alternating minimization algorithm will have small ℓ_2-error—in order to obtain a data-driven upper bound on $\|\beta^*\|_2$, we impose additional hypercontractivity assumptions on the noise distribution. As in the case of the Huber estimator (cf. Section 5.3.5), one might choose to use the LAD estimator to warm-start the algorithm and save on computation. Theorem 5.5.1 below guarantees that the LAD estimator satisfies $\|\widehat{\beta}_{LAD} - \beta^*\|_2 = O(\kappa)$ when $\mathbb{E}\,|z_i| = \kappa$; the runtime of Algorithm 13 on the shifted data $(X, y - X^\top \widehat{\beta}_{LAD})$ would then scale with $\|\widehat{\beta}_{LAD} - \beta^*\|_2 = O(\kappa)$ rather than $\|\beta^*\|_2$.*

As shown in the proof of Lemma C.8.1, we can alternatively run Algorithm 13 until $\|b^j - b^{j-1}\|_2 = O(\alpha\sqrt{n})$ to obtain a data-dependent stopping criterion. Indeed, by inequality (C.21), we have $\|b^{j+1} - b^*\| \le e_0 + \frac{1}{2}\|b^j - b^*\|_2$, so by the triangle inequality, we have $\|b^j - b^{j+1}\|_2 \ge \|b^j - b^*\|_2 - \|b^{j+1} - b^*\|_2 \ge \frac{1}{2}\|b^j - b^*\|_2 - e_0$. Thus, if the difference between successive iterates is sufficiently small, the error must be small, as well.

Finally, we emphasize that although the LTS objective function is nonconvex (5.2), our theoretical guarantees are for the output of a particular iterative algorithm which can be performed efficiently. Importantly, we need not assume that the alternating minimization algorithm converges to a global optimum of the LTS objective.

5.5 Least Absolute Deviation

We now study the least absolute deviation estimator $\widehat{\beta}_{LAD} = \operatorname{argmin}_\beta \sum_{i=1}^n |y_i - x_i^\top \beta|$. Note that the LAD estimator is parameter-free. Although the error bounds we will derive have suboptimal error rates compared to the other estimators, the LAD estimator is useful for initialization for tuning or optimizing the Huber estimator (cf. Sections 5.3.2

and 5.3.5), or initializing the alternating minimization algorithm for the LTS estimator (cf. Remark 5.4.4).

Our main result in this section is to show that under our setting, the filtered covariates satisfy the ℓ_1-stability condition of Definition 5.2.8, from which we may derive an error bound according to Lemma C.9.4. The proof is contained in Appendix C.9.2.

Theorem 5.5.1. *Let $S = \{(x_i, y_i)\}_{i=1}^n$ be i.i.d. samples from the linear model $y_i = x_i^\top \beta^* + z_i$, where the covariates satisfy Assumption 5.2.1 and the noise satisfies $\mathbb{E}\,|z_i| = \kappa$. For an $\epsilon < c^*$, let T be an ϵ-corrupted version of S. Let $\widehat{\beta}$ be the output of Algorithm 14 with input T and ϵ', where ϵ' is a small enough constant. Let τ be such that $\frac{\log(1/\tau)}{n} = O(1)$. Then with probability at least $1 - \tau$, we have $\|\widehat{\beta} - \beta^*\|_2 = O(\kappa)$, provided $n = \Omega(p \log p)$.*

Remark 5.5.2. *The guarantees of Theorem 5.5.1 hold under very general conditions. We do not require the noise distribution to have zero mean or be independent of the covariates; all we require is $\mathbb{E}\,|z_i| < \infty$. Furthermore, we can generalize this result to the case of an unknown but bounded covariance of the form $\frac{1}{2} I \preceq \mathbb{E}\, xx^\top \preceq 2I$ (cf. Section 5.3.4), as well.*

5.6 Postprocessing

We now outline a one-step estimator which, given an initial estimator $\widehat{\beta}_1$ such that $\|\widehat{\beta}_1 - \beta^*\|_2 = O(\sigma)$, returns another estimator $\widehat{\beta}_2$ that has sub-Gaussian rates. In the analysis of this section, we will assume that Assumption 5.2.3 is satisfied and the noise variance $\mathbb{E}(z_i^2) = \sigma^2$ is finite. As shown in Sections 5.4 and 5.5, the LTS or LAD estimators will then satisfy the error bound of $O(\sigma)$ with high probability and can be used for $\widehat{\beta}_1$. We note that a similar postprocessing construction has been leveraged in earlier works [BDLS17; DKS19; PSBR20].

We first state a version of the result when $\widehat{\beta}_1$ does not depend on the data. This can always be achieved by splitting the samples when either (i) there is no contamination,

or (ii) the contamination mechanism does not depend on the data, e.g., Huber's contamination model. Recall the median-of-means preprocessing algorithm [LM19a]: Given data $\{x_1, \dots, x_n\}$ and a parameter $k \in [n]$, construct $\{z_1, \dots, z_k\}$ by randomly dividing $\{x_1, \dots, x_n\}$ into k disjoint buckets of equal size (if k does not divide n, then remove some samples), and let $\{z_1, \dots, z_k\}$ be the empirical means of the points in the buckets. We have the following theorem:

Theorem 5.6.1. *Let S be a set of n i.i.d. samples from the linear model $y_i = x_i^\top \beta^* + z_i$, where the covariates satisfy Assumption 5.2.1 and the noise distribution satisfies Assumption 5.2.3. Suppose $\mathbb{E}(z_i^2) = \sigma^2$. Let $\widehat{\beta}_1$ be an estimator independent of S, satisfying $\|\widehat{\beta}_1 - \beta^*\|_2 = O(\sigma)$. Let T be an ϵ-corrupted version of S, where T might depend on $\widehat{\beta}_1$. Define the set $T_1 := \left\{\widehat{\beta}_1 + (y_i' - (x_i)'^\top \widehat{\beta}_1)x_i' : (x_i', y_i') \in T\right\}$. Suppose $\epsilon' = \Theta\left(\epsilon + \frac{\log(1/\tau)}{n}\right) = O(1)$. Then given ϵ, T_1, and τ as inputs, the mean algorithm in Theorem C.10.2 returns an output $\widehat{\beta}$ satisfying*

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \sigma\left(\sqrt{\frac{p}{n}} + \sqrt{\epsilon} + \sqrt{\frac{\log(1/\tau)}{n}}\right),$$

with probability at least $1 - \tau$.

The proof of Theorem 5.6.1 is contained in Appendix C.10.1. We now consider the case when $\widehat{\beta}_1$ might depend on the data. Such a situation might arise if we were to perform sample splitting on an adversarially contaminated data set, meaning we would estimate $\widehat{\beta}_1$ from the first half of the data and use it to initialize a postprocessing step on the other half. Since the adversary is allowed to look at the whole data set, this could lead to dependence between the two halves. In such a case, the argument used in the proof of Theorem 5.6.1 cannot be applied because we do not necessarily have an i.i.d. data set when we condition on $\widehat{\beta}_1$. However, we may still obtain a looser error bound by taking a union bound over a large enough cover of $\mathcal{S}^{p-1}$. We have the following result, proved in Appendix C.10.2:

Theorem 5.6.2. *Consider the setting and notation in Theorem 5.6.1, where $\widehat{\beta}_1$ might depend on S. Set $\epsilon' = \Theta\left(\epsilon + \frac{\log(1/\tau)}{n} + \frac{p\log(pn)}{n}\right)$, where ϵ' is less than a small constant. Then running the filtering algorithm in Theorem C.1.1 with inputs*

$$T_1 := \left\{\widehat{\beta}_1 + (y_i' - (x_i)'^{\top}\widehat{\beta}_1)x_i' : (x_i', y_i') \in T\right\}$$

and ϵ' returns a set T' such that, with probability at least $1 - 2\tau$,

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \sigma\left(\sqrt{\frac{p\log(pn)}{n}} + \sqrt{\epsilon} + \sqrt{\frac{\log(1/\tau)}{n}}\right),$$

where $\widehat{\beta}$ is the empirical mean of the vectors in T'.

Remark 5.6.3. *Compared to the error bound in Theorem 5.6.1, the error bound in Theorem 5.6.2 contains an extra factor of $\sqrt{\log(pn)}$ in the first term. This arises from a covering argument, since we cannot simply condition on $\widehat{\beta}_1$ and argue that we still have i.i.d. data.*

Remark 5.6.4. *Cherapanamjeri et al. [CATJFB20] show that when both the covariate and noise distributions are sub-Gaussian, running the post-processing step once more to the output achieved by the procedure in Theorem 5.6.2 can improve the error dependence on ϵ from $O(\sigma\sqrt{\epsilon})$ to $O(\sigma\epsilon\log(1/\epsilon))$. This is because when $\|\widehat{\beta}_1 - \beta^*\|_2 \lesssim \sigma\sqrt{\epsilon}$, the covariance matrix of $\widehat{\beta}_1 + (y_i' - (x_i)'^{\top}\widehat{\beta}_1)x_i'$ is $O(\sigma^2\epsilon)$-close to the spherical matrix $\sigma^2 I$. When covariate and noise distributions satisfy $(k, 2)$-hypercontractivity, the same argument shows that the error dependence on ϵ would improve from $O(\sigma\sqrt{\epsilon})$ to $O(\sigma\epsilon^{1-1/k})$. In comparison, the filtered Huber regression algorithm (cf. Theorem 5.3.6) provably achieves an error of the form $O(\sigma\epsilon^{1-1/k})$ under only a k^{th} moment assumption on the covariate distribution.*

5.7 Simulations

We now present the results of the simulations on synthetic data to validate our theoretical findings. We demonstrate that covariate filtering improves estimation accuracy for heavy-tailed i.i.d. data (Section 5.7.1) and heavy-tailed data with adversarial corruption (Section 5.7.2). For our simulations, we take $n = 200$ and $p = 40$, which roughly corresponds to the linear-data regime $n = O(p)$. We measure the error in the usual ℓ_2-norm, i.e., $\|\widehat{\beta} - \beta^*\|_2$. For each plot, we conduct our experiments $T = 50,000$ times, and report how the empirical quantiles of the ℓ_2-error increase with the failure probability τ. The main goal of the plots is to demonstrate the effect of covariate filtering on Huber regression and LTS.

All estimators were implemented using NumPy For Huber regression, we ran gradient descent algorithm with a line-search procedure. For LTS, we ran our algorithm (Algorithm 13) for a fixed number of 100 steps. We found that both of these estimators converged with these choices of parameters. In each experiment, we sampled β^* independently from a sphere of unit norm. We initialized all of our estimators at the same point, which is also sampled independently from a sphere of unit norm, and hence its ℓ_2-distance from β^* is at most 2. We implemented the filter so that it removed a single point at every step, which corresponds to the version in Prasad et al. [PBR19].

We used the family of (symmetrized) Pareto distributions for for both covariates and additive noise. For $\alpha > 0$, we say that a real-valued random variable X follows an α-symmetrized-Pareto distribution if the probability density function $f_X(x)$, has polynomial tails, i.e., for all $x \in \mathbb{R}$, $f_X(x) \propto \left(\frac{1}{|x|+1}\right)^{1+\alpha}$. The k^{th} moment of X exists if and only if $k < \alpha$. We say that a multivariate random variable X follows an α-symmetrized-Pareto distribution if each coordinate of X is i.i.d. with an α-symmetrized-Pareto distribution.

5.7.1 Heavy-tailed Regression

We sample i.i.d. data from a heavy-tailed distribution without any corruption. As mentioned earlier, we set $n = 200$ and $p = 40$, and $\|\beta^*\|_2 = 1$, and ran our experiments $50,000$ times to calculate the empirical quantiles of various estimators as a function of τ. For our experiments, we sampled covariates and additive noise from symmetrized-Pareto distributions with parameter 2. Note that this choice of heavy-tailed distributions does not exactly satisfy our hypercontractivity assumption (Assumption 5.2.1), because the fourth moment is infinite.

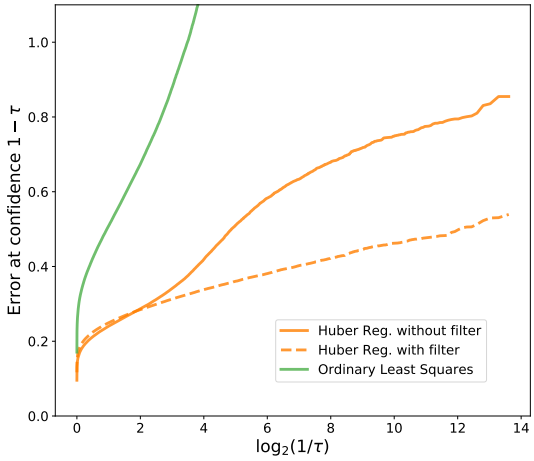

Figure 5.1: *

(a) Huber regression

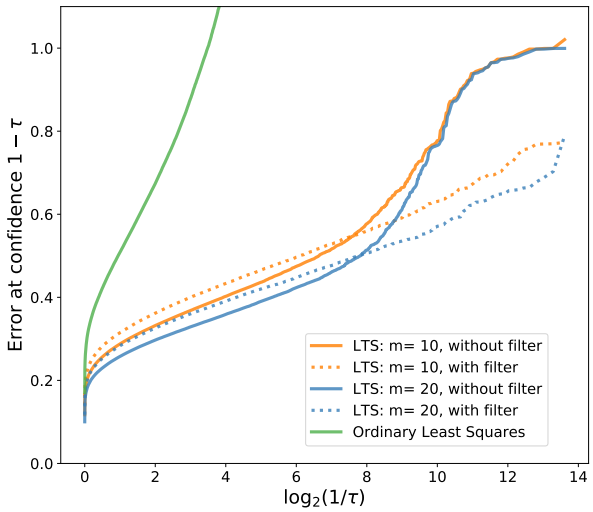

Figure 5.2: *

(b) LTS

Figure 5.3: Plots showing the effect of covariate filtering with heavy-tailed data ($n = 200, p = 40$). For Huber regression, we set $\gamma = 0.5$. In (b), m is the trimming parameter in Algorithm 13. The error is measured in terms of ℓ_2-error.

Figure 5.3 shows that covariate filtering improves the performance of Huber and LTS significantly, especially in the high-confidence regime when $\tau \to 0$. Figure 5.3 demonstrates that even removing 10 points out of 200 points can boost the accuracy of both Huber regression and LTS, where the Huber parameter is set to be 0.5. Between Huber regression and LTS with filtering step, we find that Huber regression has better

performance than LTS. Additional plots showing the effect of filtering as γ changes in Huber regression and as m changes in LTS are included in Appendix C.11 (cf. Figures C.1 and C.2). We find that the same phenomenon as in Figure 5.3 is demonstrated across a wide range of γ and m.

5.7.2 Adversarial Corruption

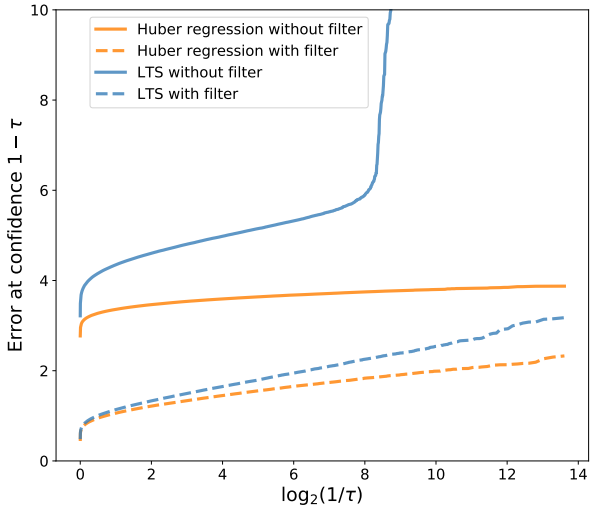

Figure 5.4: Plot showing the effect of covariate filtering on Huber regression and LTS when data are sampled from a heavy-tailed distribution and contain adversarial corruption, where $n = 200$, $p = 40$, and $\epsilon = 0.1$. The error is measured in terms of ℓ_2-error.

Once again, we set $n = 200$ and $p = 40$. We sampled covariates and responses from symmetrized-Pareto distributions with parameters 4 and 2, respectively. We consider the case $\epsilon = 0.1$, so $\epsilon n = 20$ points are corrupted in the following manner:

1. We replace the covariates $\{x_i\}$ of 10 random points by the deterministic point $10w$, where w is the vector with each coordinate equal to 1.

2. We replace the responses $\{y_i\}$ of 20 points, including the 10 points selected in the previous step, by a deterministic value 200.

We do not corrupt the covariates of all 20 points, since such a corruption scheme gives an advantage to the filtering step: if the filtering step perfectly removed all points with corrupted covariates, the data would be clean in the responses, as well. We run the filter so that it removes $1.5\epsilon n = 30$ points. For Huber regression, we again set $\gamma = 0.5$. For LTS, we set $m = 1.5\epsilon n = 30$ to handle ϵn corruption in responses. Figure 5.4 shows that the filtering step can significantly improve the performance of both Huber regression and LTS.

5.8 Discussion

We have presented several estimators that are simultaneously robust to heavy-tailed distributions and adversarial contamination. The main theme is that a simple preprocessing step applied to the covariates can be used to make classical estimators such as the Huber regression, LTS, and LAD estimators robust to contamination in both covariates and responses. Our preprocessing step leverages recent advances in algorithms for robust mean estimation, in which a filtering procedure was introduced to remove a small fraction of covariates to make the sample covariance matrix of the remaining points have a small spectral norm. The modified Huber regression estimator achieves a near-optimal error guarantee in this setting, whereas the LTS and LAD estimators can be used for initialization and/or parameter tuning, or augmented with a preprocessing step to achieve near-optimal error rates.

Aside from the filtering method analyzed in this paper, we note that other algorithms have been proposed, which—instead of returning a subset T' of the input data set T—return a distribution on T such that the weight at any point is at most $\frac{1}{(1-O(\epsilon))|T|}$ [DKKLMS16; SCV18; DHL19; CDGS20; ZJS22b]. Although we have not pursued such algorithms here, one might prove analogous results for robust regression using these alternative methods for preprocessing via one of the following two ap-

proaches: (i) discretize the distribution to obtain a set T' satisfying the conclusion in Theorem C.1.1; or (ii) study a weighted form of regression estimators (Huber regression, LAD, or LTS), where the loss at each point is weighted by the output of these algorithms.

Thinking more broadly, it would be interesting to see which other common regression estimators might benefit from covariate filtering as a preprocessing step. Another important line of future work is to extend this methodology to settings where β^* satisfies some structural assumptions, such as sparsity—this might involve proposing and analyzing a filtering step which would, with high probability, produce covariates which satisfy a restricted eigenvalue condition. Finally, we have assumed throughout the paper that the covariates and noise variables are independent, and the covariates are approximately isotropic; the question of whether our proposed algorithms could be analyzed under a more general dependency structure and unknown covariance which is not approximately isotropic remains open.

6 STATISTICAL QUERY LOWER BOUNDS FOR LIST-DECODABLE LINEAR REGRESSION

आए कुछ अब्र कुछ शराब आए
इस के बा'द आए जो अज़ाब आए

— फ़ैज़ अहमद फ़ैज़

We study the problem of list-decodable linear regression, where an adversary can corrupt a majority of the examples. Specifically, we are given a set T of labeled examples $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ and a parameter $0 < \alpha < 1/2$ such that an α-fraction of the points in T are i.i.d. samples from a linear regression model with Gaussian covariates, and the remaining $(1 - \alpha)$-fraction of the points are drawn from an arbitrary noise distribution. The goal is to output a small list of hypothesis vectors such that at least one of them is close to the target regression vector. Our main result is a Statistical Query (SQ) lower bound of $d^{\mathrm{poly}(1/\alpha)}$ for this problem. Our SQ lower bound qualitatively matches the performance of previously developed algorithms, providing evidence that current upper bounds for this task are nearly best possible.

6.1 Introduction

6.1.1 Background and Motivation

Linear regression is one of the oldest and most fundamental statistical tasks with numerous applications in the sciences [RL87; Die01; McD09]. In the standard setup, the data are labeled examples $(x^{(i)}, y^{(i)})$, where the examples (covariates) $x^{(i)}$ are i.i.d. samples from a distribution D_x on $\mathbb{R}^d$ and the labels $y^{(i)}$ are noisy evaluations of a linear function. More specifically, each label is of the form $y^{(i)} = \beta \cdot x^{(i)} + \eta^{(i)}$, where $\eta^{(i)}$ is the observation noise, for an unknown target regression vector $\beta \in \mathbb{R}^d$. The objective is to approximately

recover the hidden regression vector. In this basic setting, linear regression is well-understood. For example, under Gaussian distribution, the least-squares estimator is known to be statistically and computationally efficient.

Unfortunately, classical efficient estimators inherently fail in the presence of even a very small fraction of adversarially corrupted data. In several applications of modern data analysis, including machine learning security [BNJT10; BNL12; SKL17; DKKLSS19] and exploratory data analysis, e.g., in biology [RPWCKZF02; PLJD10; LATSCR+08], typical datasets contain arbitrary or adversarial outliers. Hence, it is important to understand the algorithmic possibilities and fundamental limits of learning and inference in such settings. Robust statistics focuses on designing estimators tolerant to a small amount of contamination, where the outliers are the *minority* of the dataset. Classical work in this field [HRRS11; HR09] developed robust estimators for various basic tasks, alas with exponential runtime. More recently, a line of work in computer science, starting with [DKKLMS16; LRV16], developed the first computationally efficient robust learning algorithms for various high-dimensional tasks. Subsequently, there has been significant progress in algorithmic robust statistics by several communities, see [DK19] for a survey on the topic.

In this paper, we study high-dimensional robust linear regression in the presence of a *majority* of adversarial outliers. As we explain below, in several applications, asking for a minority of outliers is too strong of an assumption. It is thus natural to ask what notion of learning can capture the regime when the clean data points (inliers) constitute the *minority* of the dataset. While outputting a *single* accurate hypothesis in this regime is information-theoretically impossible, one may be able to compute a *small list* of hypotheses with the guarantee that *at least one of them* is accurate. This relaxed notion is known as *list-decodable learning* [BBV08; CSV17], formally defined below.

Definition 6.1.1 (List-Decodable Learning)**.** *Given a parameter* $0 < \alpha < 1/2$ *and a dis-*

tribution family $\mathcal{D}$ on $\mathbb{R}^d$, the algorithm specifies $n \in \mathbb{Z}_+$ and observes n i.i.d. samples from a distribution $E = \alpha D + (1-\alpha)N$, where D is an unknown distribution in $\mathcal{D}$ and N is arbitrary. We say D is the distribution of inliers, N is the distribution of outliers, and E is an $(1-\alpha)$-corrupted version of D. Given sample access to an $(1-\alpha)$-corrupted version of D, the goal is to output a "small" list of hypotheses $\mathcal{L}$ at least one of which is (with high probability) close to the target parameter of D.

We note that a list of size $O(1/\alpha)$ typically suffices; an algorithm with a $\mathrm{poly}(1/\alpha)$ sized list, or even a worse function of $1/\alpha$ (but independent of the dimension d) is also considered acceptable.

Natural applications of list-decodable learning include crowdsourcing, where a majority of participants could be unreliable [SVC16; MV18], and semi-random community detection in stochastic block models [CSV17]. List-decoding is also useful in the context of semi-verified learning [CSV17; MV18], where a learner can audit a very small amount of trusted data. If the trusted dataset is too small to directly learn from, using a list-decodable learning procedure, one can pinpoint a candidate hypothesis consistent with the verified data. Importantly, list-decodable learning generalizes the task of learning mixture models, see, e.g., [DeV89; JJ94; ZJD16; LL18; KC20; CLS20; DK20] for the case of linear regression studied here. Roughly speaking, by running a list-decodable estimation procedure with the parameter α equal to the smallest mixing weight, each true cluster of points is an equally valid ground-truth distribution, so the output list must contain candidate parameters close to each of the true parameters.

In list-decodable linear regression (the focus of this paper), D is a distribution on pairs (X, y), where X is a standard Gaussian on $\mathbb{R}^d$, y is approximately a linear function of x, and the algorithm is asked to approximate the hidden regressor. The following definition specifies the distribution family $\mathcal{D}$ of the inliers for the case of linear regression with Gaussian covariates.

Definition 6.1.2 (Gaussian Linear Regression)**.** *Fix $\sigma > 0$. For $\beta \in \mathbb{R}^d$, let D_β be the distribution over (X, y), $X \in \mathbb{R}^d$, $y \in \mathbb{R}$, such that $X \sim \mathcal{N}(0, I_d)$ and $y = \beta^\top X + \eta$, where $\eta \sim \mathcal{N}(0, \sigma^2)$ independently of X. We define $\mathcal{D}$ to be the set $\{D_\beta : \beta \in S'\}$ for some set $S' \subseteq \mathbb{R}^d$.*

Recent algorithmic progress [KKK19; RY20a] has been made on this problem using the SoS hierarchy. The guarantees in [KKK19; RY20a] are very far from the information-theoretic limit in terms of sample complexity. In particular, they require $d^{\mathrm{poly}(1/\alpha)}$ samples and time to obtain non-trivial error guarantees (see Table 6.1): [KKK19] obtains an error guarantee of $O(\sigma/\alpha)$ with a list of size $O(1/\alpha)$, whereas [RY20a] obtains an error guarantee of $O(\sigma/\alpha^{3/2})$ with a list of size $(1/\alpha)^{O(\log(1/\alpha))}$.

On the other hand, as shown in this paper (see Theorem 6.1.4), $\mathrm{poly}(d/\alpha)$ samples information-theoretically suffice to obtain near-optimal error guarantees. This raises the following natural question:

What is the complexity of list-decodable linear regression?

Are there efficient algorithms with significantly better sample-time tradeoffs?

We study the above question in a natural and well-studied restricted model of computation, known as the Statistical Query (SQ) model [Kea98]. As the main result of this paper, we prove strong SQ lower bounds for this problem. Via a recently established equivalence [BBHLS21], our SQ lower bound also implies low-degree testing lower bounds for this task. Our lower bounds can be viewed as evidence that current upper bounds for this problem may be qualitatively best possible.

Before we state our contributions in detail, we give some background on SQ algorithms. SQ algorithms are a broad class of algorithms that are only allowed to query expectations of bounded functions of the distribution rather than directly access samples. Formally, an SQ algorithm has access to the following oracle.

Definition 6.1.3 (STAT Oracle)**.** *Let D be a distribution on $\mathbb{R}^d$. A statistical query is a bounded function $q : \mathbb{R}^d \to [-1, 1]$. For $\tau > 0$, the $\mathrm{STAT}(\tau)$ oracle responds to the query q with a value v such that $|v - \mathbb{E}_{X\sim D}[q(X)]| \leq \tau$. We call τ the tolerance of the statistical query.*

The SQ model was introduced by Kearns [Kea98] in the context of supervised learning as a natural restriction of the PAC model [Val84]. Subsequently, the SQ model has been extensively studied in a plethora of contexts (see, e.g., [Fel16] and references therein). The class of SQ algorithms is rather broad and captures a range of known supervised learning algorithms. More broadly, several known algorithmic techniques in machine learning are known to be implementable using SQs. These include spectral techniques, moment and tensor methods, local search (e.g., Expectation Maximization), and many others (see, e.g., [FGRVX17; FGV17]).

6.1.2 Our Results

We start by showing that $\mathrm{poly}(d/\alpha)$ samples are sufficient to obtain a near-optimal error estimator, albeit with a computationally inefficient algorithm.

Theorem 6.1.4 (Information-Theoretic Bound)**.** *There is a (computationally inefficient) list-decoding algorithm for Gaussian linear regression that uses $O(d/\alpha^3)$ samples, returns a list of $O(1/\alpha)$ many hypothesis vectors, and has ℓ_2-error guarantee of $O((\sigma/\alpha)\sqrt{\log(1/\alpha)})$. Moreover, if the dimension d is sufficiently large, any list-decoding algorithm that outputs a list of size $\mathrm{poly}(1/\alpha)$ must have ℓ_2-error at least $\Omega((\sigma/\alpha)/\sqrt{\log(1/\alpha)})$.*

The proof of this result is given in Section 6.2 (see Theorems 6.2.1 and 6.2.4). Our main result is a strong SQ lower bound for the list-decodable Gaussian linear regression problem. We establish the following theorem (see Theorem 6.3.1 for a more detailed formal statement).

Algorithmic Result	Sample Size	Running Time	List size
[KKK19]	$(d/\alpha)^{O(1/\alpha^4)}$	$(d/\alpha)^{O(1/\alpha^8)}$	$O(1/\alpha)$
[RY20a]	$d^{O(1/\alpha^4)}$	$d^{O(1/\alpha^8)}(1/\alpha)^{\log(1/\alpha)}$	$(1/\alpha)^{O(\log(1/\alpha))}$

Table 6.1: The table summarizes the sample complexity, running time, and list size of the known list-decodable linear regression algorithms in order to obtain a $1/4$-additive approximation to the hidden regression vector β in the setting of Theorem 6.1.5, i.e., when $\|\beta\|_2 \leq 1$ and σ is sufficiently small as a function of α: [KKK19] requires $\sigma = O(\alpha)$ and [RY20a] requires $\sigma = O(\alpha^{3/2})$.

Theorem 6.1.5 (SQ Lower Bound)**.** *Assume that the dimension $d \in \mathbb{Z}_+$ is sufficiently large and consider the problem of list-decodable linear regression, where the fraction of inliers is $\alpha \in (0, 1/2)$, the regression vector $\beta \in \mathbb{R}^d$ has norm $\|\beta\|_2 \leq 1$, and the additive noise has standard deviation $\sigma \leq \alpha$. Then any SQ algorithm that returns a list $\mathcal{L}$ of candidate vectors containing a $\widehat{\beta}$ such that $\|\widehat{\beta} - \beta\|_2 \leq 1/4$ does one of the following:*

- *it uses at least one query with tolerance at most $d^{-\Omega(1/\sqrt{\alpha})}/\sigma$,*
- *it makes $2^{d^{\Omega(1)}}$ queries, or*
- *it returns a list of size $|\mathcal{L}| = 2^{d^{\Omega(1)}}$.*

Informally speaking, Theorem 6.1.5 shows that no SQ algorithm can approximate β to constant accuracy with a sub-exponential in $d^{\Omega(1)}$ size list and sub-exponential in $d^{\Omega(1)}$ many queries, unless using queries of very small tolerance – that would require at least $\sigma d^{\Omega(1/\sqrt{\alpha})}$ samples to simulate. For σ not too small, e.g., $\sigma = \mathrm{poly}(\alpha)$, in view of Theorem 6.1.4, this result can be viewed as an information-computation tradeoff for the problem, within the class of SQ algorithms.

A conceptual implication of Theorem 6.1.5 is that list-decodable linear regression is harder (within the class of SQ algorithms) than the related problem of learning mixtures of linear regressions (MLR). Recent work [DK20] gave an algorithm (easily implementable in SQ) for learning MLR with k equal weight separated components

(under Gaussian covariates) with sample complexity and running time $k^{\mathrm{polylog}(k)}$, i.e., *quasi-polynomial* in k. Recalling that one can reduce k-MLR (with well-separated components) to list-decodable linear regression for $\alpha = 1/k$, Theorem 6.1.5 implies that the aforementioned algorithmic result cannot be obtained via such a reduction.

Remark 6.1.6. *While the main focus of this work is on the SQ model, our result has immediate implications to a related popular restricted computational model — that of low-degree (polynomial) algorithms [HS17; HKPRSS17; Hop18]. Recent work [BBHLS21] established that (under certain assumptions) an SQ lower bound also implies a qualitatively similar lower bound in the low-degree model. We leverage this connection to show a similar lower bound in this model (see Section 6.6).*

6.1.3 Overview of Techniques

In this section, we provide a detailed overview of our SQ lower bound construction. We recall that there exists a general methodology for establishing SQ lower bounds via an appropriate complexity measure, known as SQ dimension. Several related notions of SQ dimension exist in the literature, see, e.g., [BFJKMR94; FGRVX17; Fel17]. Here we focus on the framework introduced in [FGRVX17] for search problems over distributions, which is more natural in our setting. A lower bound on the SQ dimension of a search problem provides an unconditional lower bound on the SQ complexity of the problem. Roughly speaking, for a notion of correlation between distributions in our family $\mathcal{D}$ (Definition 6.1.9), establishing an SQ lower bound amounts to constructing a large cardinality sub-family $\mathcal{D}' \subseteq \mathcal{D}$ such that every pair of distributions in $\mathcal{D}'$ are nearly uncorrelated with respect to a given reference distribution R (see Definition 6.1.11 and Lemma 6.1.12).

A general framework for constructing SQ-hard families of distributions was introduced in [DKS17], which showed the following: Let the reference distribution R be

$\mathcal{N}(0, I)$ and A be a univariate distribution whose low-degree moments match those of the standard Gaussian (and which satisfies an additional mild technical condition). Let $P_{A,v}$ be the distribution that is a copy of A in the v-direction and standard Gaussian in the orthogonal complement (Definition 6.1.13). Then the distribution family $\{P_{A,v}\}_{v \in S}$, where S is a set of nearly orthogonal unit vectors, satisfies the pairwise nearly uncorrelated property (Lemma 6.1.14), and is therefore SQ-hard to learn.

Unfortunately, the [DKS17] framework does not suffice in the supervised setting of the current paper for the following reason: The joint distribution over labeled examples (X, y) in our setting does not possess the symmetry properties required for moment-matching with the reference $R = \mathcal{N}(0, I)$ to be possible. Specifically, the behavior of y will necessarily be somewhat different than the behavior of X. To circumvent this issue, we leverage an idea from [DKS19]. The high-level idea is to construct distributions E_v on (X, y) such that for any fixed value y_0 of y, the conditional distribution of $X \mid y = y_0$ under E_v is of the form $P_{A,v}$ described above, where A is replaced with some A_{y_0}.

We further explain this modified construction. Note that E_v should be of the form $\alpha D_v + (1-\alpha)N_v$, where D_v is the inlier distribution (corresponding to the clean samples from the linear regression model) and N_v is the outlier (noise) distribution. To understand what properties our distribution should satisfy, we start by looking at the inlier distribution D. By definition, for $(X, y) \sim D$, we have that $y = \beta^\top X + \eta$, where $X \sim N(0, I)$ and $\eta \sim N(0, \sigma^2)$ is independent of X. A good place to start here is to understand the distribution of X conditioned on $y = y_0$, for some y_0, under D. It is not hard to show (Fact 6.3.3) that this conditional distribution is already of the desired form $P_{A,\beta}$: it is a product of a $(d-1)$-dimensional standard Gaussian in directions orthogonal to β, while in the β-direction it is a much narrower Gaussian with mean proportional to y_0. To establish our SQ-hardness result, we would like to mix this conditional distribution with a carefully selected outlier distribution $N \mid y = y_0$, such that the resulting mixture

$E \mid y = y_0$ matches many of its low-degree moments with the standard Gaussian in the β-direction, while being standard Gaussian in the orthogonal directions. In the setting of minority of outliers, [DKS19] was able to provide an explicit formula for N and match *three* moments to show an SQ lower bound of $\Omega(d^2)$. The main technical difficulty in our paper is that, in order to prove the desired SQ lower bound of $\Omega(d^{\mathrm{poly}(1/\alpha)})$, we need to match $\mathrm{poly}(1/\alpha)$ many moments. We explain how to achieve this below.

Here we take a different approach and establish the existence of the desired outlier distribution $N|y = y_0$ in a non-constructive manner. We note that our problem is an instance of the moment-matching problem, where given a sequence of real numbers, the goal is to decide whether a distribution exists having that sequence as its low-degree moments. At a high-level, we leverage classical results that tackle this general question by formulating a linear program (LP) and using LP-duality to derive necessary and sufficient feasibility conditions (see [KS53] and Theorem 6.4.1). This moment-matching via LP duality approach is fairly general, but stumbles upon two technical obstacles in our setting.

The first technical issue is that our final distributions E_v on (X, y) need to have bounded χ^2-divergence with respect to the reference distribution, since the pairwise correlations scale with this quantity (see Lemma 6.1.14). To guarantee this, we can ensure that the outlier distribution in the β-direction is in fact equal to the convolution of a distribution with bounded support with a narrow Gaussian: (i) The contraction property of this convolution operator means that it can only reduce the χ^2-divergence, and (ii) the bounded support can be used in combination with tail-bounds on Hermite polynomials (Lemma 6.3.10) to bound from above the contribution to the χ^2-divergence of each Hermite coefficient of our distribution (Lemma 6.3.7). These additional constraints necessitate a modification to the moment-matching problem, but it can still be readily analyzed (Theorem 6.3.6).

The second and more complicated issue involves the fraction of outliers, i.e., the parameter "$1-\alpha$". Unfortunately, it is easy to see that the fraction of outliers necessary to make the conditional distributions match the desired number of moments must necessarily go to 1 as $|y|$ goes to infinity: As $|y|$ gets bigger, the conditional distribution of inliers moves further away from $\mathcal{N}(0, I)$ (Fact 6.3.3) and thus needs to be mixed more heavily with outliers to be corrected. This is a significant problem, since by definition we can only afford to use a $(1-\alpha)$-fraction of outliers overall. To handle this issue, we consider a reference distribution R on (X, y) that has much heavier tails in y than the distribution of inliers has. This essentially means that as $|y|$ gets large, the conditional probability that a sample is an outlier gets larger and larger. This is balanced by having slightly lower fraction of outliers for smaller values of $|y|$, in order to ensure that the total fraction of outliers is still at most $1-\alpha$. To address this issue, we leverage the fact that the probability that a clean sample has large value of $|y|$ is very small. Consequently, we can afford to make the error rates for such y quite large without increasing the overall probability of error by very much.

6.1.4 Preliminaries

Notation We use $\mathbb{N}$ to denote natural numbers and $\mathbb{Z}_+$ to denote positive integers. For $n \in \mathbb{Z}_+$ we denote $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$ and use $\mathcal{S}^{d-1}$ for the d-dimensional unit sphere. We denote by $\mathbb{I}(\mathcal{E})$ the indicator function of the event $\mathcal{E}$. We use I_d to denote the $d \times d$ identity matrix. For a random variable X, we use $\mathbb{E}[X]$ for its expectation. For $m \in \mathbb{Z}_+$, the m-th moment of X is defined as $\mathbb{E}[X^m]$. We use $\mathcal{N}(\mu, \Sigma)$ to denote the Gaussian distribution with mean μ and covariance matrix Σ. We let ϕ denote the pdf of the one-dimensional standard Gaussian. When D is a distribution, we use $X \sim D$ to denote that the random variable X is distributed according to D. For a vector $x \in \mathbb{R}^d$, we let $\|x\|_2$ denote its ℓ_2-norm. For $y \in \mathbb{R}$, we denote by δ_y the Dirac delta distribution at y, i.e., the distribution

that assigns probability mass 1 to the single point $y \in \mathbb{R}$ and zero elsewhere. When there is no confusion, we will use the same letters for distributions and their probability density functions.

Hermite Analysis Hermite polynomials form a complete orthogonal basis of the vector space $L^2(\mathbb{R}, \mathcal{N}(0,1))$ of all functions $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[f^2(X)] < \infty$. There are two commonly used types of Hermite polynomials. The *physicist's* Hermite polynomials, denoted by H_k for $k \in \mathbb{N}$ satisfy the following orthogonality property with respect to the weight function e^{-x^2}: for all $k, m \in \mathbb{N}$, $\int_{\mathbb{R}} H_k(x)H_m(x)e^{-x^2}\mathrm{dx} = \sqrt{\pi}2^k k!\,\mathbb{I}(k=m)$. The *probabilist's* Hermite polynomials H_{e_k} for $k \in \mathbb{N}$ satisfy

$$\int_{\mathbb{R}} H_{e_k}(x)H_{e_m}(x)e^{-x^2/2}\mathrm{dx} = k!\sqrt{2\pi}\,\mathbb{I}(k=m)$$

and are related to the physicist's polynomials through $H_{e_k}(x) = 2^{-k/2}H_k(x/\sqrt{2})$. We will mostly use the *normalized probabilist's* Hermite polynomials, $h_k(x) = H_{e_k}(x)/\sqrt{k!}$, $k \in \mathbb{N}$ for which $\int_{\mathbb{R}} h_k(x)h_m(x)e^{-x^2/2}\mathrm{dx} = \sqrt{2\pi}\,\mathbb{I}(k=m)$. These polynomials are the ones obtained by Gram-Schmidt orthonormalization of the basis $\{1, x, x^2, \ldots\}$ with respect to the inner product $\langle f, g\rangle_{\mathcal{N}(0,1)} = \mathbb{E}_{X\sim\mathcal{N}(0,1)}[f(X)g(X)]$. Every function $f \in L^2(\mathbb{R}, \mathcal{N}(0,1))$ can be uniquely written as $f(x) = \sum_{i\in\mathbb{N}} a_i h_i(x)$ and we have $\lim_{n\to n} \mathbb{E}_{x\sim\mathcal{N}(0,1)}[(f(x) - \sum_{i=0}^{n} a_i h_i(x))^2] = 0$ (see, e.g., [AAR99]).

Ornstein-Uhlenbeck Operator For a $\rho > 0$, we define the *Gaussian noise* (or *Ornstein-Uhlenbeck*) operator U_ρ as the operator that maps a distribution F on $\mathbb{R}$ to the distribution of the random variable $\rho X + \sqrt{1-\rho^2}Z$, where $X \sim F$ and $Z \sim \mathcal{N}(0,1)$ independently of X. A well-known property of *Ornstein–Uhlenbeck* operator is that it operates diagonally with respect to Hermite polynomials.

Fact 6.1.7 (see, e.g., [ODo14]). *For any Hermite polynomial h_i, any distribution F on $\mathbb{R}$, and $\rho \in (0,1)$, it holds that $\mathbb{E}_{X\sim U_\rho F}[h_i(X)] = \rho^i \mathbb{E}_{X\sim F}[h_i(X)]$.*

Background on the SQ Model We provide the basic definitions and facts that we use.

Definition 6.1.8 (Search problems over distributions). *Let $\mathcal{D}$ be a set of distributions over $\mathbb{R}^d$, $\mathcal{F}$ be a set called solutions, and $\mathcal{Z} : \mathcal{D} \to 2^{\mathcal{F}}$ be a map that assigns sets of solutions to distributions of $\mathcal{D}$. The* distributional search problem *$\mathcal{Z}$ over $\mathcal{D}$ and $\mathcal{F}$ is to find a valid solution $f \in \mathcal{Z}(D)$ given statistical query oracle access to an unknown $D \in \mathcal{D}$.*

The hardness of these problems is conveniently captured by the SQ dimension. For this, we first need to define the notion of correlation between distributions.

Definition 6.1.9 (Pairwise Correlation). *The pairwise correlation of two distributions with probability density functions $D_1, D_2 : \mathbb{R}^d \to \mathbb{R}_+$ with respect to a reference distribution with density $R : \mathbb{R}^d \to \mathbb{R}_+$, where the support of R contains the supports of D_1 and D_2, is defined as $\chi_R(D_1, D_2) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} D_1(x)D_2(x)/R(x)\,\mathrm{dx} - 1$. When $D_1 = D_2$, the pairwise correlation becomes the same as the χ^2-divergence between D_1 and R, i.e., $\chi^2(D_1, R) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} D_1^2(x)/R(x)\mathrm{dx} - 1$.*

Definition 6.1.10. *For $\gamma, \beta > 0$, the set of distributions $\mathcal{D} = \{D_1, \ldots, D_m\}$ is called (γ, β)-correlated relative to the distribution R if $|\chi_R(D_i, D_j)| \le \gamma$, if $i \neq j$, and $|\chi_R(D_i, D_j)| \le \beta$ otherwise.*

The statistical dimension of a search problem is based on the largest set of (γ, β)-correlated distributions assigned to each solution.

Definition 6.1.11 (Statistical Dimension). *For $\gamma, \beta > 0$, a search problem $\mathcal{Z}$ over a set of solutions $\mathcal{F}$ and a class $\mathcal{D}$ of distributions over X, we define the* statistical dimension *of $\mathcal{Z}$, denoted by $\mathrm{SD}(\mathcal{Z}, \gamma, \beta)$, to be the largest integer m such that there exists a reference distribution R over X and a finite set of distributions $\mathcal{D}_R \subseteq \mathcal{D}$ such that for any solution $f \in \mathcal{F}$, the set $\mathcal{D}_f = \mathcal{D}_R \setminus \mathcal{Z}^{-1}(f)$ is (γ, β)-correlated relative to R and $|\mathcal{D}_f| \ge m$.*

Lemma 6.1.12 (Corollary 3.12 in [FGRVX17]). *Let $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over $\mathbb{R}^d$. For $\gamma, \beta > 0$, let $s = \mathrm{SD}(\mathcal{Z}, \gamma, \beta)$ be the statistical dimension of the problem. For any $\gamma' > 0$, any SQ algorithm for $\mathcal{Z}$ requires either $s\gamma'/(\beta - \gamma)$ queries or at least one query to $\mathrm{STAT}(\sqrt{\gamma + \gamma'})$ oracle.*

We continue by recalling the machinery from [DKS17] that will be used for our construction.

Definition 6.1.13 (High-Dimensional Hidden Direction Distribution). *For a unit vector $v \in \mathbb{R}^d$ and a distribution A on the real line with probability density function $A(x)$, define $P_{A,v}$ to be a distribution over $\mathbb{R}^d$, where $P_{A,v}$ is the product distribution whose orthogonal projection onto the direction of v is A, and onto the subspace perpendicular to v is the standard $(d-1)$-dimensional normal distribution. That is, $P_{A,v}(x) := A(v^\top x)\phi_{\perp v}(x)$, where $\phi_{\perp v}(x) = \exp\left(-\|x - (v^\top x)v\|_2^2/2\right)/(2\pi)^{(d-1)/2}$.*

The distributions $\{P_{A,v}\}$ defined above are shown to be nearly uncorrelated as long as the directions where A is embedded are pairwise nearly orthogonal.

Lemma 6.1.14 (Lemma 3.4 in [DKS17]). *Let $m \in \mathbb{Z}_+$. Let A be a distribution over $\mathbb{R}$ that agrees with the first m moments of $\mathcal{N}(0,1)$. For any v, let $P_{A,v}$ denote the distribution from Definition 6.1.13. For all $v, u \in \mathbb{R}^d$, we have that $\chi_{\mathcal{N}(0,I_d)}(P_{A,v}, P_{A,u}) \leq |u^\top v|^{m+1}\chi^2(A, \mathcal{N}(0,1))$.*

The following result shows that there are exponentially many nearly-orthogonal unit vectors.

Lemma 6.1.15 (see, e.g., Lemma 3.7 in [DKS17]). *For any $0 < c < 1/2$, there is a set S, of at least $2^{\Omega(d^c)}$ unit vectors in $\mathbb{R}^d$, such that for each pair of distinct $v, v' \in S$, it holds $|v^\top v'| \leq O(d^{c-1/2})$.*

6.1.5 Prior and Related Work

Early work in robust statistics, starting with the pioneering works of Huber and Tukey [Hub64; Tuk75], pinned down the sample complexity of high-dimensional robust estimation with a minority of outliers. In contrast, until relatively recently, even the most basic computational questions in this field were poorly understood. Two concurrent works [DKKLMS16; LRV16] gave the first provably robust and efficiently computable estimators for robust mean and covariance estimation. Since the dissemination of these works, there has been a flurry of activity on algorithmic robust estimation in a variety of high-dimensional settings; see [DK19] for a recent survey on the topic. Notably, the robust estimators developed in [DKKLMS16] are scalable in practice and yield a number of applications in exploratory data analysis [DKKLMS17] and adversarial machine learning [TLM18; DKKLSS19]

The list-decodable learning setting studied in this paper was first considered in [CSV17] with a focus on mean estimation. [CSV17] gave a polynomial-time algorithm with near-optimal statistical guarantees for list-decodable mean estimation under a bounded covariance assumption on the clean. Subsequent work has led to significantly faster algorithms for the bounded covariance setting [DKK20; CMY20; DKKLT21; DKKLT22] and polynomial-time algorithms with improved error guarantees under stronger distributional assumptions [DKS18; KSS18]. More recently, a line of work developed list-decodable learners for more challenging tasks, including linear regression [KKK19; RY20a] and subspace recovery [RY20b; BK21].

6.2 Information-Theoretic Bounds

6.2.1 Upper Bound on Sample Complexity

In this section, we show that $n = \mathrm{poly}(d, 1/\alpha)$ samples suffice for list-decodable linear regression.

Theorem 6.2.1. *There is a (computationally inefficient) algorithm that uses $O(d/\alpha^3)$ samples from a $(1-\alpha)$-corrupted version of a Gaussian linear regression model of Definition 6.1.2 with $S' = \mathbb{R}^d$, and returns a list $\mathcal{L}$ of $|\mathcal{L}| = O(1/\alpha)$ many hypotheses such that with high probability at least one of them is within ℓ_2-distance $O((\sigma/\alpha)\sqrt{\log(1/\alpha)})$ from the regression vector.*

The proof strategy is similar to [DKS18]. When S is a set, we use the notation $X \sim_u S$ to denote that X is distributed according to the uniform distribution on S. We require the following theorem:

Fact 6.2.2 (VC Inequality)**.** *Let $\mathcal{F}$ be a class of Boolean functions with finite VC dimension $\mathrm{VC}(\mathcal{F})$ and let a probability distribution D over the domain of these functions. For a set S of n independent samples from D*

$$\sup_{f\in\mathcal{F}} \left| \mathop{\mathbb{P}}_{X\sim_u S}[f(X)] - \mathop{\mathbb{P}}_{X\sim D}[f(X)] \right| \lesssim \sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} \,,$$

with probability at least $1-\tau$.

Proof of Theorem 6.2.1. Recall the notation in Definitions 6.1.1 and 6.1.2. Let T be the set of points generated by the $(1-\alpha)$-corrupted version of D_{β^*} for some unknown $\beta^* \in \mathbb{R}^d$. Let S_1 be the set of points that are sampled from D_{β^*}. Since inliers are sampled with probability α, we have that $|S_1| \geq \alpha|T|/2$ with high probability. For a $t \geq 0$, define $\mathcal{H}_{t,\gamma}$

as follows:

$$\mathcal{H}_{t,\gamma} := \Bigg\{ \beta \in \mathbb{R}^d : \ \exists T' \subset T, \ |T'| = \alpha|T|/2, \tag{6.1}$$

$$\mathbb{P}_{(X,y)\sim_u T'}[|y - X^\top \beta| > \sigma t] \leq \alpha/20, \tag{6.2}$$

$$\forall v \in \mathcal{S}^{d-1}, \gamma' \geq \gamma : \ \mathbb{P}_{(X,y)\sim_u T'}[|y - X^\top \beta - \gamma' v^\top X| \leq \sigma t] \leq \ \alpha/20 \Bigg\}. \tag{6.3}$$

Recall that the distribution of inliers is $X \sim \mathcal{N}(0, I_d)$ and $y = X^\top \beta^* + \eta$, where $\eta \sim \mathcal{N}(0, \sigma^2)$ independent of X. If $|S_1| \geq Cd/\alpha^2$ for a sufficiently large constant C, then we claim that $\beta^* \in \mathcal{H}_{t,\gamma}$ with $t = \Theta(\sqrt{\log(1/\alpha)})$ and $\gamma = 40\sigma t/\alpha = \Theta((\sigma/\alpha)\sqrt{\log(1/\alpha)})$. Let S' be a set of i.i.d. points sampled from D_{β^*} with $|S'| = |T|\alpha/2$. We first argue that conditions (6.2) and (6.3) hold under $(X, y) \sim D_{\beta^*}$, even after replacing $\alpha/20$ with $\alpha/40$ in conditions (6.2) and (6.3), with the claimed bounds on t and γ, and then the required result on $(X, y) \sim_u S'$ will follow from the VC inequality. Since $y - X^\top \beta^* \sim \mathcal{N}(0, \sigma^2)$ under D_{β^*}, we get that $\mathbb{P}[|y - X^\top \beta^*| > \sigma t] \leq \alpha/40$ because of Gaussian concentration. Let $G \sim \mathcal{N}(0, 1)$ independent of η. For condition (6.3), the expression again reduces to concentration of a Gaussian distribution:

$$\mathbb{P}_{\eta\sim\mathcal{N}(0,\sigma^2), G\sim\mathcal{N}(0,1)}[|\eta + \gamma' G| \leq \sigma t] = \mathbb{P}_{Z\sim\mathcal{N}(0,\sigma^2+\gamma'^2)}[|Z| \ \leq \sigma t] \lesssim \frac{\sigma t}{\gamma'},$$

which is less than $\alpha/40$ for $\gamma' \geq \gamma = 40t\sigma/\alpha$. The desired conclusion now follows by noting that conditions (6.2) and (6.3) follow by uniform concentration of linear threshold functions on (X, y), which have VC dimension $O(d)$ and the condition that $|S'| = \Omega(d/\alpha^2)$.

We then show that any γ-packing of the set $\mathcal{H}_{t,\gamma}$ has size $O(1/\alpha)$. Having this, it follows that there exists a γ-cover of size $O(1/\alpha)$ and the output of the algorithm, $\mathcal{L}$, consists of returning any such cover. The key claim for bounding the size of any γ-

packing is that the pairwise intersections between the sets T' from condition (6.1) are small.

Claim 6.2.3. *Let $\beta_1, \dots, \beta_k \in \mathcal{H}_{t,\gamma}$ such that $\|\beta_i - \beta_j\|_2 > \gamma$ for all $i, j \in [k]$ and $i \neq j$. Let T_i' be the corresponding subsets of T satisfying the condition (6.1). Then $|T_i' \cap T_j'| \leq \alpha(|T_i'| + |T_j'|)/20$.*

Proof. Fix an $i \neq j$. Let $\beta_i - \beta_j = v\gamma'$, where $v \in \mathcal{S}^{d-1}$ and $\gamma' \geq \gamma$. Let $\mathcal{E}$ be the event $\{(X, y) : |y - X^\top \beta_j| \leq \sigma t\}$ and $\mathcal{E}^c$ be its complement. As T_i' and T_j' are sets of size $\alpha|T|/2$, we have that

$$
\begin{aligned}
|T_i' \cap T_j'| &= \frac{|T_i'| + |T_j'|}{2} \left(\frac{|T_i' \cap T_j' \cap \mathcal{E}|}{|T_i'|} + \frac{|T_i' \cap T_j' \cap \mathcal{E}^c|}{|T_j'|} \right) \\
&\leq \frac{|T_i'| + |T_j'|}{2} \left(\frac{|T_i' \cap \mathcal{E}|}{|T_i'|} + \frac{|T_j' \cap \mathcal{E}^c|}{|T_j'|} \right) = \frac{|T_i'| + |T_j'|}{2} \left(\mathop{\mathbb{P}}_{(X,y) \sim_u T_i'}[\mathcal{E}] + \mathop{\mathbb{P}}_{(X,y) \sim_u T_j'}[\mathcal{E}^c] \right).
\end{aligned}
$$

As $\beta_j \in \mathcal{H}_{t,\gamma}$, we have that $\mathbb{P}_{(X,y) \sim_u T_j'}[\mathcal{E}^c] \leq \alpha/20$ by condition (6.2). We now bound the first term.

$$
\mathop{\mathbb{P}}_{(X,y) \sim_u T_i'}[\mathcal{E}] = \mathop{\mathbb{P}}_{(X,y) \sim_u T_i'}[|y - X^\top \beta_i - \gamma' v^\top X| \leq \sigma t],
$$

which is less than $\alpha/20$ by the condition (6.3). This completes the proof of the claim. □

We use this to show that there cannot exist a γ-packing of size $k \geq 4/\alpha$. To see this, assume that $k = 4/\alpha$, then

$$
|T| \geq \sum_{i=1}^{k} |T_i'| - \sum_{1 \leq i < j \leq k} |T_i' \cap T_j'| \geq \left(1 - \frac{\alpha}{20}(k-1)\right) \sum_{i=1}^{k} |T_i'| \geq \frac{4}{5} k\alpha \frac{|T|}{2} > |T| \,.
$$

This yields a contradiction, completing the proof of Theorem 6.2.1. □

6.2.2 Information-Theoretic Lower Bound on Error

We establish the following lower bound on the error of any list-decoding algorithm for linear regression.

Theorem 6.2.4. *Let $0 < \alpha < 1/2$, $\sigma > 0$, $k > 1$ such that $k = O(1/(\alpha^2 \log(1/\alpha)))$, and $d \in \mathbb{Z}_+$ such that $d > (\log(1/\alpha^k))^C$, where C is a sufficiently large constant. Any list-decodable algorithm that receives a $(1-\alpha)$-corrupted version of D_β (defined in Definition 6.1.2) for some unknown $\beta \in \mathbb{R}^d$, and returns a list $\mathcal{L}$ of size $|\mathcal{L}| = O((1/\alpha)^k)$ has error bound $\Omega\left(\frac{\sigma}{\alpha\sqrt{k \log(1/\alpha)}}\right)$ with high probability.*

Proof. Let $\rho > 0$ to be decided later. We will take β to be of the form ρv for some unit vector v. By abusing notation, let $D_v(x, y)$ be the joint distribution on (X, y) from the linear model $X \sim \mathcal{N}(0, I_d)$, $y = \beta^\top X + \eta$, where $\eta \sim \mathcal{N}(0, \sigma^2)$ independently of X and $\beta = \rho v$. As d is large enough, let S' be a subset of the set S of nearly orthogonal unit vectors of $\mathbb{R}^d$ from Lemma 6.1.15 with $|S'| = \lfloor 0.5(1/\alpha)^k \rfloor$ for $k > 1$. Consider the set of distributions $\{D_v\}_{v \in S'}$ and note that for every distinct pair $u, v \in S$ we have that $\|\rho u - \rho v\|_2 \geq c\rho$ for some $c > 0$. We want to show that after adding $(1 - \alpha)$-fraction of outliers these distributions become indistinguishable, i.e., there exists some distribution that is pointwise greater than αD_v for every $v \in S'$. This will lead to a lower bound on error of the form $\Omega(\rho)$. Let P be the joint pseudo-distribution on (X, y) such that $P(x, y) = \max_{v \in S} D_v(x, y)$ and denote by $\|P\|_1$ the normalizing factor $\int_{\mathbb{R}} \int_{\mathbb{R}^d} P(x, y) \mathrm{dxdy}$. We will show that $P/\|P\|_1 \geq \alpha D_v$ pointwise. To this end, it suffices to show that $\|P\|_1 \leq 1/\alpha$. Denote $z := v^\top x$. Noting that D_v's marginal on x is $\mathcal{N}(0, I_d)$ and the conditional $D_v(y|x)$ is $\mathcal{N}(\rho z, \sigma^2)$ we can write

$$\begin{aligned} D_v(x, y) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y - \rho z|^2}{2\sigma^2}\right) \frac{1}{(\sqrt{2\pi})^d} \exp\left(-\frac{\|x\|^2}{2}\right) \\ &= \frac{1}{(\sqrt{2\pi})^{d+1}\sigma} \exp\left(-\frac{|y - \rho z|^2}{2\sigma^2} - \frac{\|x\|^2}{2}\right). \end{aligned}$$

For some σ_1 to be defined later, take R to be the reference distribution where $X \sim \mathcal{N}(0, I_d)$ and $y \sim \mathcal{N}(0, \sigma_1^2)$ independently. We now calculate the ratio of density of R with D_v at arbitrary (x, y):

$$\begin{aligned}
\frac{R(x,y)}{D_v(x,y)} &= \frac{R(y)R(x|y)}{D_v(y)D_v(x|y)} \\
&= \frac{\frac{1}{(\sqrt{2\pi})^{d+1}\sigma_1}\exp\left(-0.5\|x\|^2 - 0.5y^2/\sigma_1^2\right)}{\frac{1}{(\sqrt{2\pi})^{d+1}\sigma}\exp\left(-0.5\|x\|^2 - 0.5\frac{\rho^2}{\sigma^2}\left(z - \frac{y}{\rho}\right)^2\right)} \\
&= \frac{\sigma}{\sigma_1}\exp\left(-\frac{y^2}{2\sigma_1^2} + \frac{\rho^2}{2\sigma^2}\left(z - \frac{y}{\rho}\right)^2\right) \\
&\geq \frac{\sigma}{\sigma_1}\exp\left(-\frac{y^2}{2\sigma_1^2}\right).
\end{aligned}$$

As we will show later, it suffices to show that this expression is greater than 2α with high probability under D_v. As $y \sim \mathcal{N}(0, \sigma_y^2)$ under D_v, with probability $1 - \alpha^{k-1}$, $|y| \leq 10\sqrt{k}\sigma_y\sqrt{\log(1/\alpha)}$. Setting $\sigma_1 = 10\sqrt{k}\sigma_y\sqrt{\log(1/\alpha)}$, we get that with the same probability,

$$\frac{R(x,y)}{D_v(x,y)} \geq \frac{1}{100\sqrt{k}}\frac{\sigma}{\sigma_y\sqrt{\log(1/\alpha)}}.$$

We can now try to maximize ρ (and thus σ_y) so that the expression on the right-hand side is greater than 2α. This holds as long as ρ satisfies the following:

$$\sigma_y^2 = \sigma^2 + \rho^2 \leq \frac{\sigma^2}{C'k\alpha^2\log(1/\alpha)},$$

As $k = O(1/(\alpha^2\log(1/\alpha)))$, the condition above shows that ρ can be as large as $\frac{\sigma}{\alpha\sqrt{k\log(1/\alpha)}}$

up to constants. Finally we show that $\|P\|_1$ is less than $1/\alpha$ as follows:

$$\begin{aligned}
\|P\|_1 &= \int_{\mathbb{R}} \int_{\mathbb{R}^d} P(x,y)\mathrm{dxdy} \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}^d} P(x,y)\,\mathbb{I}(|y| \le 10\sqrt{k}\sigma_y\sqrt{\log(1/\alpha)})\mathrm{dxdy} \\
&\quad + \int_{\mathbb{R}} \int_{\mathbb{R}^d} P(x,y)\,\mathbb{I}(|y| > 10\sqrt{k}\sigma_y\sqrt{\log(1/\alpha)})\mathrm{dxdy} \\
&\le \frac{1}{2\alpha} \int_{\mathbb{R}} \int_{\mathbb{R}^d} R(x,y)\mathrm{dxdy} + \int_{\mathbb{R}} \int_{\mathbb{R}^d} P(x,y)\,\mathbb{I}(|y| > 10\sqrt{k}\sigma_y\sqrt{\log(1/\alpha)})\mathrm{dxdy} \\
&\le \frac{1}{2\alpha} + \sum_{v \in S'} \mathbb{P}_{(X,y)\sim D_v}\left[|y| > 10\sqrt{k}\sigma_y\sqrt{\log(1/\alpha)}\right] \\
&\le \frac{1}{2\alpha} + |S'|\alpha^{k-1} \le 1/\alpha,
\end{aligned}$$

where the last inequality follows by noting that $|S'| \le 0.5(1/\alpha)^k$. □

6.3 Main Result: Proof of Theorem 6.1.5

In this section, we present the main result of this paper: SQ hardness of list-decodable linear regression (Definitions 6.1.1 and 6.1.2). We consider the setting when β has norm less than 1, i.e., $\beta = \rho v$ for $v \in \mathcal{S}^{d-1}$ and $\rho \in (0,1)$.[9] Note that the marginal distribution of the labels is $\mathcal{N}(0, \sigma_y^2)$, where $\sigma_y^2 = \rho^2 + \sigma^2$. We ensure that the labels y have unit variance by using $\sigma^2 = 1 - \rho^2$. Specifically, the choice of parameters will be such that obtaining a $\rho/2$-additive approximation of the regressor β is possible information-theoretically with $\mathrm{poly}(d/\alpha)$ samples (cf. Section 6.2.1), but the complexity of any SQ algorithm for the task must necessarily be at least $d^{\mathrm{poly}(1/\alpha)}/\sigma$. We show the following more detailed statement of Theorem 6.1.5.

Theorem 6.3.1 (SQ Lower Bound). *Let $c \in (0, 1/2)$, $d \in \mathbb{Z}_+$ with $d = 2^{\Omega(1/(1/2-c))}$, $\alpha \in (0, 1/2)$, $\rho \in (0,1)$, $\sigma^2 = 1 - \rho^2$, and $m \in \mathbb{Z}_+$ with $m \le c_1/\sqrt{\alpha}$ for some sufficiently small constant $c_1 > 0$. Any list-decoding algorithm that, given statistical query access to a $(1-\alpha)$-*

[9] This is a standard assumption and considered by existing works [KKK19; RY20a] (cf. Remark 6.3.11).

corrupted version of the distribution described by the model of Definition 6.1.2 with $\beta = \rho v$ for $v \in \mathcal{S}^{d-1}$, returns a list $\mathcal{L}$ of hypotheses vectors that contains a $\widehat{\beta}$ such that $\|\widehat{\beta} - \beta\|_2 \leq \rho/2$, does one of the following:

- *it uses at least one query to* $\mathrm{STAT}\left(\Omega(d)^{-(2m+1)(1/4-c/2)} e^{O(m)}/\sigma\right)$,
- *it makes* $2^{\Omega(d^c)} d^{-(2m+1)(1/2-c)}$ *many queries, or*
- *it returns a list $\mathcal{L}$ of size* $2^{\Omega(d^c)}$.

In the rest of this section, we will explain the hard-to-learn construction for our SQ lower bound, i.e., a set of distributions with large statistical dimension. The proof would then follow from Lemma 6.1.12. We begin by describing additional notation that we will use.

Notation: As $\beta = \rho v$ for a fixed ρ, we will slightly abuse notation by using $D_v(x, y)$ to denote the joint distribution of the inliers and we use $E_v(x, y)$ to denote the $(1-\alpha)$-corrupted version of $D_v(x, y)$. To avoid using multiple subscripts, we use $D_v(x|y)$ to denote the conditional distribution of $X|y$ according to the distribution D_v and similarly for the other distributions. In addition, we use $D_v(y)$ to denote the marginal distribution of y under D_v and similarly for other distributions.

Following the general construction of [DKS17], we will specify a *reference* joint distribution $R(x, y)$ where X and y are independent, and $X \sim \mathcal{N}(0, I_d)$. We will find a marginal distribution $R(y)$ and a way to add the outliers so that the following hold for each E_v (where $m = \Theta(1/\sqrt{\alpha})$):

(I) E_v is indeed a valid distribution of (X, y) in our corruption model (i.e., can be written as a mixture $\alpha D_v(x, y) + (1-\alpha) N_v(x, y)$ for some noise distribution N_v). Moreover, the marginal of E_v on the labels, $E_v(y)$, coincides with $R(y)$.

(II) For every $y \in \mathbb{R}$, the conditional distribution $E_v(x|y)$ is of the form $P_{A_y,v}$ of Definition 6.1.13, with A_y being a distribution that matches the first $2m$ moments with $\mathcal{N}(0,1)$.[10]

(III) For A_y defined above, $\mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]$ is bounded.

We first briefly explain why a construction satisfying the above properties suffices to prove our main theorem (postponing a formal proof for the end of this section). We start by noting the following decomposition.

Lemma 6.3.2. *For $u, v \in \mathcal{S}^{d-1}$, if E_u and E_v have the same marginals $R(y)$ on the labels, they satisfy* $\chi_{R(x,y)}(E_v(x,y), E_u(x,y)) = \mathbb{E}_{y\sim R(y)}\left[\chi_{\mathcal{N}(0,I_d)}\left(E_v(x|y), E_u(x|y)\right)\right]$.

Proof. Let ϕ denote the density of $\mathcal{N}(0,1)$. Using the fact that E_v and E_u have the same marginal $R(y)$ we have that

$$\begin{aligned}
\chi_{R(x,y)}(E_v(x,y), E_u(x,y)) + 1 &= \int_{\mathbb{R}}\int_{\mathbb{R}^d} \frac{E_v(x,y)E_u(x,y)}{\phi(x)R(y)}\mathrm{dxdy} \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}^d} \frac{E_v(x|y)E_u(x|y)}{\phi(x)}R(y)\mathrm{dxdy} \\
&= \int_{\mathbb{R}} \left(1 + \chi_{\mathcal{N}(0,I_d)}(E_v(x|y)E_u(x|y))\right) R(y)\mathrm{dy} \\
&= 1 + \mathop{\mathbb{E}}_{y\sim R(y)}\left[\chi_{\mathcal{N}(0,I_d)}\left(E_v(x|y), E_u(x|y)\right)\right]. \qquad \square
\end{aligned}$$

Using the decomposition in Lemma 6.3.2 for E_u and E_v satisfying Property (II), Lemma 6.1.14 implies that $|\chi_{R(x,y)}(E_v(x,y), E_u(x,y))| \le |u^\top v|^{2m+1}\,\mathbb{E}_{y\sim R(y)}[\chi^2(A_y, N(0,1))]$. Letting $\mathcal{D} = \{E_v : v \in S\}$, where S is the set of nearly uncorrelated unit vectors from Lemma 6.1.15, we get that $\mathcal{D}$ is (γ, b)-correlated relative to R, for $b = \mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]$ and $\gamma \le d^{-\Omega(m)}$b. As $|S| = 2^{\Omega(d^c)}$, b is bounded by Property (III), and the list size is much smaller than $|S|$, we can show that the statistical dimension of the list-decodable linear regression is large.

[10] We use even number of moments for simplicity. The analysis would slightly differ for odd number.

Thus, in the rest of the section we focus on showing that such a construction exists. We first note that according to our linear model of Definition 6.1.2, the conditional distribution of X given y for the inliers is Gaussian with unit variance in all but one direction.

Fact 6.3.3. *Fix $\rho > 0$, $v \in \mathcal{S}^{d-1}$, and consider the regression model of Definition 6.1.2 with $\beta = \rho v$. Then the conditional distribution $X|y$ of the inliers is $\mathcal{N}(y\rho v, I_d - \rho^2 vv^\top)$, i.e., independent standard Gaussian in all directions perpendicular to v and $\mathcal{N}(\rho y, 1-\rho^2)$ in the direction of v.*

Proof. This is due to the following fact for the conditional distribution of the Gaussian distribution.

Fact 6.3.4. *If* $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$, *then* $y_1|y_2 \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$, *with* $\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2)$ *and* $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

We apply this fact for the pair (X, y) by setting $y_1 = X, y_2 = y, \mu_1 = \mu_2 = 0$ and $\Sigma_{11} = I_d, \Sigma_{12} = \beta, \Sigma_{21} = \beta^\top, \Sigma_{22} = \sigma^2 + \|\beta\|_2^2$. □

Since Fact 6.3.3 states that $D_v(x|y)$ is already of the desired form (standard normal in all directions perpendicular to v and $\mathcal{N}(y\rho, 1-\rho^2)$ in the direction of v), the problem becomes one-dimensional. More specifically, for every $y \in \mathbb{R}$, we need to find a one-dimensional distribution Q_y and appropriate values $\alpha_y \in [0, 1]$ such that the mixture $A_y = \alpha_y \mathcal{N}(y\rho, 1-\rho^2) + (1-\alpha_y)Q_y$ matches the first $2m$ moments with $\mathcal{N}(0, 1)$. Then, multiplying by $\phi_{\perp v}$ (which denotes the contribution of the space orthogonal to v to the density of standard Gaussian, as defined in Definition 6.1.13) yields the d-dimensional mixture distribution $\alpha_y D_v(x|y) + (1-\alpha_y)Q_y(v^\top x)\phi_{\perp v}(x)$. We show that an appropriate selection of α_y can ensure that this is a valid distribution for our contamination model.

Lemma 6.3.5. *Let R be a distribution on pairs $(x, y) \in \mathbb{R}^{d+1}$ such that $\alpha_y := \alpha D_v(y)/R(y) \in [0, 1]$ for all $y \in \mathbb{R}$. Suppose that for every $y \in \mathbb{R}$ there exists a univariate distribution Q_y such that $A_y := \alpha_y \mathcal{N}(y\rho, 1-\rho^2) + (1-\alpha_y)Q_y$ matches the first $2m$ moments with $\mathcal{N}(0, 1)$. If the distribution of the outliers is $N_v(x, y) = ((1-\alpha_y)/(1-\alpha))Q_y(v^\top x)\phi_{\perp v}(x)R(y)$, Properties (I) and (II) hold.*

Proof. First note that the noise distribution N_v is indeed a valid distribution since it is non-negative everywhere because of the assumption $\alpha_y \in [0, 1]$ and it integrates to one:

$$\begin{aligned}
&\frac{1}{1-\alpha}\int_{\mathbb{R}}\int_{\mathbb{R}^d}(1-\alpha_y)Q_y(v^\top x)\phi_{\perp v}(x)R(y)\mathrm{dxdy} \\
&\quad= \frac{1}{1-\alpha}\left(\int_{\mathbb{R}}\int_{\mathbb{R}^d} R(y)Q_y(v^\top x)\phi_{\perp v}(x)\mathrm{dxdy} - \alpha\int_{\mathbb{R}}\int_{\mathbb{R}^d} D_v(y)Q_y(v^\top x)\phi_{\perp v}(x)\mathrm{dxdy}\right) \\
&\quad= 1\ .
\end{aligned}$$

The joint distribution $E_v(x, y)$ can be written as

$$\begin{aligned}
E_v(x, y) &= \alpha D_v(x, y) + (1-\alpha)N_v(x, y) \\
&= \alpha D_v(x, y) + (1-\alpha)\frac{1-\alpha_y}{1-\alpha}Q_y(v^\top x)\phi_{\perp v}(x)R(y) \\
&= \left(\alpha_y D_v(x|y) + (1-\alpha_y)Q_y(v^\top x)\phi_{\perp v}(x)\right)R(y)\ .
\end{aligned}$$

This means that the marginal of y under E_v is $R(y)$, which establishes Property (I), and the conditional distribution of $X|y$ under E_v is $E_y(x|y) = \alpha_y D_v(x|y) + (1-\alpha_y)Q_y(v^\top x)\phi_{\perp v}(x)$.

The moment matching part of Property (II) holds by assumption. For the other part of Property (II), we note that $E_v(x|y)$ is standard Gaussian in directions perpendicular to v because of Fact 6.3.3 and the form of the term $Q_y(v^\top x)\phi_{\perp v}(x)$ that corresponds to the outliers. □

We will choose the reference distribution $R(x, y)$ to have $X \sim \mathcal{N}(0, I_d)$ and $y \sim$

$\mathcal{N}(0, 1/\alpha)$ independently, which makes the corresponding value of α_y to be

$$\alpha_y = \alpha D_v(y)/R(y) = \sqrt{\alpha}\exp(-y^2(1-\alpha)/2)\,.$$

This satisfies the condition in Lemma 6.3.5 that $\alpha_y \in [0, 1]$. Our choice of $R(y)$ being $\mathcal{N}(0, 1/\alpha)$ is informed by Properties (II) and (III), and will be used later on in the proofs of Theorem 6.3.6 and Lemma 6.3.7 (also see the last paragraph of Section 6.1.3 for more intuition). Going back to our goal, i.e., making $A_y = \alpha_y\mathcal{N}(y\rho, 1-\rho^2) + (1-\alpha_y)Q_y$ match moments with $\mathcal{N}(0, 1)$, we will argue that it suffices to only look for Q_y of the specific form $U_\rho F_y$, where U_ρ is the Ornstein-Uhlenbeck operator. This suffices because $U_\rho\delta_y = \mathcal{N}(y\rho, 1-\rho^2)$ and the operator U_ρ preserves the moments of a distribution if they match with $\mathcal{N}(0, 1)$ (see Lemma 6.3.7 (i) below). Letting $A_y = U_\rho(\alpha_y\delta_y + (1-\alpha_y)F_y)$, the new goal is to show that the argument of U_ρ matches moments with $\mathcal{N}(0, 1)$. We show the following structural result:

Theorem 6.3.6. *Let $y \in \mathbb{R}$, $B \in \mathbb{R}$, $\alpha \in (0, 1/2)$, and define $\alpha_y := \sqrt{\alpha}\exp(-y^2(1-\alpha)/2)$. For any $m \in \mathbb{Z}_+$ such that $m \le C_1/\sqrt{\alpha}$ and $B \ge C_2\sqrt{m}$, with $C_1 > 0$ being a sufficiently small constant and C_2 being a sufficiently large constant, there exists a distribution F_y that satisfies the following:*

1. *The mixture distribution $\alpha_y\delta_y + (1-\alpha_y)F_y$ matches the first $2m$ moments with $\mathcal{N}(0, 1)$.*
2. *F_y is a discrete distribution supported on at most $2m+1$ points, all of which lie in $[-B, B]$.*

The proof of Theorem 6.3.6 is the bulk of the technical work of this paper and is deferred to Section 6.4. As mentioned before, applying U_ρ preserves the required moment-matching property. More crucially, it allows us to bound the χ^2-divergence: the following result bounds $\chi^2(A_y, \mathcal{N}(0, 1))$ using contraction properties of U_ρ, tail bounds of Hermite polynomials, and the discreteness of F_y.

Lemma 6.3.7. *In the setting of Theorem 6.3.6, let $\rho > 0$ and $Q_y = U_\rho F_y$. Then the following holds for the mixture $A_y = \alpha_y \mathcal{N}(y\rho, 1-\rho^2) + (1-\alpha_y)Q_y$: (i) A_y matches the first $2m$ moments with $\mathcal{N}(0,1)$, and (ii) $\chi^2(A_y, \mathcal{N}(0,1)) \le \alpha O(e^{y^2(\alpha-1/2)})/(1-\rho^2) + O(e^{B^2/2})/(1-\rho^2)$.*

Proof. The first property follows by noting that $A_y = \alpha_y \mathcal{N}(y\rho, 1-\rho^2) + (1-\alpha_y)Q_y = U_\rho(\alpha_y \delta_y + (1-\alpha_y)F_y)$ and using the eigenvalue property of Hermite polynomials (Fact 6.1.7). This implies that for all $i \le 2m$ we have that

$$\mathop{\mathbb{E}}_{X \sim U_\rho(\alpha_y \delta_y + (1-\alpha_y)F_y)}[h_i(X)] = \rho^i \mathop{\mathbb{E}}_{X \sim \alpha_y \delta_y + (1-\alpha_y)F_y}[h_i(X)] = \rho^i \mathop{\mathbb{E}}_{X \sim \mathcal{N}(0,1)}[h_i(X)] = \mathop{\mathbb{E}}_{X \sim \mathcal{N}(0,1)}[h_i(X)],$$

where the last equation uses that $\mathbb{E}_{X \sim \mathcal{N}(0,1)}[h_i(X)] = 0$ for $i > 0$ and $\mathbb{E}_{X \sim \mathcal{N}(0,1)}[h_0(X)] = 1$. Since $\{h_i : i \in [2m]\}$ form a basis of $\mathcal{P}(2m)$, the space of all polynomials of degree at most $2m$, it follows that A_y continues to matches $2m$ moments with $\mathcal{N}(0,1)$.

The χ^2 bound is due to the bounded support in $[-B, B]$ and the Gaussian smoothing operation and can be shown as follows. First, we need the following fact whose proof is included in Appendix D.1 for completeness.

Fact 6.3.8. *For any one-dimensional distribution P that matches the first m moments with $\mathcal{N}(0,1)$ and has $\chi^2(P, \mathcal{N}(0,1)) < \infty$ the following identity is true:*

$$\chi^2(P, \mathcal{N}(0,1)) = \sum_{i=m+1}^{\infty} \left(\mathop{\mathbb{E}}_{X \sim P}[h_i(X)] \right)^2 .$$

Let M_y denote the distribution $\alpha_y \delta_y + (1-\alpha_y)F_y$, i.e., the mixture before applying the Ornstein-Uhlenbeck operator. In order to apply Fact 6.3.8 to M_y, we need to argue that its χ^2-divergence is finite. As F_y is a discrete distribution, the U_ρ operator will transform it to a finite sum of Gaussians with variances strictly less than 2. We defer the proof of the following claim to Appendix D.1.

Claim 6.3.9. *If $P = \sum_{i=1}^{k} \lambda_i N(\mu_i, \sigma_i^2)$ with $\mu_i \in \mathbb{R}$, $\sigma_i < \sqrt{2}$ and $\lambda_i \geq 0$ such that $\sum_{i=1}^{k} \lambda_i = 1$, we have that $\chi^2(P, \mathcal{N}(0,1)) < \infty$.*

Using the formula of Fact 6.3.8 and Fact 6.1.7 for the individual terms, we get that

$$\begin{aligned}
\chi^2(A_y, \mathcal{N}(0,1)) &= \sum_{i=2m+1}^{\infty} \mathop{\mathbb{E}}_{X \sim U_\rho M_y} [h_i(X)]^2 = \sum_{i=2m+1}^{\infty} \rho^{2i} \mathop{\mathbb{E}}_{X \sim M_y} [h_i(X)]^2 \\
&= \sum_{i=2m+1}^{\infty} \rho^{2i} \left(\alpha_y h_i(y) + (1-\alpha_y) \mathop{\mathbb{E}}_{x \sim F_y} [h_i(X)] \right)^2 \\
&\leq 2\alpha_y^2 \sum_{i=2m+1}^{\infty} \rho^{2i} h_i^2(y) + 2(1-\alpha_y)^2 \sum_{i=2m+1}^{\infty} \rho^{2i} \mathop{\mathbb{E}}_{x \sim F_y} [h_i(X)]^2 \,, \qquad (6.4)
\end{aligned}$$

where the inequality uses that $(a+b)^2 \leq 2(a^2+b^2)$ for all $a, b \in \mathbb{R}$. To bound this expression from above we will use the following tail bound for Hermite polynomials.

Lemma 6.3.10 ([Kra04])**.** *Let h_i be the i-th normalized probabilist's Hermite polynomial. Then $\max_{x \in \mathbb{R}} h_k^2(x) e^{-x^2/2} = O(k^{-1/6})$.*

More details on how Lemma 6.3.10 follows from the result of [Kra04] can be found in Section 6.4.3. For the first term of Equation (6.4), we have that

$$\begin{aligned}
\sum_{i=2m+1}^{\infty} \rho^{2i} \alpha_y^2 h_i^2(y) &\leq \sum_{i=2m+1}^{\infty} \rho^{2i} \alpha e^{-y^2 + \alpha y^2} O(e^{y^2/2}) \\
&\leq \alpha O(e^{y^2(\alpha - 1/2)}) \sum_{i=2m+1}^{\infty} \rho^{2i} \\
&\leq \alpha O(e^{y^2(\alpha-1/2)}) \rho^{2(2m+1)} / (1-\rho^2) \,,
\end{aligned}$$

where the first inequality uses Lemma 6.3.10 and the definition of α_y. For the second term, we use the bounded support of F_y in $[-B, B]$ along with the bound of Lemma 6.3.10 to obtain

$$\sum_{i=2m+1}^{\infty} \rho^{2i} \mathop{\mathbb{E}}_{x \sim F_y} [h_i(X)]^2 \leq \sum_{i=2m+1}^{\infty} \rho^{2i} O(e^{B^2/2}) \leq O(e^{B^2/2}) \sum_{i=2m+1}^{\infty} \rho^{2i} \leq O(e^{B^2/2}) \frac{\rho^{2(2m+1)}}{1-\rho^2} \,.$$

This completes the proof of Lemma 6.3.7. □

Putting everything together, we now prove our main theorem.

Proof of Theorem 6.3.1. We will show that the following search problem $\mathcal{Z}$ has large statistical dimension: $\mathcal{D}$ is the set of distributions of the form $E_v(x,y) = \alpha D_v(x,y) + (1-\alpha)N_v(x,y)$ for every $v \in \mathcal{S}^{d-1}$ and noise distribution N_v as in Lemma 6.3.5. The reference distribution R is $R = \mathcal{N}(0, I_d) \times \mathcal{N}(0, 1/\alpha)$. Let $\beta(v) = \rho v$ denote the regression vector corresponding to E_v. The set of solutions $\mathcal{F}$ is the set of all lists of size ℓ containing vectors of norm ρ in $\mathbb{R}^d$ and the solution set $\mathcal{Z}(E_v)$ for the distribution E_v is exactly the set of lists from $\mathcal{F}$ having at least one element u at distance $\|u - \beta(v)\|_2 \leq \rho/2$. The appropriate subset of $\mathcal{D}$ that we will consider is the one corresponding to the set S of nearly orthogonal vectors of Lemma 6.1.15, $\mathcal{D}_R = \{E_v\}_{v \in S}$.

Note that for any $u \in \mathcal{F}$, there exists at most one element E_v in $\mathcal{D}_R$ that satisfies $\|u - \beta(v)\|_2 \leq \rho/2$, since if there exists another v' with $\|u - \beta(v')\|_2 \leq \rho/2$, then by triangle inequality $\|\beta(v) - \beta(v')\|_2 \leq \rho$. However, this cannot happen because $|v^\top(v')| \leq O(d^{c-1/2})$ for all $v, v' \in S$ together with $d = 2^{\Omega(1/(1/2-c))}$ implies that $\|\beta(v) - \beta(v')\|_2 \geq \rho\sqrt{2(1 - v^\top(v'))} \geq \rho$. This implies that for any solution list L, $|\mathcal{D}_R \setminus \mathcal{Z}^{-1}(L)| \geq |\mathcal{D}_R| - \ell$. We choose $\ell = |\mathcal{D}_R|/2$. We now calculate the pairwise correlation of the set $\mathcal{D}_R$. Let a pair of $u, v \in \mathcal{S}^{d-1}$.

$$
\begin{aligned}
&\chi_{R(x,y)}(E_v(x,y), E_u(x,y)) \\
&\qquad = \mathop{\mathbb{E}}_{y \sim R(y)} \left[\chi_{\mathcal{N}(0,I_d)} \left(E_v(x|y), E_u(x|y) \right) \right] \\
&\qquad \leq |u^\top v|^{2m+1} \mathop{\mathbb{E}}_{y \sim R(y)} \left[\chi^2(A_y, \mathcal{N}(0,1)) \right] \\
&\qquad = |u^\top v|^{2m+1} \left(O(e^{B^2/2})/(1-\rho^2) + \int_{\mathbb{R}} \alpha O(e^{y^2(\alpha - 1/2)}) \sqrt{\alpha} e^{-y^2 \alpha} \mathrm{dy} \right) \\
&\qquad \leq |u^\top v|^{2m+1} O(e^{B^2/2})/(1-\rho^2) \\
&\qquad \leq \Omega(d)^{-(2m+1)(1/2-c)} O(e^{B^2/2})/(1-\rho^2) \,,
\end{aligned}
$$

where the first line is due to Lemma 6.3.2, the second line is from Lemma 6.1.14 along with the observation that $E_v(x|y)$ is of the form $P_{A_y,v}$, the third line comes from the second part of Lemma 6.3.7, and the last one uses Lemma 6.1.15. Thus, by recalling that we can choose $B = C_2\sqrt{m}$ for a sufficiently large constant C_2, the set $\mathcal{D}_R$ is (γ, b)-correlated with respect to R, where $\gamma := \Omega(d)^{-(2m+1)(1/2-c)}e^{O(m)}/(1-\rho^2)$ and $b := e^{O(m)}/(1-\rho^2)$. The proof is concluded by applying Lemma 6.1.12 with $\gamma' = \gamma$. □

We conclude this section with a note on the model and existing algorithmic results (extending the relevant discussion of Section 6.1.1).

Remark 6.3.11 (Comparison of SQ Lower Bound to Existing Upper Bounds)**.** *We remark that the model used in Theorem 6.1.5 (i.e., having a regressor with norm at most one and additive noise with small variance) is considered in both recent works [KKK19; RY20a] that provided list-decoding algorithms for the problem. In particular, these works give the following upper bounds:*

- *[KKK19] considers the model where $\|\beta\|_2 \le 1$ and gives an algorithm that for every $\epsilon > 0$, runs in time $(d/(\alpha\epsilon))^{O\left(\frac{1}{\alpha^8\epsilon^8}\right)}$ and outputs a list of size $O(1/\alpha)$ containing a $\widehat{\beta}$ such that $\|\widehat{\beta} - \beta\|_2 \le O(\sigma/\alpha) + \epsilon$. Note that this guarantee is better than the trivial upper bound of 1 only if $\sigma = O(\alpha)$. To achieve error $1/4$, this algorithm runs in time $(d/\alpha)^{O\left(\frac{1}{\alpha^8}\right)}$. On the other hand, our lower bound for the complexity of any SQ algorithm becomes $\alpha d^{\Omega\left(1/\sqrt{\alpha}\right)}$.*

- *[RY20a] does not impose any constraint on $\|\beta\|_2$ and gives an algorithm that runs in time $(\|\beta\|_2/\sigma)^{\log(1/\alpha)}d^{O(1/\alpha^4)}$ and outputs a list of size $O((\|\beta\|_2/\sigma)^{\log(1/\alpha)})$ including a $\widehat{\beta}$ with the guarantee that $\|\widehat{\beta} - \beta\|_2 \le O(\sigma/\alpha^{3/2})$. For the special case where $\|\beta\|_2 \le 1$ (and $\sigma = O(\alpha^{3/2})$ in order for the error guarantee to be meaningful), this algorithm can achieve error $1/4$ in time $(1/\alpha^{3/2})^{\log(1/\alpha)}d^{O(1/\alpha^4)}$. In comparison, our lower bound becomes $\alpha^{3/2}d^{\Omega\left(1/\sqrt{\alpha}\right)}$.*

6.4 Duality for Moment Matching: Proof of Theorem 6.3.6

We now prove the existence of a bounded distribution F_y such that the mixture $\alpha_y\delta_y + (1-\alpha_y)F_y$ matches the first $2m$ moments with $\mathcal{N}(0,1)$. The proof follows a non-constructive argument based on the duality between the space of moments and the space of non-negative polynomials.

Let $B > 0$ and $m \in \mathbb{Z}_+$. Let $\mathcal{P}(m)$ denote the class of all polynomials $p : \mathbb{R} \to \mathbb{R}$ with degree at most m. Let $\mathcal{P}^{\geq 0}(2m, B)$ be the class of polynomials that can be represented in either the form $p(t) = (\sum_{i=0}^{m} a_i t^i)^2$ or the form $p(t) = (B^2 - t^2)(\sum_{i=0}^{m-1} b_i t^i)^2$. The intuition for $\mathcal{P}^{\geq 0}(2m, B)$ is that every polynomial of degree at most $2m$ that is non-negative in $[-B, B]$ can be written as a finite sum of polynomials from $\mathcal{P}^{\geq 0}(2m, B)$. By slightly abusing notation, for a polynomial $p(t) = \sum_{i=0}^{m} p_i t^i$, we also use p to denote the vector in $\mathbb{R}^{m+1}$ consisting of the coefficients $(p_0, \ldots, p_m)$. The following classical result characterizes when a vector is realizable as the moment sequence of a distribution with support in $[-B, B]$ (for simplicity, we focus on matching an even number of moments in the rest of this section).

Theorem 6.4.1 (Theorem 16.1 of [KS53])**.** *Let $B > 0$, $k \in \mathbb{Z}_+$, and $x = (x_0, x_1, \ldots, x_{2k}) \in \mathbb{R}^{2k+1}$ with $x_0 = 1$. There exists a distribution with support in $[-B, B]$ having as its first $2k$ moments the sequence $(x_1, \ldots, x_{2k})$ if and only if for all $p \in \mathcal{P}^{\geq 0}(2k, B)$ it holds that $\sum_{i=0}^{2k} x_i p_i \geq 0$.*

As we require the distribution to be discrete, we prove the following result using Theorem 6.4.1:

Proposition 6.4.2. *Fix $y \in \mathbb{R}$, $\alpha_y \in (0,1)$, $B > 0$, and $m \in \mathbb{Z}_+$. There exists a discrete distribution F_y supported on at most $2m+1$ points in $[-B, B]$ such that $\alpha_y\delta_y + (1-\alpha_y)F_y$*

matches the first $2m$ *moments with* $\mathcal{N}(0,1)$ *if and only if* $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[p(X)] \geq \alpha_y p(y)$ *for all* $p \in \mathcal{P}^{\geq 0}(2m, B)$.

The proof of Proposition 6.4.2 is deferred to Section 6.4.1. To prove Theorem 6.3.6, we need to establish the condition of Proposition 6.4.2. To this end, we first need the following two technical lemmas, whose proofs are given in Sections 6.4.2 and 6.4.3.

Lemma 6.4.3. *Let* $m \in \mathbb{Z}_+$. *If* $B \geq C\sqrt{m}$ *for some sufficiently large constant* $C > 0$, *then for every* $q \in \mathcal{P}(m)$, *it holds that* $B^2\,\mathbb{E}_{X\sim\mathcal{N}(0,1)}[q^2(X)] \geq 2\,\mathbb{E}_{X\sim\mathcal{N}(0,1)}[X^2 q^2(X)]$.

Lemma 6.4.4. *Let* $y \in \mathbb{R}$, $\alpha \in (0, 1/2)$, $m \in \mathbb{Z}_+$, *and* $\alpha_y = \sqrt{\alpha}\exp(-y^2(1-\alpha)/2)$. *Suppose* $m \leq C/\sqrt{\alpha}$ *for some sufficiently small constant* $C > 0$. *Then for all* $r \in \mathcal{P}(m), r \not\equiv 0$: $r^2(y)/(\mathbb{E}_{X\sim\mathcal{N}(0,1)}[r^2(X)]) \leq 1/(2\alpha_y)$.

Proof of Theorem 6.3.6. By Proposition 6.4.2, it remains to show that if $B \geq C_2\sqrt{m}$, then the condition $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[p(X)] \geq \alpha_y p(y)$ holds for all $p \in \mathcal{P}^{\geq 0}(2m, B)$. Thus, it suffices to ensure that the following two inequalities hold for $X \sim \mathcal{N}(0,1)$:

$$\sup_{r\in\mathcal{P}(m),r\not\equiv 0} \frac{r^2(y)}{\mathbb{E}[r^2(X)]} \leq \frac{1}{\alpha_y} \quad \text{and} \quad \sup_{q\in\mathcal{P}(m-1),q\not\equiv 0} \frac{(B^2-y^2)q^2(y)}{\mathbb{E}[(B^2-X^2)q^2(X)]} \leq \frac{1}{\alpha_y}, \tag{6.5}$$

where we use Lemma 6.4.3 to show that $\mathbb{E}[(B^2 - X^2)q^2(X)] > 0$ for all non-zero polynomials $q \in \mathcal{P}(m-1)$. The first expression can be bounded using Lemma 6.4.4 when $m \leq C_1/\sqrt{\alpha}$. We now focus on the second expression. By Lemma 6.4.3, $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[(B^2 - X^2)q^2(X)] \geq 0.5\,\mathbb{E}_{X\sim\mathcal{N}(0,1)}[B^2 q^2(X)]$. Therefore, we have that

$$\begin{aligned}\sup_{q\in\mathcal{P}(m-1),q\not\equiv 0} \frac{(B^2-y^2)q^2(y)}{\mathbb{E}_{X\sim\mathcal{N}(0,1)}[(B^2-X^2)q^2(X)]} &\leq \sup_{q\in\mathcal{P}(m-1),q\not\equiv 0} \frac{B^2 q^2(y)}{\mathbb{E}_{X\sim\mathcal{N}(0,1)}[(B^2-X^2)q^2(X)]} \\ &\leq \sup_{q\in\mathcal{P}(m-1),q\not\equiv 0} \frac{B^2 q^2(y)}{\mathbb{E}_{X\sim\mathcal{N}(0,1)}[0.5B^2 q^2(X)]} = 2 \sup_{q\in\mathcal{P}(m-1),q\not\equiv 0} \frac{q^2(y)}{\mathbb{E}_{X\sim\mathcal{N}(0,1)}[q^2(X)]},\end{aligned}$$

where the first inequality uses that the denominator is positive and $y^2 q^2(y) \geq 0$ and the second inequality uses that $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[(B^2 - X^2)q^2(X)] \geq 0.5\,\mathbb{E}_{X\sim\mathcal{N}(0,1)}[B^2 q^2(X)]$. The

expression above is of the same form as the first expression in Equation (6.5), and thus is also bounded above by $1/\alpha_y$ when $m \leq C_1/\sqrt{\alpha}$ using Lemma 6.4.4. This completes the proof of Theorem 6.3.6. □

6.4.1 Proof of Proposition 6.4.2

We require the following result stating that for every distribution Q with bounded support, there exists a discrete distribution P with bounded support that matches the low-degree moments of Q.

Lemma 6.4.5. *Let $B > 0$, $k \in \mathbb{Z}_+$, and Q be any distribution with support in $[-B, B]$. Then there exists a discrete distribution P with the following properties: (i) the support of P is contained in $[-B, B]$, (ii) the first k moments of P agree with the first k moments of Q, and (iii) P is supported on at most $k+1$ points.*

Proof. Let $\mathcal{Q}$ be the set of distributions on $\mathbb{R}$ that are supported in $[-B, B]$ and let $\mathcal{Q}' \subset \mathcal{Q}$ be the set of Dirac delta distributions supported in $[-B, B]$, i.e., $\mathcal{Q}' = \{\delta_y : y \in [-B, B]\}$. Let $\mathcal{C} \subset \mathbb{R}^k$ and $\mathcal{C}' \subset \mathbb{R}^k$ be the set of all vectors $(x_1, \ldots, x_k)$ whose coordinates $x_1, \ldots, x_k$ are the moments of a distribution in $\mathcal{Q}$ and $\mathcal{Q}'$ respectively, i.e.,

$$\mathcal{C} := \{x \in \mathbb{R}^k : \exists Q \in \mathcal{Q} : \forall i \in [k], x_i = \mathop{\mathbb{E}}_{X \sim Q}[X^i]\},$$
$$\mathcal{C}' := \{x \in \mathbb{R}^k : \exists Q' \in \mathcal{Q}' : \forall i \in [k], x_i = \mathop{\mathbb{E}}_{X \sim Q'}[X^i]\}.$$

Note that there is a bijection between $\mathcal{C}'$ and $\mathcal{Q}'$. We now recall the following classical result stating convexity properties of $\mathcal{C}$ and its relation with $\mathcal{C}'$. We say a set M is a convex hull of a set M' if every $x \in M$ can be written as $x = \sum_{i=1}^{j} \lambda_i y_i$, where $j \in \mathbb{Z}_+$, $\sum_{i=1}^{j} \lambda_i = 1$, and for all $i \in [j]$: $\lambda_i \geq 0$, $y_i \in M'$.

Lemma 6.4.6 (Theorem 7.2 and 7.3 of [KS53])**.** *$\mathcal{C}$ is convex, closed, and bounded. Moreover, $\mathcal{C}$ is a convex hull of $\mathcal{C}'$.*

Let $x^* = (x_1^*, \ldots, x_k^*)$ be the first k moments of Q. Since $x^* \in \mathcal{C}$, Caratheodory theorem and Lemma 6.4.6 implies that x^* can be written as a convex combination of at most $k+1$ elements of $\mathcal{C}'$. This implies that there is a distribution, which is a convex combination of at most $k+1$ Dirac delta distributions in $\mathcal{Q}'$, that matches the first k moments with x^*. This completes the proof. □

We can now prove the main result of this section.

Proof of Proposition 6.4.2. Let $X \sim \mathcal{N}(0,1)$. We note that F_y should have the moment sequence $x = (x_1, \ldots, x_{2m})$ where $x_i = (\mathbb{E}_{X\sim\mathcal{N}(0,1)}[X^i] - \alpha_y y^i)/(1-\alpha_y)$ for $i \in [2m]$. Theorem 6.4.1 implies that this happens if and only if for all $p = (p_0, \ldots, p_{2m}) \in \mathcal{P}^{\geq 0}(2m, B)$, we have that $\sum_{i=0}^{2m} x_i p_i \geq 0$. The desired expression follows by noting that $\sum_{i=0}^{2m} x_i p_i = (\sum_{i=0}^{2m} p_i \mathbb{E}_{X\sim\mathcal{N}(0,1)}[X^i] - \alpha_y p_i y^i)/(1-\alpha_y) = (\mathbb{E}_{X\sim\mathcal{N}(0,1)}[p(X)] - \alpha_y p(y))/(1-\alpha_y)$. The result that F_y is discrete follows from Lemma 6.4.5. □

6.4.2 Proof of Lemma 6.4.3

The proof of Lemma 6.4.3 is a relatively straightforward application of Hölder's inequality and the Gaussian Hypercontractivity Theorem (stated below). For $p \in (0, \infty)$, we define the L^p-norm of a random variable X to be $\|X\|_{L^p} := (\mathbb{E}[|X|^p])^{1/p}$.

Fact 6.4.7 (Gaussian Hypercontractivity [Bog98; Nel73]). *Let $X \sim \mathcal{N}(0,1)$. If $p \in \mathcal{P}(d)$ and $t \geq 2$, then $\|p(X)\|_{L^t} \leq (t-1)^{d/2} \|p(X)\|_{L^2}$.*

Proof of Lemma 6.4.3. Let $X \sim \mathcal{N}(0,1)$. We can assume that q is a non-zero polynomial. Then it suffices to bound B from above by $\sqrt{2}$ times the following expression:

$$\sup_{q\in\mathcal{P}(m),q\not\equiv 0} \sqrt{\frac{\mathbb{E}[X^2 q^2(X)]}{\mathbb{E}[q^2(X)]}} \leq \sup_{q\in\mathcal{P}(m),q\not\equiv 0} \sqrt{\frac{(\mathbb{E}[(X^2)^{m+1}])^{1/(m+1)} \left(\mathbb{E}[(q^2(X))^{\frac{m+1}{m}}]\right)^{\frac{m}{m+1}}}{\mathbb{E}[q^2(X)]}}$$
$$= \sup_{q\in\mathcal{P}(m),q\not\equiv 0} \frac{\|X\|_{L^{2m+2}} \|q(X)\|_{L^{\frac{2m+2}{m}}}}{\|q(X)\|_{L^2}},$$

where the first step uses Hölder's inequality. Using standard concentration bounds for the standard Gaussian (or Fact 6.4.7 with $p(x) = x$), we get that $\|X\|_{L^{2m+2}} = O(\sqrt{m})$. Gaussian Hypercontractivity (Fact 6.4.7) implies that for any polynomial of degree at most m and $r > 2$, $\|q(X)\|_{L^r} \leq (r-1)^{m/2} \|q(X)\|_{L^2}$. For $r = (2m+2)/m$, we get that

$$\frac{\|q(X)\|_{L^{\frac{2m+2}{m}}}}{\|q(X)\|_{L^2}} \leq \left(\frac{2m+2}{m} - 1\right)^{\frac{m}{2}} = \left(1 + \frac{2}{m}\right)^{\frac{m}{2}} \leq \exp(1).$$

Therefore, $B \geq C\sqrt{m}$ suffices for a sufficiently large constant C. □

6.4.3 Proof of Lemma 6.4.4

We first recall the result on the tails of Hermite polynomials.

Lemma 6.4.8 ([Kra04])**.** *Let h_k be the k-th normalized probabilist's Hermite polynomial. Then* $\max_{x\in\mathbb{R}} h_k^2(x)e^{-x^2/2} = O(k^{-1/6})$.

For completeness, we give an explicit calculation that translates the result of [Kra04] in our setting.

Proof of Lemma 6.4.8. We will split the analysis in two cases. First suppose the case when $k < 6$. As $h_k(\cdot)$ is a constant degree polynomial, we get that $\max_{x\in\mathbb{R}} h_k^2(x)\exp(-x^2/2)$ is a constant. For the rest of the proof, we will assume that $k \geq 6$.

For brevity, we will only consider the case where k is even. The case where k is odd is similar. Let $H_k(\cdot)$ be the physicist's Hermite polynomial. Recall that we can relate $h_k(\cdot)$ with $H_k(\cdot)$ with the following change of variable: $H_k(x) = \sqrt{2^k k!}h_k(\sqrt{2}x)$.

[Kra04, Theorem 1] implies the following:

$$\max_{x\in\mathbb{R}} \left((H_k(x))^2 e^{-x^2}\right) = O\left(\sqrt{k}k^{-1/6}\binom{k}{0.5k}k!\right) . \tag{6.6}$$

From Equation (6.6) we obtain:

$$\max_{x\in\mathbb{R}} 2^k k! h_k^2(\sqrt{2}x)e^{-x^2} = \max_{x\in\mathbb{R}} 2^k k! h_k^2(x)e^{-x^2/2} = O\left(\sqrt{k}k^{-1/6}\binom{k}{0.5k}k!\right) \quad .$$

This implies the following:

$$\max_{x\in\mathbb{R}} h_k^2(x)e^{-x^2/2} = O\left(k^{-1/6}\sqrt{k}\binom{k}{0.5k}2^{-k}\right) = O(k^{-1/6}),$$

where we use that $\binom{k}{0.5k}2^{-k}/\sqrt{k} = O(1)$. □

Proof of Lemma 6.4.4. Let h_i be the i-th normalized probabilist's Hermite polynomial. Since r is a polynomial of degree at most m and $\{h_i, i \in [m]\}$ form a basis for $\mathcal{P}(m)$, we can represent $r(x) = \sum_{i=1}^m a_i h_i(x)$ for some $a_i \in \mathbb{R}$. Using orthonormality of h_i under the Gaussian measure, we get that $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[r^2(X)] = \sum_{i=1}^m a_i^2$. Since r is a non-zero polynomial, we have that $\sum_{i=1}^m a_i^2 > 0$. We thus have that

$$\begin{aligned}
\sup_{r\in\mathcal{P}(m),r\not\equiv 0} \frac{r^2(y)}{\mathbb{E}_{X\sim\mathcal{N}(0,1)}[r^2(X)]} &= \sup_{a_1,\ldots,a_m\in\mathbb{R},\sum_{i=1}^m a_i^2>0} \frac{\sum_{i=1}^m\sum_{j=1}^m a_i a_j h_i(y)h_j(y)}{\sum_{i=1}^m a_i^2} \\
&= \sup_{a_1,\ldots,a_m\in\mathbb{R},\sum_{i=1}^m a_i^2>0} \frac{\sqrt{\sum_{i=1}^m\sum_{j=1}^m a_i^2 a_j^2}\sqrt{\sum_{i=1}^m\sum_{j=1}^m h_i^2(y)h_j^2(y)}}{\sum_{i=1}^m a_i^2} \\
&= \sum_{i=1}^m h_i^2(y).
\end{aligned}$$

Therefore, we need to show that, for all $y \in \mathbb{R}$, $\sum_{i=1}^m \alpha_y h_i^2(y) \le 1/2$ whenever $m \le C/\sqrt{\alpha}$ for a sufficiently small constant $C > 0$. We will now split the analysis in two cases:

Case 1: $|y| \le 1/\sqrt{\alpha}$**.** Using Lemma 6.4.8 and the assumption that $|y|^2\alpha \le 1$, we can bound the desired expression as follows:

$$\max_{|y|\le 1/\sqrt{\alpha}} \alpha_y h_i^2(y) = \max_{|y|\le 1/\sqrt{\alpha}} \sqrt{\alpha}\exp(y^2\alpha/2)\exp(-y^2/2)h_i^2(y)$$

$$\leq \sqrt{\alpha e} \sup_{y \in \mathbb{R}} \exp(-y^2/2) h_i^2(y)$$
$$= O(\sqrt{\alpha} i^{-1/6}).$$

Therefore, we get the following bound on $\sum_i h_i^2(y)$.

$$\sum_{i=1}^{m} \alpha_y h_i^2(y) = O\left(\sqrt{\alpha} \sum_{i=1}^{m} i^{-1/6}\right) = O(\sqrt{\alpha} m^{5/6}) \ .$$

The last expression is less than $1/2$ when $m = O(1/\alpha^{3/5})$.

Case 2: $|y| \geq 1/\sqrt{\alpha}$**.** We will use rather crude bounds here. We have the following explicit expression of $h_i(\cdot)$ (see, for example, [AAR99; Sze89]):

$$|h_i(x)| = \left|\frac{He_i(x)}{\sqrt{i!}}\right| = \left|\sqrt{i!} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{(-1)^j}{j!(i-2j)!} \frac{x^{i-2j}}{2^j}\right| = \left|\sqrt{i!} x^i \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{(-1)^j}{(2j)!(i-2j)!} x^{-2j} \frac{(2j)!}{j! 2^j}\right|$$
$$\leq \sqrt{i!} |x|^i \sum_{k=0}^{i} \frac{i!}{k!(i-k)!} |x|^{-k} \leq (i|x|)^i (1 + |x|^{-1})^i = i^i (1 + |x|)^i.$$

Therefore, we get the following relation for all $|y| > 1$, $\alpha < 0.5$, and $i \in \mathbb{N}$:

$$\alpha_y h_i^2(y) = \sqrt{\alpha} \exp(-y^2(1-\alpha)/2) h_i^2(y)$$
$$\leq \sqrt{\alpha} \exp(-y^2/4)(2i)^i |y|^i$$
$$= \sqrt{\alpha} \exp(-y^2/4 + i \log(2i|y|)).$$

The expression above is at most $C'\sqrt{\alpha}$ for a constant $C' > 0$ for all $|y| \geq c'\sqrt{i \log i}$ for a constant $c' > 0$. The latter condition holds whenever $1/\sqrt{\alpha} \geq c'\sqrt{i \log i}$. It suffices that $i = O(1/\alpha^{0.9})$. Overall, we get the following bound when $m = O(1/\alpha^{0.9})$:

$$\sup_{|y| > 1/\sqrt{\alpha}} \sum_{i=1}^{m} \alpha_y h_i^2(y) = O(\sqrt{\alpha} m).$$

The last expression is less than $1/2$ when $m \leq C/\sqrt{\alpha}$ for some constant $C > 0$. This completes the proof of Lemma 6.4.4. □

6.5 Hypothesis Testing Version of List-Decodable Linear Regression

Organization We introduce Problem 6.5.2, which is a hypothesis testing problem related to the search problem we discussed in Section 6.3. We first show the SQ-hardness of Problem 6.5.2 in Theorem 6.5.3. In Section 6.5.2, we give an efficient reduction from Problem 6.5.2 to list-decodable linear regression, showing that Problem 6.5.2 is indeed not harder than list-decodable linear regression. In Section 6.6, we also show the hardness of Problem 6.5.2 against low-degree polynomial tests.

We begin by formally defining a hypothesis problem.

Definition 6.5.1 (Hypothesis testing)**.** *Let a distribution D_0 and a set $\mathcal{S} = \{D_u\}_{u \in S}$ of distributions on $\mathbb{R}^d$. Let μ be a prior distribution on the indices S of that family. We are given access (via i.i.d. samples or oracle) to an* underlying *distribution where one of the two is true:*

- *H_0: The underlying distribution is D_0.*
- *H_1: First u is drawn from μ and then the underlying distribution is set to be D_u.*

We say that a (randomized) algorithm solves the hypothesis testing problem if it succeeds with non-trivial probability (i.e., greater than 0.9).

We now introduce the following hypothesis testing variant of the $(1-\alpha)$-contaminated linear regression problem:

Problem 6.5.2. *Let $\alpha \in (0, 1/2)$, $\rho \in (0, 1)$. Let S be the set of d-dimensional nearly orthogonal vectors from Lemma 6.1.15. We are given access (via i.i.d. samples or oracle) to an* underlying *distribution where one of the two is true:*

- *H_0: The underlying distribution is $R = \mathcal{N}(0, I_d) \times \mathcal{N}(0, 1/\alpha)$.*
- *H_1: First, a vector v is chosen uniformly at random from S. The underlying distribution is set to be E_v, i.e., the $(1-\alpha)$-additively corrupted linear model of Definition 6.1.2 with $\beta = \rho v$, $\sigma^2 = 1 - \rho^2$, and a fixed noise distribution N_v as specified in Lemma 6.3.5.*

Using the reduction outlined in Lemma 6.5.9, it follows that $O(d/\alpha^3)$ samples suffice to solve Problem 6.5.2 when $\sigma \leq O(\alpha/\sqrt{\log(1/\alpha)})$. On the other hand, the following result shows an SQ lower bound of $d^{\mathrm{poly}(1/\alpha)}$.

Theorem 6.5.3 (SQ Hardness of Problem 6.5.2)**.** *Let $0 < c < 1/2$, $m \in \mathbb{Z}_+$ with $m \leq c_1/\sqrt{\alpha}$ for some sufficiently small constant $c_1 > 0$ and $d = m^{\Omega(1/c)}$. Every SQ algorithm that solves Problem 6.5.2 either performs $2^{\Omega(d^{c/4})}$ queries or performs at least one query to $\mathrm{STAT}\left(\Omega(d)^{-(2m+1)(1/4-c/2)} e^{O(m)}/\sigma\right)$.*

We note that the lower bound on the (appropriate) statistical dimension implies SQ hardness of the (corresponding) hypothesis testing problem. As the Problem 6.5.2 differs slightly from the kind of hypothesis testing problems considered in [FGRVX17], we provide the proof of Theorem 6.5.3 in Section 6.5.1, where we introduce the relevant statistical dimension from [BBHLS21] (Definition 6.5.4 in this paper).

6.5.1 Hardness of Problem 6.5.2 in the SQ Model

We need the following variant of the statistical dimension from [BBHLS21], which is closely related to the hypothesis testing problems considered in this section. Since this is a slightly different definition from the statistical dimension (SD) used so far, we will assign the distinct notation (SDA) for it.

Notation For $f : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$ and a distribution D, we define the inner product $\langle f, g \rangle_D = \mathbb{E}_{X \sim D}[f(X)g(X)]$ and the norm $\|f\|_D = \sqrt{\langle f, f \rangle_D}$.

Definition 6.5.4 (Statistical Dimension)**.** *For the hypothesis testing problem of Definition 6.5.1, we define the* statistical dimension $\mathrm{SDA}(\mathcal{S}, \mu, n)$ *as follows:*

$$\mathrm{SDA}(\mathcal{S}, \mu, n) = \max \left\{ q \in \mathbb{N} : \mathop{\mathbb{E}}_{u,v\sim\mu}[|\langle \bar{D}_u, \bar{D}_v \rangle_{D_0} - 1| \mid \mathcal{E}] \leq \frac{1}{n} \right. \\ \left. \text{for all events } \mathcal{E} \text{ s.t. } \mathop{\mathbb{P}}_{u,v\sim\mu}[\mathcal{E}] \geq \frac{1}{q^2} \right\}.$$

We will omit writing μ when it is clear from the context.

Theorem 6.5.5 (Theorem A.5 of [BBHLS21])**.** *Let $\mathcal{S} = \{D_u\}_{u\in S}$ vs. D_0 be a hypothesis testing problem with prior μ on $\mathcal{S}$. If $\mathrm{SDA}(\mathcal{S}, \mu, 3/t) > q$, then every SQ algorithm that solves the hypothesis testing problem either makes at least q queries, or makes at least one query to $\mathrm{STAT}(\sqrt{t})$.*

In order to prove Problem 6.5.2, we will prove a lower bound on the SDA of Problem 6.5.2. As we will show later, Problem 6.5.2 is a special case of the following hypothesis testing problem:

Problem 6.5.6 (Non-Gaussian Component Hypothesis Testing)**.** *Let R be the joint distribution R over the pair $(X, y) \in \mathbb{R}^{d+1}$ where $X \sim \mathcal{N}(0, I_d)$ and $y \sim R(y)$ independently of X. Let E_v be the joint distribution over pairs $(X, y) \in \mathbb{R}^{d+1}$ where the marginal on y is again $R(y)$ but the conditional distribution $E_v(x|y)$ is of the form $P_{A_y,v}$ (with $P_{A_y,v}$ as in Definition 6.1.13). Define $\mathcal{S} = \{E_v\}_{v\in S}$ for S being the set of d-dimensional nearly orthogonal vectors from Lemma 6.1.15 and let the hypothesis testing problem be distinguishing between R vs. $\mathcal{S}$ with prior μ being the uniform distribution on S.*

The following lemma translates the (γ, β)-correlation of $\mathcal{S}$ to a lower bound for the statistical dimension of the hypothesis testing problem. The proof is very similar to that of Corollary 8.28 of [BBHLS21] but it is given below for completeness.

Lemma 6.5.7. *Let $0 < c < 1/2$ and $d, m \in \mathbb{Z}_+$ such that $d = m^{\Omega(1/c)}$. Consider the hypothesis testing problem of Problem 6.5.6 where for every $y \in \mathbb{R}$ the distribution A_y matches the first m moments with $\mathcal{N}(0,1)$ and $\mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))] < \infty$. Then, for any $q \geq 1$,*

$$\mathrm{SDA}\left(\mathcal{D}, \frac{\Omega(d)^{(m+1)(1/2-c)}}{\mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]\left(\frac{q^2}{2^{\Omega(d^{c/2})}}+1\right)}\right) \geq q\ .$$

Proof. The first part is to calculate the correlation of the set $\mathcal{S}$ exactly as we did in the proof of Theorem 6.3.1. By Lemma 6.1.15, Lemma 6.1.14 and Lemma 6.3.2 we know that the set $\mathcal{S}$ is (γ, β)-correlated with $\gamma = \Omega(d)^{-(m+1)(1/2-c)}\,\mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]$ and $\beta = \mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]$.

We next calculate the SDA according to Definition 6.5.4. We denote by $\bar{E}_v$ the ratios of the density of E_v to the density of R. Note that the quantity $\langle \bar{E}_u, \bar{E}_{,v}\rangle - 1$ used there is equal to $\langle \bar{E}_u - 1, \bar{E}_v - 1\rangle$. Let $\mathcal{E}$ be an event that has $\mathbb{P}_{u,v\sim\mu}[\mathcal{E}] \geq 1/q^2$. For d sufficiently large we have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{u,v\sim\mu}[|\langle \bar{E}_u, \bar{E}_v\rangle - 1|\mathcal{E}] &\leq \min\left(1, \frac{1}{|\mathcal{S}|\,\mathbb{P}[\mathcal{E}]}\right) \mathop{\mathbb{E}}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))] \\
&\quad + \max\left(0, 1 - \frac{1}{|\mathcal{S}|\,\mathbb{P}[\mathcal{E}]}\right) \frac{\mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]}{\Omega(d)^{(m+1)(1/2-c)}} \\
&\leq \mathop{\mathbb{E}}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]\left(\frac{q^2}{2^{\Omega(d^c)}} + \frac{1}{\Omega(d)^{(m+1)(1/2-c)}}\right) \\
&= \mathop{\mathbb{E}}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]\frac{q^2\Omega(d)^{(m+1)(1/2-c)} + 2^{\Omega(d^c)}}{2^{\Omega(d^c)}\Omega(d)^{(m+1)(1/2-c)}} \\
&= \mathop{\mathbb{E}}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]\left(\frac{\Omega(d)^{(m+1)(1/2-c)}}{q^2\Omega(d)^{(m+1)(1/2-c)}/2^{\Omega(d^c)} + 1}\right)^{-1} \\
&= \mathop{\mathbb{E}}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]\left(\frac{\Omega(d)^{(m+1)(1/2-c)}}{q^2/2^{\Omega(d^{c/2})} + 1}\right)^{-1},
\end{aligned}$$

where the first inequality uses that $\mathbb{P}[u = v|\mathcal{E}] = \mathbb{P}[u = v, \mathcal{E}]/\,\mathbb{P}[\mathcal{E}]$ and bounds the numerator in two different ways: $\mathbb{P}[u = v, \mathcal{E}]/\,\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[u = v]/\,\mathbb{P}[\mathcal{E}] = 1/(|\mathcal{D}|\,\mathbb{P}[\mathcal{E}])$ and $\mathbb{P}[u = v, \mathcal{E}]/\,\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{E}]/\,\mathbb{P}[\mathcal{E}] = 1$. □

We note that the lemma above and Theorem 6.5.5 show SQ hardness of Problem 6.5.6. In the remainder of this section, we will apply these results to Problem 6.5.2.

Corollary 6.5.8. *Let $0 < c < 1/2$, $m \in \mathbb{Z}_+$ with $m \le c_1/\sqrt{\alpha}$ for some sufficiently small constant $c_1 > 0$ and $d = m^{\Omega(1/c)}$. Consider the hypothesis testing problem of Problem 6.5.2. Then, for any $k < d^{c/4}$:*

$$\mathrm{SDA}\left(\mathcal{D}, \frac{\Omega(d)^{(2m+1)(1/2-c)}}{e^{O(m)}/(1-\rho^2)}\right) \ge 100^k \ .$$

Proof. We note that Problem 6.5.2 is a special case of Problem 6.5.6 (see Fact 6.3.3 and Lemma 6.3.5 which show that the conditional distributions are of the form $P_{A_y,v}$). In Lemma 6.5.7 we use $q = \sqrt{2^{\Omega(d^{c/2})}(n/n')}$ with $n' = n = \frac{\Omega(d)^{(2m+1)(1/2-c)}}{\mathbb{E}_{y\sim R(y)}[\chi^2(A_y,\mathcal{N}(0,1))]}$ to get that $\mathrm{SDA}(\mathcal{D}, n) > 100^k$ for $k < d^{c/4}$. The first part of Lemma 6.3.7 states that the distributions A_y's match the first $2m$ moments with $\mathcal{N}(0,1)$ for $m \le c_1/\sqrt{\alpha}$ and the second part implies that $\mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))] = O(e^m)/(1-\rho^2)$. This completes the proof. □

We conclude by noting the hardness of Problem 6.5.6 and thus Problem 6.5.2 in the SQ model. The proof of Theorem 6.5.3 follows from Corollary 6.5.8 and Theorem 6.5.5.

6.5.2 Reduction of Hypothesis Testing to List-Decodable Linear Regression

We now show that any list-decoding algorithm for robust linear regression can be efficiently used to solve Problem 6.5.2, that is, hypothesis testing efficiently reduces to list-decodable estimation. For a list $\mathcal{L}$ and $i \in [|\mathcal{L}|]$, we use $\mathcal{L}(i)$ to denote the i-th element of $\mathcal{L}$.

Lemma 6.5.9. *Let $d \in \mathbb{Z}_+$ with $d = 2^{\Omega(1/(1/2-c))}$. Consider the $(1-\alpha)$-corrupted linear regression model of Definition 6.1.2 with $\beta = \rho v$ for $v \in \mathcal{S}^{d-1}$, $\rho \in (0,1)$, $\sigma^2 = 1-\rho^2$. There*

exists an algorithm List_Regression_To_Testing *that, given a list-decoding algorithm $\mathcal{A}$ with the guarantee of returning a list $\mathcal{L}$ of candidate vectors such that for some $i \in \{1, \dots, |\mathcal{L}|\}$, $\|\mathcal{L}(i) - \beta\|_2 \le \rho/4$, solves the hypothesis testing Problem 6.5.2 with probability at least $1 - |\mathcal{L}|^2 e^{-\Omega(d^{2c})}$. The running time of this reduction is quadratic in $|\mathcal{L}|$.*

Proof. The reduction is described in Section 6.5.2. To see correctness, first assume that

Algorithm 2 Reduction from Hypothesis Testing to List-Decodable Linear Regression.

$\mathcal{A}(\rho, (X_1, y_1), \dots, (X_n, y_n))$: List-decoding algorithm returning a list L such that $\|\mathcal{L}(i) - \beta\|_2 \le \rho/4$ for some $i \in \{1, \dots, |\mathcal{L}|\}$.

function List_Regression_To_Testing($\rho, (X_1, y_1), \dots, (X_{2n}, y_{2n})$)
  Split dataset into two equally sized parts $\{(X_i, y_i)\}_{i=1}^n, \{(X'_i, y'_i)\}_{i=1}^n$.
  Let A be a random rotation matrix independent of data.
  $\mathcal{L}_1 \leftarrow \mathcal{A}(\rho, (X_1, y_1), \dots, (X_n, y_n))$.
  $\mathcal{L}_2 \leftarrow \mathcal{A}(\rho, (AX'_1, y'_1), \dots, (AX'_{2n}, y'_n))$.
  for $i \leftarrow 1$ to $|\mathcal{L}_1|$ **do**
    for $j \leftarrow 1$ to $|\mathcal{L}_2|$ **do**
      if $\|\mathcal{L}_1(i)\|_2, \|\mathcal{L}_2(j)\|_2 \in [3\rho/4, 5\rho/4]$ and $\|\mathcal{L}_1(i) - A^\top \mathcal{L}_2(j)\|_2 \le \rho/2$ **then**
        return H_1
      end if
    end for
  end for
  return H_0
end function

the alternative hypothesis holds. We note that the rotated points $(AX'_1, y'_1), \dots, (AX'_n, y'_n)$ come from the Gaussian linear regression model of Definition 6.1.2 having $\beta' = A\beta$ as the regressor. Thus $\mathcal{A}$ finds lists $\mathcal{L}_1, \mathcal{L}_2$ such that there exist $i^* \in \{1, \dots, |\mathcal{L}_1|\}$ with $\|\mathcal{L}_1(i^*) - \beta\|_2 \le \rho/4$ and $j^* \in \{1, \dots, |\mathcal{L}_2|\}$ with $\|A^\top \mathcal{L}_2(j^*) - \beta\|_2 \le \rho/4$, where we use that $A^\top A = I$. Moreover, since we are considering the regression model with $\|\beta\|_2 = \rho$, $\mathcal{L}_1(i^*)$ and $A^\top \mathcal{L}_2(j^*)$ must have norms belonging in $[3\rho/4, 5\rho/4]$. By the triangle inequality we get that $\|\mathcal{L}_1(i^*) - A^\top \mathcal{L}_2(j^*)\|_2 \le \rho/2$ and thus the algorithm correctly outputs H_1.

Now assume that the null hypothesis holds, where the marginal on points is $\mathcal{N}(0, I_d)$ and labels are independently distributed as $\mathcal{N}(0, 1/\alpha)$. Fix a pair $i \in [|\mathcal{L}_1|]$, $j \in [|\mathcal{L}_2|]$ for which $\|\mathcal{L}_1(i)\|_2, \|\mathcal{L}_2(j)\|_2 \in [3\rho/4, 5\rho/4]$. Note that, by rotation invariance of the

standard Gaussian distribution and the independence between covariates and response under the null distribution, the input $\{(AX'_i, y'_i)\}_{i=1}^n$ for the second execution of the list-decoding algorithm is independent of A. Thus the list $\mathcal{L}_2$ is independent of A (and also independent of $\mathcal{L}_1$). Thus, $A^\top \mathcal{L}_2(j)$ is a random vector selected uniformly from the sphere of radius $\|\mathcal{L}_2(j)\|_2$ and independently of $\mathcal{L}_1(i)$. Recall that two random vectors are almost orthogonal with high probability.

Lemma 6.5.10 (see, e.g., [CFJ13])**.** *Let θ be the angle between two random unit vectors uniformly distributed over $\mathcal{S}^{d-1}$. Then we have that $\mathbb{P}[|cos\theta| \geq \Omega(d^{c-1/2})] \leq e^{-\Omega(d^{2c})}$ for any $0 < c < 1/2$.*

Taking a union bound over the $|\mathcal{L}_1| \cdot |\mathcal{L}_2|$ possible pairs of candidate vectors, we have that with probability at least $1 - |\mathcal{L}_1| \cdot |\mathcal{L}_2| e^{-\Omega(d^{2c})}$, for all $i \in [|\mathcal{L}_1|], j \in [|\mathcal{L}_2|]$ we have that

$$\begin{aligned}\|\mathcal{L}_1(i) - A^\top \mathcal{L}_2(j)\|_2 &= \sqrt{\|\mathcal{L}_1(i)\|_2^2 + \|A^\top \mathcal{L}_2(j)\|_2^2 - 2(\mathcal{L}_1(i))^\top (A^\top \mathcal{L}_2(j))} \\ &\geq \sqrt{2(3\rho/4)^2(1 - \Omega(d^{c-1/2}))} > \rho \,,\end{aligned}$$

where in the last inequality we used that $d = 2^{\Omega(1/(1/2-c))}$. This concludes correctness for the case of the null hypothesis. □

We note that the Section 6.5.2 can be implemented in both of the models of computation that we consider: SQ model and low-degree polynomial test (Section 6.6). For the SQ model, we can simulate the queries on the rotated X by modifying the queries to explicitly perform the rotation on X by a matrix A. For the low-degree polynomial model, Remark 6.6.5 shows that this reduction can be implemented as a low-degree polynomial algorithm.

6.6 Hardness Against Low-Degree Polynomial Algorithms

In this section, we recall the recently established connection between the statistical query framework and low-degree polynomials, shown in [BBHLS21], and deduce hardness results in the latter model. Section 6.6.1 and Section 6.6.2 are dedicated to the hypothesis problem. In Section 6.6.3, we show that the reduction of Section 6.5.2 can be expressed as a low-degree polynomial test.

6.6.1 Preliminaries: Low-Degree Method

We begin by recording the necessary notation, definitions, and facts. This section mostly follows [BBHLS21].

Notation For a distribution D, we denote by $D^{\otimes n}$ the joint distribution of n independent samples from D. For $f : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$ and a distribution D, we define the inner product $\langle f, g \rangle_D = \mathbb{E}_{X \sim D}[f(X)g(X)]$ and the norm $\|f\|_D = \sqrt{\langle f, f \rangle_D}$. We will omit the subscripts when they are clear from the context.

Low-Degree Polynomials A function $f : \mathbb{R}^a \to \mathbb{R}^b$ is a polynomial of degree at most k if it can be written in the form

$$f(x) = (f_1(x), f_2(x), \ldots, f_b(x)) \ ,$$

where each $f_i : \mathbb{R}^a \to \mathbb{R}$ is a polynomial of degree at most k. We allow polynomials to have random coefficients as long as they are independent of the input x. When considering *list-decodable estimation* problems, an algorithm in this model of computation is a polynomial $f : \mathbb{R}^{d_1 \times n} \to \mathbb{R}^{d_2 \times \ell}$, where d_1 is the dimension of each sample, n is the

number of samples, d_2 is the dimension of the output hypotheses, and ℓ is the number of hypotheses returned. On the other hand, [BBHLS21] focuses on *binary hypothesis testing* problems defined in Definition 6.5.1.

A degree-k polynomial test for Definition 6.5.1 is a degree-k polynomial $f : \mathbb{R}^{d\times n} \to \mathbb{R}$ and a threshold $t \in \mathbb{R}$. The corresponding algorithm consists of evaluating f on the input $x_1, \ldots, x_n$ and returning H_0 if and only if $f(x_1, \ldots, x_n) > t$.

Definition 6.6.1 (n-sample ϵ-good distinguisher)**.** *We say that the polynomial $p : \mathbb{R}^{d\times n} \to \mathbb{R}$ is an n-sample ϵ-distinguisher for the hypothesis testing problem in Definition 6.5.1 if*

$$\left| \mathop{\mathbb{E}}_{X\sim D_0^{\otimes n}}[p(X)] - \mathop{\mathbb{E}}_{u\sim\mu} \mathop{\mathbb{E}}_{X\sim D_u^{\otimes n}}[p(X)] \right| \geq \epsilon \sqrt{\mathop{\mathbf{Var}}_{X\sim D_0^{\otimes n}}[p(X)]}\,.$$

We call ϵ the advantage *of the distinguisher.*

Let $\mathcal{C}$ be the linear space of polynomials with degree at most k. The best possible advantage is given by the *low-degree likelihood ratio*

$$\max_{\substack{p\in\mathcal{C} \\ \mathbb{E}_{X\sim D_0^{\otimes n}}[p^2(X)]\leq 1}} \left| \mathop{\mathbb{E}}_{u\sim\mu} \mathop{\mathbb{E}}_{X\sim D_u^{\otimes n}}[p(X)] - \mathop{\mathbb{E}}_{X\sim D_0^{\otimes n}}[p(X)] \right| = \left\| \mathop{\mathbb{E}}_{u\sim\mu} \left[(\bar{D}_u^{\otimes n})^{\leq k} \right] - 1 \right\|_{D_0^{\otimes n}},$$

where we denote $\bar{D}_u = D_u/D_0$ and the notation $f^{\leq k}$ denotes the orthogonal projection of f to $\mathcal{C}$.

Another notation we will use regarding a finer notion of degrees is the following: We say that the polynomial $f(x_1, \ldots, x_n) : \mathbb{R}^{d\times n} \to \mathbb{R}$ has *samplewise degree* (r, k) if it is a polynomial, where each monomial uses at most k different samples from $x_1, \ldots, x_n$ and uses degree at most d for each of them. In analogy to what was stated for the best degree-k distinguisher, the best distinguisher of samplewise degree (r, k)-achieves advantage $\left\| \mathbb{E}_{u\sim\mu}[(\bar{D}_u^{\otimes n})^{\leq r,k}] - 1 \right\|_{D_0^{\otimes n}}$ the notation $f^{\leq r,k}$ now means the orthogonal projection of f to the space of all samplewise degree-(r, k) polynomials with unit norm.

6.6.2 Hardness of Hypothesis Testing Against Low-Degree Polynomials

In this section, we show the following result:

Theorem 6.6.2. *Let $0 < c < 1/2$ and $m \in \mathbb{Z}_+$ with $m \le c_1/\sqrt{\alpha}$ for some sufficiently small constant $c_1 > 0$. Consider the hypothesis testing problem of Problem 6.5.2. For $d \in \mathbb{Z}_+$ with $d = m^{\Omega(1/c)}$, any $n \le \Omega(d)^{(2m+1)(1/2-c)} e^{-O(m)}(1-\rho^2)$ and any even integer $k < d^{c/4}$, we have that*

$$\left\| \mathop{\mathbb{E}}_{u\sim\mu} \left[(\bar{E}_u^{\otimes n})^{\le \infty, \Omega(k)} \right] - 1 \right\|^2_{R^{\otimes n}} \le 1 \, .$$

We prove Theorem 6.6.2 by using the lower bound on SDA in Corollary 6.5.8 and the relation between SDA and low-degree polynomials established in [BBHLS21]. In [BBHLS21], the following relation between SDA and low-degree likelihood ratio is established.

Theorem 6.6.3 (Theorem 4.1 of [BBHLS21])**.** *Let $\mathcal{D}$ be a hypothesis testing problem on $\mathbb{R}^d$ with respect to null hypothesis D_0. Let $n, k \in \mathbb{N}$ with k even. Suppose that for all $0 \le n' \le n$, $\mathrm{SDA}(\mathcal{S}, n') \ge 100^k (n/n')^k$. Then, for all r, $\left\| \mathbb{E}_{u\sim\mu} \left[(\bar{D}_u^{\otimes n})^{\le r, \Omega(k)} \right] - 1 \right\|^2_{D_0^{\otimes n}} \le 1$.*

We first apply Theorem 6.6.3 to the more general Problem 6.5.6. In Lemma 6.5.7 we set $n = \frac{\Omega(d)^{(m+1)(1/2-c)}}{\mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]}$ and $q = \sqrt{2^{\Omega(d^{c/2})}(n/n')}$. Then, $\mathrm{SDA}(\mathcal{S}, n') \ge \sqrt{2^{\Omega(d^{c/2})}(n/n')} \ge (100n/n')^k$ for $k < d^{c/4}$. Thus, we have shown the following.

Corollary 6.6.4. *Let $0 < c < 1/2$ and the hypothesis testing problem of Problem 6.5.6 where for every $y \in R$ the distribution A_y matches the first m moments with $\mathcal{N}(0,1)$. For any $d \in \mathbb{Z}_+$ with $d = m^{\Omega(1/c)}$, any $n \le \Omega(d)^{(m+1)(1/2-c)} / \mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))]$ and any even integer*

$k < d^{c/4}$, we have that

$$\left\| \mathop{\mathbb{E}}_{u\sim\mu} \left[(\bar{D}_u^{\otimes n})^{\leq \infty, \Omega(k)} \right] - 1 \right\|_{R^{\otimes n}}^2 \leq 1 \ .$$

Proof of Theorem 6.6.2. We now apply the Corollary 6.6.4 to Problem 6.5.2, which is a special case of Problem 6.5.6. The first part of Lemma 6.3.7 states that the distributions A_y's match the first $2m$ moments with $\mathcal{N}(0,1)$ for $m \leq c_1/\sqrt{\alpha}$ and the second part implies that $\mathbb{E}_{y\sim R(y)}[\chi^2(A_y, \mathcal{N}(0,1))] = O(e^m)/(1-\rho^2)$. An application of Corollary 6.6.4 completes the proof. □

6.6.3 Low-Degree Polynomial Reduction to List-Decodable Regression

Remark 6.6.5. *We note that the reduction of Lemma 6.5.9 is an algorithm that can be expressed in the low-degree polynomials model. The modification of the algorithm is the following: First note that the ℓ_2-norm of a vector is indeed a polynomial of degree two in each coordinate. Second, one can check whether there exists a pair $i \in [|\mathcal{L}_1|], j \in [|\mathcal{L}_2|]$ with $\|\mathcal{L}_1(i)\|_2, \|\mathcal{L}_2(j)\|_2 \in [3\rho/4, 5\rho/4]$ for which $\|\mathcal{L}_1(i) - A^\top \mathcal{L}_2(j)\|_2 \leq \rho/2$ using the condition*

$$\sum_{i\in 1}^{|\mathcal{L}_1|} \sum_{j\in 1}^{|\mathcal{L}_2|} \mathbb{I}(\|\mathcal{L}_1(i)\|_2^2 \geq (3\rho/4)^2) \cdot \mathbb{I}(\|A^\top \mathcal{L}_2(j)\|_2^2 \leq (5\rho/4)^2) \cdot \mathbb{I}(\|\mathcal{L}_1(i) - A^\top \mathcal{L}_2(j)\|_2^2 \leq \rho^2/4) = 0 \ ,$$

and use a polynomial approximation for the step function in order to express each term as a polynomial. The degree needed for a uniform ϵ-approximation has been well-studied [*GR08; Gan02; EY07*].

Lemma 6.6.6 ([EY07])**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be the step function defined as $f(x) = 1$ for all $x \geq 0$ and $f(x) = 0$ otherwise. The minimum $k \in \mathbb{Z}_+$ for which there exists a degree-k polynomial $p : \mathbb{R} \to \mathbb{R}$ such that $\max_{x\in[-1,1]} |f(x) - p(x)| \leq \epsilon$ is $k = \Theta(1/\epsilon^2)$.*

For our purpose, it suffices to approximate the step function up to error $\epsilon = \Theta(1/(|\mathcal{L}_1| \cdot |\mathcal{L}_2|))$*, thus the resulting polynomial test has degree* $\Theta(|\mathcal{L}_1|^2 \cdot |\mathcal{L}_2|^2)$*.*

7 ESTIMATING LOCATION PARAMETERS IN SAMPLE-HETEROGENEOUS DISTRIBUTIONS

रात को जीत तो पाता नहीं लेकिन ये चराग़
कम से कम रात का नुक़सान बहुत करता है

इरफ़ान सिद्दीक़ी

Estimating the mean of a probability distribution using i.i.d. samples is a classical problem in statistics, wherein finite-sample optimal estimators are sought under various distributional assumptions. In this paper, we consider the problem of mean estimation when independent samples are drawn from d-dimensional non-identical distributions possessing a common mean. When the distributions are radially symmetric and unimodal, we propose a novel estimator, which is a hybrid of the modal interval, shorth, and median estimators, and whose performance adapts to the level of heterogeneity in the data. We show that our estimator is near-optimal when data are i.i.d. and when the fraction of "low-noise" distributions is as small as $\Omega\left(\frac{d \log n}{n}\right)$, where n is the number of samples. We also derive minimax lower bounds on the expected error of any estimator that is agnostic to the scales of individual data points. Finally, we extend our theory to linear regression. In both the mean estimation and regression settings, we present computationally feasible versions of our estimators that run in time polynomial in the number of data points.

7.1 Introduction

Heterogeneity is prevalent in many modern data sets, leading to new challenges in estimation and prediction. The i.i.d. assumption imposed in much of classical statistics is unlikely to hold in practice, creating a need to develop new theory under relaxed

assumptions allowing for heterogeneous data [Liu88; DKBR07; SC09; ZXLZ15; FNS16]. In this paper, we consider the problem of estimating a common mean when independent data are drawn from non-identical distributions.

A version of this problem for Gaussian distributions was recently studied in Chierichetti et al. [CDKL14], who motivated their work using the following crowdsourcing application: Suppose the quality of an item is obtained by soliciting ratings from several agents, who are assumed to provide unbiased ratings. However, the rating distributions may vary across agents depending, e.g., on their expertise. In the Gaussian setting, this translates into data drawn from independent distributions with a common mean but possibly different variances. Chierichetti et al. [CDKL14] proposed a mean estimator based on calculating the "shortest gap" between samples, and derived upper bounds on the estimation error of their algorithm. Naturally, one might ask whether the estimators proposed by Chierichetti et al. [CDKL14] also perform provably well for non-Gaussian settings; furthermore, although Chierichetti et al. [CDKL14] derived some lower bounds for the behavior of the best possible estimator in the unknown variance setting, the optimality of their proposed estimator was only partially addressed.

The work of this paper revisits the problem of common mean estimation and generalizes the case of Gaussian mixtures considered in Chierichetti et al. [CDKL14] to settings where the component distributions are only assumed to be symmetric and unimodal about a common mean. Although the estimators studied in our paper resemble the estimators proposed by Chierichetti et al. [CDKL14], our method of analysis is substantially different and allows us to obtain bounds without assuming Gaussianity, sub-Gaussianity, or even finite variances of individual distributions. In the multivariate mean estimation setting, this leads to sharper estimation error rates than those obtained in Chierichetti et al. [CDKL14] for isotropic Gaussian data. The upper bounds we derive

are stated in terms of percentiles of the overall mixture distribution and may be finite even in the case of heavy-tailed distributions.

The aforementioned model of non-i.i.d. data has even older roots in the statistics literature, under the name of *sample heterogeneity*. Initial research in sample heterogeneity focused on understanding the asymptotic distribution of order statistics and linear functions thereof [Hoe56; Wei69; Sen68; Sen70; Sti76; SW09]. More recent work has established necessary and sufficient conditions for consistency of the sample median [HM97; MW98; HM01]. In particular, Hallin and Mizera [HM01] established the optimality of the median over a certain class of M-estimators (having a bounded, non-decreasing, skew-symmetric score function). However, as explained in more detail later (cf. Section 7.3.2), certain cases exist where the median itself is not optimal in comparison to more complicated estimators. For example, redescending M-estimators do not lie in the class studied by Hallin and Mizera [HM01]. We show that, under certain conditions, the modal interval estimator (Estimator 1)—which may be viewed as an extreme case of a redescending M-estimator—has smaller error than the median (cf. Table 7.1). Sample heterogeneity was also studied in the linear regression setting, where previous work focused on the least absolute deviation estimator [EH99; Kni99]. Inspired by the modal estimator, we propose and analyze a related estimator for linear regression (cf. Section 7.8). Note that we are chiefly interested in estimators which have minimal assumptions and allow the fraction of low-variance points to be as small as $\frac{\log n}{n}$, whereas the estimators studied in previous work required the fraction to be $\Omega\left(\frac{\sqrt{n}}{n}\right)$ [HM01; HM01; EH99; Kni99; MW98].

We also briefly mention classical work on the modal interval estimator [Che64] and shorth estimator [ABHHRT72], which are used as building blocks for our hybrid estimator. Notably, previous analysis has focused on asymptotic results for i.i.d. data, where both the modal interval and shorth estimators were proven to have an $n^{-\frac{1}{3}}$ convergence

rate [KP90], in contrast to the faster $n^{-\frac{1}{2}}$ convergence rate of the sample mean. The results of this paper show the benefit of these estimators when a substantial fraction of the component distributions have large (or even infinite) variances, underscoring the general fact that robustness may need to be traded off for efficiency in clean-data settings.

The main contributions of our paper may be summarized as follows:

- Provide a rigorous analysis of the modal interval (Theorems 7.3.1, 7.4.1, and 7.4.3), shorth (Theorems 7.3.3 and 7.4.5), and hybrid (Theorems 7.3.5 and 7.4.6) estimators for multivariate, radially symmetric distributions. We also show how to relax the symmetry conditions further (Theorem 7.7.1). These estimation error guarantees hold with high probability.

- Derive upper bounds on the expected error of the estimators (Theorem 7.5.3). Along the way, we demonstrate the need for additional conditions on the tails of the mixture components in order to derive expected error bounds of the same order as the high-probability results.

- Derive minimax lower bounds on the error rate of any estimator (Theorem 7.5.5), and prove that the hybrid estimator is nearly optimal in various regimes of interest (Theorem 7.5.7).

- Extend the methodology for multivariate mean estimation to linear regression (Theorem 7.8.3).

- Provide computationally efficient versions of the multivariate mean estimator (Theorem 7.6.2) and linear regression estimator (Theorem 7.8.4) in high dimensions.

We also note that while our work vastly generalizes the results of Chierichetti et al. [CDKL14] for mean estimation in Gaussian mixtures, our derivations bypass some

critical technical gaps in their proofs using a very different approach via empirical process theory. Finally, we comment that preliminary work on this topic appeared in our earlier conference paper [PJL19b], but was limited to the univariate case (Theorems 7.3.1, 7.3.3, 7.3.5, and 7.4.3) and did not discuss optimality, regression, or any computational aspects[11]. Furthermore, all examples and counterexamples illustrating various phenomena, including the detailed theoretical derivations (Propositions 7.3.9–7.5.2), are new to this paper.

We end with a few remarks regarding parameter estimation in mixture models. The setting studied in our paper is markedly different from the canonical setting [Lin95; Das99; AK01; KSV05; AM05], since the number of components in the mixture distribution is allowed to be as large as the number of observations. Furthermore, the parameters of the component mixtures are "entangled" in the sense that they share a common mean, which we wish to estimate. Notably, this allows us to obtain meaningful error guarantees without imposing strong distributional assumptions such as Gaussianity or log-concavity, which are prevalent in the literature on parameter estimation for mixture models.

The roadmap of the paper is as follows: In Section 7.2, we define notation and the basic estimators we will consider in the univariate case, which are subsequently analyzed in Section 7.3. In Section 7.4, we present results for the multivariate analog of these estimators. In Section 7.5, we derive expected error bounds on the performance of our estimators, and also present minimax lower bounds on the estimation error of any estimator, thus providing settings in which our proposed estimators are provably optimal. In Section 7.6, we present computationally feasible variants of our estimators in higher dimensions, and prove that the error rates of these estimators are of the same order as those derived earlier. In Section 7.7, we discuss various relaxations of the

[11]We also mention follow-up papers by Liang and Yuan [LY20] and Devroye et al. [DLLZ23], which appeared after the initial posting of our conference paper.

symmetry assumptions on the mixture components. In Section 7.8, we describe our results for linear regression. Simulation results reporting the relative performance of different estimators are contained in Section 7.9. All proofs are contained in the supplementary appendix.

Notation: We regularly use the standard big-O notation: For two real-valued non-negative functions $f(n)$ and $g(n)$, we write $f = O(g)$, when there exists constants n_0 and $C > 0$ such that for all $n \geq n_0$, $f(n) \leq Cg(n)$. We say $f = \Omega(g)$ if $g = O(f)$, and say $f = \Theta(g)$ when $f = O(g)$ and $g = O(f)$. We write $f = \omega(g)$ if for every real constant $c > 0$, there exists $n_0 \geq 1$ such that $f(n) > c \cdot g(n)$ for every integer $n \geq n_0$. We write $f = o(g)$, when $g = \omega(f)$. We use $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$, and $\tilde{\omega}(\cdot)$ to hide polylogarithmic factors. We write w.h.p., or "with high probability," to mean with probability tending to 1 as the sample size increases. We use C, c, C', and c' to represent absolute positive constants which may vary from place to place, and their exact values can be found in the proofs. Similarly, we use C_t to represent positive numbers that depend only on t. For a real-valued random variable X, we use $\mathbf{Var}\, X$ to denote its variance.

We will use $\|\cdot\|_2$ to denote the Euclidean norm. We use $B(x, r)$ to denote the Euclidean ball of radius r centered around x, and we also write B_r in place of $B(0, r)$. We denote the $d \times d$ identity matrix by I_d. We use $P(X, \epsilon)$ to denote the ϵ-packing number of a set X with respect to Euclidean distance, and we use $N(X, \epsilon)$ to denote the ϵ-covering number. We write $\mathrm{Diam}(X)$ to denote the diameter of the set with respect to Euclidean distance, i.e., $\mathrm{Diam}(X) := \sup_{x,y \in X} \|x - y\|_2$.

7.2 Problem Setup

We begin by introducing the entangled mean estimation problem. Suppose we have n independent samples $X_i \sim P_i$, where each P_i is a distribution in $\mathbb{R}^d$ with a density. Furthermore, we assume that each density p_i is radially symmetric and unimodal with

a common mean (and median) μ^*. Our goal is to estimate the location parameter μ^* from the n samples, where the P_i's are unknown a priori and may even come from different classes of (non)parametric distributions. Since the estimators we consider are translation-invariant, we can assume without loss of generality that $\mu^* = 0$, so the error of an estimator $\widehat{\mu}$ is measured by $\|\widehat{\mu}\|_2$.

A natural estimator to use is the empirical mean, which is certainly an unbiased estimator of μ^*. However, it is a well-known fact that the mean is not "robust," in the sense that one outlying observation can have a massive impact on the estimation error of the mean. In our setting, one P_i with a very large variance can dramatically inflate the error of the mean, even if the remaining $n - 1$ distributions are well-behaved. Due to the symmetry assumption on the P_i's, we could consider a (multivariate) median as a more robust alternative. Our theory in Section 7.3 below shows that using a median estimator can somewhat improve the estimation error so that it depends only on the spread of the $\sqrt{n} \log n$ distributions with the smallest quantiles; however, other more cleverly constructed estimators can reduce this dependence to $Od \log n)$ distributions, meaning that the remaining mixture components may have arbitrarily large (or even infinite) variances, yet have a bounded effect on the behavior of the estimator.

Another potential estimator when the mixing components come from a sufficiently nice parametric family (e.g., Gaussians) is the maximum likelihood estimator. However, since we do not assume knowledge of which observations are drawn from which mixture components, the MLE calculation becomes considerably more complicated. Nonetheless, it is sometimes informative to compare the error rate of the MLE—assuming side information of which observations correspond to which mixture components—to the error rates obtained using various agnostic estimators. In particular, if the former error rate diverges with n, we know that a diverging error rate for a proposed estimator is reasonable.

We will focus on the simple setting where the overall mixture distribution is radially symmetric, e.g., we have multivariate Gaussian observations $X_i \sim \mathcal{N}(0_d, \sigma_i^2 I_d)$. Throughout this paper, we focus on the setting where $d = O \log n)$; as shown in Chierichetti et al. [CDKL14], when $d = \Omega(\log n)$, the problem reduces to the case of known variances, since these can be estimated accurately. We shall discuss how to replace the spherical symmetry assumption by log-concavity in Section 7.7. As the covariance matrix of a radially symmetric distribution is of the form $\sigma^2 I_d$, we denote the covariance matrix of X_i by $\sigma_i^2 I_d$.

We now define the central objects in our analysis:

Definition 7.2.1 (Order statistics)**.** *Let the covariance of X_i be $\sigma_i^2 I_d$. Define $\sigma_{(i)}$ to be the corresponding order statistic. Let s_i denote the interquartile range of X_i, so that $\mathbb{P}(\|X_i - \mu^*\|_2 \leq s_i) = \frac{1}{2}$. Define $s_{(i)}$ to be the corresponding order statistic.*

Definition 7.2.2 (Indicator functions on balls)**.** *For $x \in \mathbb{R}^d$ and $r \in \mathbb{R}$, let $f_{x,r}(z) := \mathbb{1}_{\|x-z\|_2 \leq r}$ denote the indicator function of the ℓ_2-ball $B(x, r)$. For $s \in \mathbb{R}$, we will also use $f_{s,r}(z)$ to denote the indicator of the ball of radius r centered at the vector with first coordinate s and all other coordinates equal to 0.*

Note that when $d = 1$, the function $f_{x,r}$ is simply the indicator function of the interval $[x - r, x + r]$.

Definition 7.2.3 (Function class)**.** *Let*

$$\mathcal{H}_r := \{f_{x,r'} : x \in \mathbb{R}^d, r' \in \mathbb{R}, 0 \leq r' \leq r\}.$$

Note that $\mathcal{H}_r$ has VC dimension $d + 1$ [WD81].

As in prior analysis of sample heterogeneous models [SW09; Sen68], most of our arguments will be in terms of the mixture distribution $\overline{P} := \frac{1}{n} \sum_{i=1}^n P_i$, which is again

unimodal and symmetric. We will write $\overline{P}_n$ to denote the empirical distribution of $X_1, \ldots, X_n$.

Definition 7.2.4 (Risk). *For a function f, we use $R_n(f) := \frac{1}{n}\sum_{i=1}^n f(X_i)$ to denote the expectation of f with respect to the empirical distribution of $X_1, \ldots, X_n$. Let*

$$R(f) := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\, f(X_i).$$

Thus, $R(f)$ is the expectation of f with respect to $\overline{P}$. Define

$$R_r^* := \sup_{f \in \mathcal{H}_r} R(f) = R(f_{0,r}),$$

where the second equality follows by symmetry and unimodality.

Note that $R(f_{0,r})$ also equals the probability of the ball $B(0,r)$ under $\overline{P}$. The spherical symmetry assumption readily gives $R(f_{x,r}) = R(f_{s,r})$ for all x such that $\|x\|_2 = s$.

We first state several useful properties of radially symmetric distributions. The proof of the following result is contained in Appendix E.1.1.

Lemma 7.2.5. *Recall Definitions 7.2.1, 7.2.2, and 7.2.4 of $\sigma_{(i)}$, $s_{(i)}$, $f_{x,r}$, and R_r^*. Suppose the density of $\overline{P}$ is radially symmetric and unimodal. We have the following properties:*

(*i*) *For any $r > 0$ and $x, x' \in \mathbb{R}^d$, if $\|x\|_2 < \|x'\|_2$, then $R(f_{x,r}) \geq R(f_{x',r})$.*

(*ii*) *For any $x \in \mathbb{R}^d$, if $r < r'$, then $R(f_{x,r}) \leq R(f_{x,r'})$.*

(*iii*) *If $0 < r_1 < r_2$, then $\frac{R_{r_1}^*}{r_1^d} > \frac{R_{r_2}^*}{r_2^d}$.*

(*iv*) *If $0 < r_1 < r_2$, then*

$$R(f_{r_2,r_1}) < \frac{1}{P(B_{r_2-r_1}, r_1)} R_{r_2}^* \leq \left(\frac{2r_1}{r_2 - r_1}\right)^d R_{r_2}^*,$$

where $P(B_{r_2-r_1}, r_1)$ denotes the packing number of $B_{r_2-r_1}$ with respect to B_{r_1}. In particular, if $r_1 \le \frac{r_2}{2}$, then $R(f_{r_2,r_1}) \le \left(\frac{4r_1}{r_2}\right)^d R^*_{r_2}$.

(v) *If $1 \le k \le n/2$, then $\frac{k}{n} < R^*_{s_{(2k)}}$ and $\frac{k}{n} < R^*_{2\sqrt{d}\sigma_{(2k)}}$.*

7.2.1 Estimators

We now proceed to define the estimators that will be studied in our paper.

Estimator 1 (r-modal interval)**.** *The r-modal interval estimator, introduced for the (univariate) i.i.d. setting by Chernoff [Che64], outputs the center of the most populated ball of radius r, with ties broken arbitrarily:*

$$\widehat{\mu}_{M,r} \in \arg\max_x R_n(f_{x,r}). \tag{7.1}$$

Estimator 2 (k-shortest gap / shorth estimator)**.** *For $k \ge 2$, the k-shortest gap (k-shorth) estimator, $\widehat{\mu}_{S,k}$, outputs the center of the smallest ball containing at least k points. More precisely, we define*

$$\widehat{r}_k := \inf\left\{r : \sup_x R_n(f_{x,r}) \ge \frac{k}{n}\right\}, \qquad \widehat{\mu}_{S,k} := \widehat{\mu}_{M,\widehat{r}_k}. \tag{7.2}$$

The traditional (univariate) shorth estimator [ABHHRT72; KP90] corresponds to $k = \frac{n}{2}$, whereas choosing $k = 2$ outputs the midpoint of the shortest interval between any two points. As we will see, the choice of $k = C \log n$ will be convenient for our setting, and is more suitable than $k = \frac{n}{2}$ if data are not i.i.d.

Note that a type of "shortest interval" estimator has also been employed in the work on mean estimation for contaminated i.i.d. data [LRV16], but was used as an outlier screening step in that context, rather than a mean estimator. Incidentally, our hybrid

estimator to be introduced later will employ a different screening approach based on the median, and then use the shorth estimator to return a more accurate mean estimate.

Definition 7.2.6. *Recall Definitions 7.2.2 and 7.2.4 for the quantity $R(f_{x,r})$. Define*

$$r_k := \inf\left\{r : \sup_x R(f_{x,r}) \geq \frac{k}{n}\right\} = \inf\left\{r : R(f_{0,r}) \geq \frac{k}{n}\right\},$$

where the second equality follows from unimodality and radial symmetry.

The quantity r_k measures the spread of $\overline{P}$, and $r_{n/2}$ is the interquartile range of $\overline{P}$. Furthermore, since $\overline{P}$ has a density, we have $R^*_{r_k} = \frac{k}{n}$. Note that r_k is problem-dependent, since its magnitude depends on the relative dispersion of the mixing components; in particular, we will be interested in $r_{\Theta(d\log n)}$. As the fraction of "nice" points increases, r_k becomes smaller. However, r_k does not depend too strongly on the high-variance distributions (cf. Lemma (i) and Proposition 7.3.9).

The univariate k-median outputs an element from the centermost k points of the data. According to our definition, the k-median outputs a set rather than a point estimator, which will be used as a preprocessing step before applying the modal interval or shorth estimators to obtain a hybrid estimator with superior rates.

Estimator 3 (k-median)**.** *In the univariate setting, the k-median estimator outputs an arbitrary element $\widehat{\mu}_{med,k}$ from the subset S_k, defined as $X_i \in S_k$ if and only if $\widehat{\theta}_{med,-k} \leq X_i \leq \widehat{\theta}_{med,k}$, where*

$$\widehat{\theta}_{med,k} := \inf\left\{\theta : \psi_n(\theta) \geq \frac{k}{n}\right\},$$
$$\widehat{\theta}_{med,-k} := \sup\left\{\theta : \psi_n(\theta) \leq \frac{-k}{n}\right\},$$

and $\psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n sign(\theta - X_i)$. The sample median corresponds to taking $k = 0$.

Various multivariate extensions of the median exist, with different robustness properties and computational complexity; for our purposes, it will suffice to consider the simplest version of the

multivariate median, which simply operates componentwise on the data points.

Definition 7.2.7 (Multivariate median)**.** *Define the set $S_{k,i}$ as follows: For each dimension i, consider the k median points in that dimension; i.e.,*

$$S_{k,i} := \{X_j(i) : X_j(i) \text{ belongs to the } k\text{-median of } (X_j(i))_{j=1}^n\},$$

where $X_j(i)$ denotes the i^{th} coordinate of the vector X_j. Define S_k^∞ to be the cuboid based on $S_{k,i}$, for each dimension i:

$$S_k^\infty := \prod_{i=1}^d [\min(S_{k,i}), \max(S_{k,i})].$$

Estimator 4 (Hybrid estimator)**.** *The hybrid algorithm consists of the following steps, summarized in Algorithm 3:*

(*i*) *Compute the cuboid $S_{k_1}^\infty$ with $k_1 = \sqrt{n}\log n$.*

(*ii*) *Compute the k_2-shorth estimator $\widehat{\mu}_{S,k_2}$ with $k_2 = Cd\log n$.*

(*iii*) *If $\widehat{\mu}_{S,k_2} \notin S_{k_1}^\infty$, return the projection of $\widehat{\mu}_{S,k_2}$ on $S_{k_1}^\infty$. Otherwise, return $\widehat{\mu}_{S,k_2}$.*

Algorithm 3 Hybrid mean estimator (d-dimensional)

function HYBRIDMULTIDIMENSIONAL($X_{1:n}, k_1, k_2, d$)
  $S_{k_1}^\infty \leftarrow$ kCuboid($X_{1:n}, k_1$).
  $\widehat{\mu}_{S,k_2} \leftarrow$ Shorth($X_{1:n}, k_2$).
  if $\widehat{\mu}_{S,k_2} \in S_{k_1}^\infty$ **then**
    $\widehat{\mu}_{k_1,k_2} \leftarrow \widehat{\mu}_{S,k_2}$
  else
    $\widehat{\mu}_{k_1,k_2} \leftarrow \arg\min_{x \in S_{k_1}^\infty} \|x - \widehat{\mu}_{S,k_2}\|_2$
  end if
  return $\widehat{\mu}_{k_1,k_2}$
end function

Note that the projection in step (iii) is easy to accomplish, since ℓ_2-projection onto the cuboid may be done componentwise, hence computed in $O(d)$ time. Our theoretical results show that replacing the shorth estimator by the modal interval estimator produces similar statistical error rates.

7.2.2 Concentration Inequality

The following concentration inequality will be a key technical ingredient for deriving results concerning our estimators. The proof is contained in Appendix E.1.2.

Lemma 7.2.8. *Recall the Definitions 7.2.3 and 7.2.4 of the terms $R_n(f), R(f), R_r^*$, and $\mathcal{H}_r$. For any fixed $t \in (0, 1]$ and $n > 1$, we have*

$$\mathbb{P}\left\{\sup_{f \in \mathcal{H}_r} |R_n(f) - R(f)| \geq t R_r^*\right\} \leq 2 \exp\left(-cn R_r^* t^2\right),$$

provided r is large enough so that $nR_r^* \geq C_t \frac{d+1}{2} \log n$, where $C_t = \left(\frac{144}{t}\right)^2$ and $c = \frac{1}{200}$.

This theorem is useful because the bounds rely on R_r^*; i.e., they are adaptive to the problem, compared to the traditional $O\left(\frac{1}{\sqrt{n}}\right)$ distribution-independent bound. We also note that Lemma 7.2.8 requires the mass R_r^* lying around the true mode to be sufficiently large, and while the theorem requires R_r^* to increase with d, we will work in settings where $d = O \log n)$.

7.3 Univariate Mean Estimation

We now state several theoretical guarantees for the aforementioned estimators in the univariate setting. Some of these results appeared in our preliminary work [PJL19b], but we provide complete proofs of all statements in Appendix E.2.

In the univariate setting, we assume that we have n independent samples $X_i \sim P_i$, where each P_i is a univariate distribution with a density p_i which is symmetric and decreasing around μ^*. Let q_i and σ_i denote the interquartile range and standard deviation of P_i, respectively, and recall that the interquartile range satisfies $\mathbb{P}(|X_i - \mu^*| \leq q_i) = \frac{1}{2}$. We use $q_{(i)}$ and $\sigma_{(i)}$ to denote the i^{th} smallest interquartile range and standard deviation, respectively (cf. Definition 7.2.1). By Lemma E.2.1(v) below, we have $r_k \leq q_{(2k)}$ and $r_k \leq 2\sigma_{(2k)}$, although these bounds may be loose (for instance, r_k could be finite even if $\sigma_{(1)}$ is infinite). However, we are guaranteed that r_k will be small if $2k$ points come from "nice" (low-variance) distributions.

Theorem 7.3.1 (Theorem 2 of Pensia et al. [PJL19b])**.** *Recall Definitions 7.2.2 and 7.2.4 of the terms R_r^*, $R(\cdot)$, and $f_{r',r}$. Let r be a fixed number such that $R_r^* = \Omega\left(\frac{\log n}{n}\right)$. Then with probability at least $1 - 2\exp(-c'nR_r^*)$, the modal interval estimator (Estimator 1) satisfies*

$$|\widehat{\mu}_{M,r}| \leq r', \tag{7.3}$$

for any r' that satisfying $R(f_{r',r}) < \frac{R_r^*}{2}$. In particular, we can always choose $r' = \frac{2r}{R_r^*}$ to obtain the bound

$$|\widehat{\mu}_{M,r}| \leq \frac{2r}{R_r^*}. \tag{7.4}$$

The proof of Theorem 7.3.1 is contained in Appendix E.2.1, and proceeds by using Lemma 7.2.8 to bound the ratio between $R(f_{\widehat{\mu}_{M,r},r})$ and R_r^*, and then using Lemma E.2.1 to turn this into a deviation bound on $|\widehat{\mu}_{M,r}|$. Although the bound (7.4) in Theorem 7.3.1 is simple to state, it may be looser than the bound (7.3).

Remark 7.3.2. *Importantly, by Lemma E.2.1(v), we know that the choice $r = \sigma_{(C\log n)}$ always*

guarantees the condition $R_r^* = \Omega\left(\frac{\log n}{n}\right)$. *Hence, inequality* (7.4) *implies that*

$$|\widehat{\mu}_{M,r}| \leq \frac{2\sigma_{(C \log n)}}{R_r^*} \leq \frac{2n\sigma_{(C \log n)}}{\log n}, \tag{7.5}$$

with a similar inequality involving $q_{(C' \log n)}$. *Note that this bound holds regardless of the magnitude of the standard deviations of the latter* $n - C \log n$ *mixture components.*

At the same time, one might be wary of the fact that the bound in inequality (7.5) *could* increase *with* n *if we fix* $\sigma_{(C \log n)}$ *; for i.i.d. data,* $R_r^* = \Theta(1)$, *so even the first expression in the bound is of constant order. This is rather alarming, compared to the* $On^{-1/2}$) *error rate of the median. However, it should be noted that if the variances of the mixture components increase sufficiently rapidly with* n, *even the error rate of the MLE in the Gaussian case* (*which knows the distribution of each sample*) *will have a diverging error rate. Thus, although the error bounds of the modal interval estimator in Theorem 7.3.1 may be rather unsatisfactory in the case of i.i.d. data, they can lead to more meaningful error bounds when the mixture distribution involves a sizable portion of high-variance points. We will explore the question of optimality in more detail in Section 7.5.2 below.*

Guarantees for the shorth estimator are similar to the modal interval estimator. Further note that as the proofs of the results in this section reveal, the technical machinery we have developed to derive guarantees for the error of the modal interval estimator may also be used to derive estimation error bounds for the shorth estimator.

The proof of the following result is provided in Appendix E.2.2.

Theorem 7.3.3 (Theorem 4 of Pensia et al. [PJL19b]). *Recall Definitions 7.2.1 and 7.2.6 of the terms* $\sigma_{(i)}, q_{(i)}$, *and* r_k. *Suppose* $2k \geq C_{0.25} \log n$. *With probability at least* $1 - 2\exp(-c'k)$, *the shorth estimator* (*Estimator 2*) *satisfies*

$$|\widehat{\mu}_{S,k}| \leq \frac{2nr_{2k}}{k} < \frac{2n \min\left(q_{(4k)}, 2\sigma_{(4k)}\right)}{k}.$$

Remark 7.3.4. *Lemma E.2.1(iii) shows that the upper bound is actually tighter for small k: for $k' > k$, we have $kr_{2k'} > k'r_{2k}$. The smallest value permissible from our theory would be $k = \Theta(\log n)$. Also note that the upper bound in Theorem 7.3.3 for the shorth estimator resembles the bound in Theorem 7.4.3, except for the fact that the bound for the modal interval estimator involves the quantity $r_{C_{0.25}\log n}$ rather than $r_{2C_{0.25}\log n}$, and the latter could be larger depending on the spread of $\overline{P}$. Furthermore, both upper bounds in Theorem 7.3.3 may sometimes be loose: In particular, if the X_i's were i.i.d., r_{2k} would be of order $\Theta\left(\frac{k}{n}\right)$ for small k, so the bound $\frac{nr_{2k}}{k}$ would be of constant order, whereas it is known [KP90] that the shorth estimator is consistent for $k = 0.5n$.*

We now turn to theoretical guarantees from the hybrid estimator, which combines the shorth and k-median in order to obtain superior performance for both fast and slow decay of $\overline{P}$. Recall from Table 7.1 that the median has superior performance when there is less heterogeneity in the data and $\overline{P}$ decays fast enough. However, the superior performance of the modal interval estimator is apparent in the presence of large number of high-variance points. It is then desirable to have an estimator that adapts to the problem and enjoys the best of both worlds without any prior information. Indeed, as outlined in Proposition 7.3.14, the hybrid estimator achieves this rate. The key point is that if the true mean lies inside a convex set (defined with respect to the k-median), then projecting any other point (e.g., the shorth) onto the set will only move the point closer to the mean, so the hybrid estimator can leverage the better of the two rates enjoyed by the median and shorth.

Algorithm 4 specializes the hybrid estimator of Algorithm 3 to the univariate setting. The algorithm proceeds by separately computing the k_1-shorth estimator and k_2-median. If the shorth estimator lies within the median interval, the algorithm outputs the shorth; otherwise, it outputs the closest endpoint of the median interval. Note that this estimator resembles the estimator proposed by Chierichetti et al. [CDKL14] since it employs the

median as a screening step for points with very large variance. However, the shorth estimator is computed separately and then projected onto an interval around the median. In contrast, the estimator proposed by Chierichetti et al. [CDKL14] first computes the k_2-median and then computes the shorth on the remaining points, leading to a delicate conditioning argument in the analysis and creating some technical gaps in the proofs.

Algorithm 4 Hybrid mean estimator

function HYBRIDMEANESTIMATOR($X_{1:n}, k_1, k_2$)
  $S_{k_1} \leftarrow$ kMedian$(X_{1:n}, k_1)$.
  $\widehat{\mu}_{S,k_2} \leftarrow$ Shorth$(X_{1:n}, k_2)$.
  if $\widehat{\mu}_{S,k_2} \in [\min(S_{k_1}), \max(S_{k_1})]$ **then**
    $\widehat{\mu}_{k_1,k_2} \leftarrow \widehat{\mu}_{S,k_2}$
  else
    $\widehat{\mu}_{k_1,k_2} \leftarrow$ closestPoint$(S_{k_1}, \widehat{\mu}_{S,k_2})$
  end if
  return $\widehat{\mu}_{k_1,k_2}$
end function

Theorem 7.3.5 (Theorem 5 of Pensia et al. [PJL19b])**.** *Recall the Definition 7.2.6and Estimator 3 for the terms r_k, S_k, and $\widehat{\mu}_{S,k}$. If $k_1 = \sqrt{n}\log n$ and $k_2 \geq C \log n$, the error of the hybrid estimator (Estimator 4) in Algorithm 4 is bounded by*

$$|\widehat{\mu}_{k_1,k_2}| \leq \min\left(\mathrm{Diam}(S_{k_1}), |\widehat{\mu}_{S,k_2}|\right) \leq \frac{4\sqrt{n}\log n}{k_2} r_{2k_2},$$

with probability at least $1 - 2\exp(-c'k_2) - 4\exp(-c\log^2 n)$.

The proof of Theorem 7.3.5 is provided in Appendix E.2.3. Importantly, the bound in Theorem 7.3.5 is finite even for heavy-tailed distributions with infinite variance. Finally, note that in Algorithm 4, we could replace the shorth estimator by the modal interval estimator with adaptively chosen interval width to obtain similar error guarantees.

7.3.1 Examples

We illustrate this below in several cases for $r_{\log n}$, assuming Gaussian distributions for simplicity. The following examples will reappear throughout the paper to illustrate the error of our proposed estimators in various regimes of interest:

Example 7.3.6. (*i.i.d. observations*). *$P_i = \mathcal{N}(0, \sigma^2)$, so $\overline{P}$ is again $\mathcal{N}(0, \sigma^2)$.*

Example 7.3.7. (*quadratic variance*). *$P_i = \mathcal{N}(0, c^2 i^2)$, for some small $c > 0$.*

Example 7.3.8. (α*-mixture distributions*).

$$P_i = \begin{cases} \mathcal{N}(0,1), & \text{if } i \le c\lceil \log n \rceil, \\ \mathcal{N}(0, n^{2\alpha}), & \text{otherwise}, \end{cases}$$

for some $\alpha > 0$ and some large $c > 0$.

Example 7.3.8 is similar to the "contamination model" in prior work [SW09; Sti76], but with a specific scaling of variances to highlight the difference between multiple estimators by varying α. The following proposition, proved in Appendix E.3.1, will be useful in our development:

Proposition 7.3.9. *We have the following bounds for $r_{\log n}$ when $n = \Omega(1)$:*

1. *For Example 7.3.6 (i.i.d. observations), we have $r_{\log n} = \Theta\left(\frac{\sigma \log n}{n}\right)$.*
2. *For Example 7.3.7 (quadratic variance) and sufficiently small $c > 0$, we have $r_{\log n} = \Theta(1)$.*
3. *For Example 7.3.8 (α-mixture distributions) and sufficiently large $c > 0$, we have*

$$r_{\log n} = \begin{cases} \Theta\left(\frac{\log n}{n^{1-\alpha}}\right), & \text{if } \alpha < 1, \\ \Theta(1), & \text{if } \alpha \ge 1. \end{cases}$$

Note that these bounds are tighter than the ones provided by Lemma 7.2.5(v); the latter states that $r_k \leq \sigma_{(2k)}$. This is because Lemma 7.2.5(v) is a worst-case bound which does not account for the contributions of high-variance points.

7.3.1.1 Guarantees for Individual Estimators

We now revisit the examples above and calculate the bounds that follow from Theorem 7.3.1 by choosing $r = r_{C \log n}$ for a large constant $C > 0$. We also mention the cases where the bound (7.4) is weaker than the bound (7.3). The proof of the following proposition is contained in Appendix E.3.2.

Proposition 7.3.10. *Recall Definition 7.2.6 of r_k. Suppose $r = r_{C \log n}$. We have the following bounds for the modal interval estimator $|\widehat{\mu}_{M,r}|$ (Estimator 1):*

1. *For Example 7.3.6 (i.i.d. observations), we have $|\widehat{\mu}_{M,r}| \leq \Theta(\sigma)$, w.h.p.*
2. *For Example 7.3.7 (quadratic variance), we have $|\widehat{\mu}_{M,r}| \leq On^{\epsilon})$, w.h.p., for any $\epsilon > 0$. Inequality (7.4) results in a weaker bound of the form $On)$, w.h.p.*
3. *For Example 7.3.8 (α-mixture distributions), we have*

$$|\widehat{\mu}_{M,r}| = \begin{cases} On^{\alpha}), & \text{if } \alpha < 1 \\ O1), & \text{if } \alpha \geq 1, \end{cases}$$

w.h.p. For $\alpha \geq 1$, inequality (7.4) results in a weaker bound of the form $On^{\alpha})$.

Remark 7.3.11. *As discussed in Remark 7.3.2 above, the guarantees for the modal interval estimator are somewhat unsatisfactory for i.i.d. data, since Proposition 7.3.10(i) gives an error rate of $\Theta(\sigma)$, rather than the optimal rate $\Theta\left(\frac{\sigma}{\sqrt{n}}\right)$ achievable by the sample mean. On the other hand, Proposition 7.3.10 shows that for other problem settings with more widely varying variances—such as the α-mixture with $\alpha \geq 1$—the modal interval estimator results in constant*

error, whereas the sample mean would have $\Theta(n^{\alpha-0.5})$ error. These differences are summarized in more detail in Table 7.1 below.

The modal interval estimator is a "local" estimator that only considers the value of $\overline{P}_n$ in small windows. As we increase the variance of noisy points, the distribution $\overline{P}$ approaches 0 around μ^*. The modal interval estimator makes mistakes when $\overline{P}$ is flat after normalization, meaning that the density at $x + \mu^*$ is within a $(1-\epsilon)$-factor of its density at μ^*, for $\epsilon = o(1)$. If this is the case, $\overline{P}_n$ might assign higher mass at $x + \mu^*$ than μ^* due to stochasticity introduced by sampling, so a local method would mistakenly choose $x + \mu^*$ over μ^*.

More concretely, consider the setting of Example 7.3.8. If an adversary tried to alter the estimator by making the variance of the points very high $(\alpha \gg 1)$, then although $\overline{P}$ would approach 0, the normalized density would not be flat. An extreme example of this can be seen when variance of noisy points is "∞": Near μ^*, the distribution $\overline{P}$ would behave like $\mathcal{N}(\mu^*, 1)$ scaled by $O\left(\frac{\log n}{n}\right)$, which is *not* flat after normalization although $\overline{P}$ approaches 0 very rapidly, so that the mean or median would behave poorly. As Proposition 7.3.10 shows, the modal interval estimator would only suffer $O1)$ error in this case.

Remark 7.3.12. *Examining the bound in Proposition 7.3.10 for Example 7.3.8, we see the possible emergence of a "phase transition" phenomenon: For $\alpha < 1$, the modal interval estimator has error growing with n, whereas for $\alpha \geq 1$, the modal interval estimator only incurs constant error. This suggests that for $\alpha < 1$, high-variance points are more effectively hidden within the mixture distribution, so the accuracy of the modal interval estimator is more severely compromised than in the case when $\alpha \geq 1$, where the modal interval estimator can distinguish between low-variance and high-variance points. This phase transition phenomenon is established rigorously in Section 7.3.1.2 below, where we prove a lower bound of $\Omega(n^{\alpha})$ in the case when $\alpha < 1$.*

The performance of the $\Theta(\log n)$-shorth estimator is similar to the modal interval

estimator with $r = r_{\Theta(\log n)}$ (cf. inequality (7.5) in Remark 7.3.2). Consequently, the error guarantees derived for the running examples in Proposition 7.3.10 also hold for the $\Theta(\log n)$-shorth.

For completeness, we calculate the bounds of the $(\sqrt{n}\log n)$-median estimator on the recurring examples, proved in Appendix E.3.3:

Proposition 7.3.13. *We have the following bounds on the $(\sqrt{n}\log n)$-median estimator (Estimator 3):*

1. *For Example 7.3.6 (i.i.d. observations), $|\widehat{\mu}_{med,\sqrt{n}\log n}| = O\left(\frac{\sigma \log n}{\sqrt{n}}\right)$, w.h.p.*
2. *For Example 7.3.7 (quadratic variance), $|\widehat{\mu}_{med,\sqrt{n}\log n}| = On^{0.5}\log n)$, w.h.p.*
3. *For Example 7.3.8 (α-mixture distributions), $|\widehat{\mu}_{med,\sqrt{n}\log n}| = On^{\alpha-0.5}\log n)$, w.h.p.*

The following proposition translates the error guarantees of Theorem 7.3.5 into our running examples. These bounds are a direct result of Theorem 7.3.5 and Propositions 7.3.10 and 7.3.13.

Proposition 7.3.14. *When k_1 and k_2 are chosen as in Theorem 7.3.5, we have the following bounds on the hybrid estimator (Estimator 4):*

1. *For Example 7.3.6 (i.i.d. observations), $|\widehat{\mu}_{k_1,k_2}| = O\left(\frac{\sigma \log n}{\sqrt{n}}\right)$, w.h.p.*
2. *For Example 7.3.7 (quadratic variance), $|\widehat{\mu}_{k_1,k_2}| = On^{\epsilon})$, w.h.p., for any $\epsilon > 0$.*
3. *For Example 7.3.8 (α-mixture distributions), with high probability,*

$$|\widehat{\mu}_{k_1,k_2}| = \begin{cases} On^{\alpha-0.5}), & \text{if } \alpha < 1, \\ O1), & \text{if } \alpha \geq 1. \end{cases}$$

	Mean	Median	Modal/Shorth	Hybrid
Example 1(i.i.d. samples)	$\boldsymbol{n^{-0.5}}$	$\boldsymbol{n^{-0.5}}$	$n^{-1/3}$	$\boldsymbol{n^{-0.5}}$
Example 2 (quadratic variances)	$\sqrt{n}$	$\sqrt{n}$	$\boldsymbol{n^{\epsilon}}$	$\boldsymbol{n^{\epsilon}}$
Example 3 ($\alpha < 1$-mixture distributions)	$\boldsymbol{n^{\alpha-0.5}}$	$\boldsymbol{n^{\alpha-0.5}}$	n^{α}	$\boldsymbol{n^{\alpha-0.5}}$
Example 3 ($\alpha \geq 1$-mixture distributions)	$n^{\alpha-0.5}$	$n^{\alpha-0.5}$	$\boldsymbol{c}$	$\boldsymbol{c}$

Table 7.1: The table above summarizes the performance of various estimators on our three running examples. We have ignored poly-logarithmic factors for simplicity, and we use n^{ϵ} to denote On^{ϵ}) error for any $\epsilon > 0$, and c to denote an error bounded by a constant. The radius for the modal estimator and the k for the shorth estimator are adjusted to be optimal for each particular example; i.e., the estimators are assumed to know which example data are coming from. Observe that mean and median estimators outperform the modal and shorth estimators when the outliers have relatively small variances. On the other hand, the modal and shorth estimators are better when the outliers have large variances. Simulations in Section 7.9 show that the rates provided above are indeed observed in practice. Our hybrid estimator achieves the best performance in all cases *without knowing which example is under consideration.*

7.3.1.2 Phase Transition Behavior

In this subsection, we focus on verifying the statement in Remark 7.3.12 above, namely the existence of a phase transition for the modal interval estimator depending on whether $\alpha < 1$ or $\alpha \geq 1$. This phenomenon is illustrated via simulations in the plots of Figure 7.3.

For ease of analysis, we tweak the setting of Example 7.3.8 slightly: Instead of having different distributions for high variance and low variance points, we assume that the points are sampled i.i.d. from a mixture distribution, with weights resembling their original fraction in Example 7.3.8. Moreover, we assume that individual distributions are uniform rather than Gaussian.

Example 7.3.15. (*Modified α-mixture distributions*). *Let $c > 0$ be a large enough constant. For each i, $P_i = Q_n$, where*

$$Q_n = \frac{c \log n}{n} U[-1, 1] + \frac{n - c \log n}{n} U[-n^{\alpha}, n^{\alpha}],$$

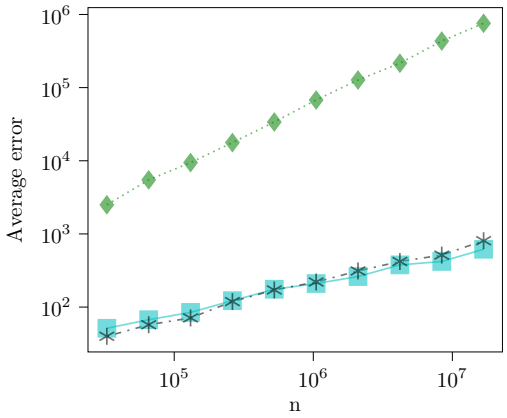

Figure 7.1: *

(a) $\alpha = 0.9$

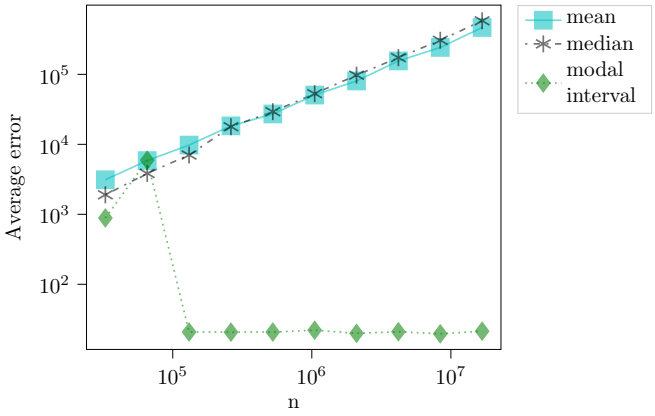

Figure 7.2: *

(b) $\alpha = 1.3$

Figure 7.3: Plots comparing average error of the mean, median, and modal interval estimators on Example 7.3.8 (α-mixture distributions) for different values of α. As shown in Proposition 7.3.10, the modal interval estimator undergoes a phase transition at $\alpha = 1$, where the error of modal interval estimator drops from the increasing function $\Omega(n^\alpha)$ to the constant function $\Theta(1)$. Moreover, as shown in Proposition 7.3.13, the median has better performance than the modal interval estimator for $\alpha < 1$, motivating our hybrid estimator. The average error, $\frac{1}{T}\sum_{i=1}^{T} |\widehat{\mu} - \mu^*|$, is calculated using $T = 200$ runs for each n. Both of the axes are on the log scale. More details can be found in Section 7.9.

and $U[-a, a]$ *is the uniform distribution on* $[-a, a]$.

Note that if we sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Q_n$, the number of points with variance $\Theta(1)$ is $\Theta(\log n)$, w.h.p. It is easy to see that the upper bounds for Example 7.3.15 are the same as that of Example 7.3.8 in Proposition 7.3.10, i.e.,

$$|\widehat{\mu}_{M,r_{C\log n}}| = \begin{cases} On^\alpha), & \text{if } \alpha < 1 \\ O1), & \text{if } \alpha \geq 1, \end{cases}$$

w.h.p. The following proposition, proved in Appendix E.3.4, establishes a lower bound of $\Omega(n^\alpha)$ on the error:

Proposition 7.3.16. *For* $\frac{1}{3} \leq \alpha < 1$ *in Example 7.3.15, the* 1*-modal interval estimator (Estimator 1) incurs* $\Omega(n^\alpha)$ *error , with a constant non-zero probability.*

Proposition 7.3.16 proves rigorously that the apparent phase transition of the modal interval estimator is not simply an artifact of the argument used to prove Proposition 7.3.10. Indeed, the modal interval estimator experiences a sharp phase transition depending on the relative variance of the mixture component with the higher variance, which is governed by the parameter α. Moreover, this phase transition is not specific to just modal interval estimator. As stated in Theorem 7.5.5, all agnostic estimators must have error $\Omega(n^{\alpha-0.5})$ for $\alpha < 1$. Thus Example 7.3.8 is indeed a difficult problem for $\alpha < 1$, but a surprisingly easy one for $\alpha > 1$.

As a final remark, note that in Examples 7.3.8 and 7.3.15, the sample median and even the mean would have an error of $\tilde{O}(n^{\alpha-0.5})$. When $\alpha < 1$, this rate is much better than the $On^{\alpha})$ guarantee of the modal interval estimator. This motivates the hybrid estimator proposed above, which is able to combine the "best of both worlds" for the modal interval and median estimators.

7.3.2 Comparison to Common Estimators

We briefly mention some common univariate estimators and contrast their performance with the performance guarantees of our proposed estimators. For simplicity, we focus on mixtures of univariate Gaussian distributions in which $\Theta(\log n)$ of the samples are drawn from distributions with bounded variance. The primary reason why the estimators mentioned below have suboptimal guarantees is because they are designed to guard against a constant fraction of arbitrarily corrupted or heavy-tailed points. In such cases, the sample median is the optimal estimator; in contrast, the sample median can be shown to be suboptimal in our setting (see Table 7.1 or Figure 7.4).

1. *Sample median:* Hallin and Mizera [HM01] established necessary and sufficient conditions for the consistency of the median for sample heterogeneous distributions. Although the sample median is more robust than the sample mean, this result

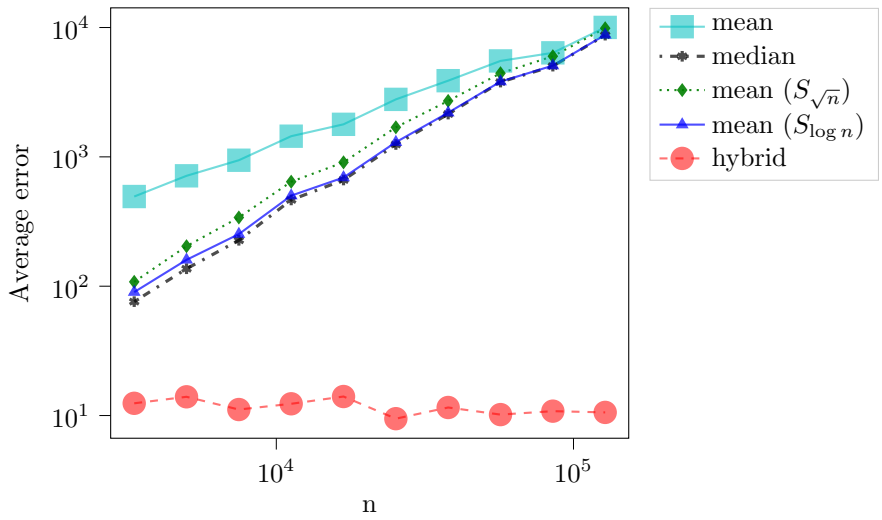

Figure 7.4: Plot comparing the average error of various estimators on Example 7.3.8 with $\alpha = 1.3$. Both of the axes are on the $\log$ scale to show the rate. As mentioned in Table 7.1, both the mean and median have an $n^{\alpha-0.5}$ error rate. The error rates of the α-trimmed mean, with $\alpha = \frac{1}{2} - \frac{\sqrt{n}}{n}$ and $\alpha = \frac{1}{2} - \frac{\log n}{n}$, are similar to the median. Note that the hybrid estimator has far superior performance. More simulations and details are available in Section 7.9.

shows that sample median is consistent if and only if $R^*_\epsilon = \omega\left(\frac{1}{\sqrt{n}}\right)$ for every $\epsilon > 0$. In particular, it implies that if the median is consistent, then $r_{\sqrt{n}} \to 0$. Focusing on particular Example 7.3.8, the error rate of the median is $O(n^{\alpha-0.5})$ (cf. Table 7.1 and Figure 7.4).

Moreover, Hallin and Mizera [HM01] established the optimality of the median among all M-estimators with score functions $\psi(\cdot)$ satisfying the following conditions:

a) $\psi(\cdot)$ is non-decreasing and skew-symmetric.

b) $\psi(\infty) = 1$.

c) The set of discontinuity points of $\psi(\cdot)$ is finite.

Therefore, one must consider broader classes of estimators beyond this family of M-estimators in order to obtain better error guarantees than the median.

2. *Huber's M-estimator* [HR09; Hub64]*:* For any finite truncation parameter, Huber's M-estimator falls in the class of M-estimators considered by Hallin and Mizera [HM01], since normalizing the score function of a bounded score function does not change the final estimate. Thus, the error rate of Huber's M-estimator cannot be any better than the error rate of the median.

3. *k-median of means:* This estimator [LM19d] divides the n data points into k disjoint blocks $(B_1, \dots, B_k)$ of equal size (assuming n/k is an integer). For each $i \in \{1, \dots, k\}$, we define Z_i to be the mean of the samples in B_i, and then define the estimator

$$\widehat{\mu} := \text{Median}_{1 \le i \le k}(Z_i) = \text{Median}_{1 \le i \le k} \left(\frac{X_{(i-1)\frac{n}{k}+1} + \cdots + X_{i\frac{n}{k}}}{n/k} \right).$$

   The median of means is robust to a constant fraction of outliers and sub-Gaussian tails even for heavy tailed i.i.d. distributions [DL22a]; however, we argue that the median of means estimator is not robust to substantial sample heterogeneity. If each $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$, then $Z_i \sim \mathcal{N}\left(\mu, \frac{\sum_{l=(i-1)k+1}^{ik} \sigma_l^2}{n^2/k^2}\right)$. Therefore, the final estimator behaves like the median of independent Gaussian samples. Furthermore, a single "high-variance" point in the set B_i can increase the variance of Z_i arbitrarily, hiding the signal from "low-variance" samples. The best case would thus be when each block contains either all "low-variance" samples or all "high-variance" samples. However, in that case, $\widehat{\mu}$ behaves essentially like the median of a smaller set with rescaled standard deviations. As argued above, regimes exist where the median estimator is suboptimal.

4. *α-trimmed mean:* Let $\alpha \in [0, 0.5)$ be such that αn is an integer. Given samples $\{X_1, \dots, X_n\}$, the α-trimmed mean [Hub64; LM19d] discards the largest and

smallest αn samples and returns the mean of the remaining $(1-2\alpha)n$ samples:

$$\widehat{\mu}_\alpha = \frac{1}{(1-2\alpha)n} \sum_{i=\alpha n+1}^{n-\alpha n} X_{(i)}.$$

The trimmed means estimator is robust to a constant fraction of outliers and has sub-Gaussian tails even for heavy-tailed distributions [LM19d]. As the fraction of "low-variance" points can be as small as $\frac{\log n}{n}$ in our sample-heterogeneous setting, the estimator $\widehat{\mu}_\alpha$ would have a large variance for any constant $\alpha > 0$. Thus, our choice of α should depend on n, going to 0.5 as $n \to \infty$.

Recall that in the definition of the k-median (cf. Estimator 3), S_k was defined as the k centermost points of the data. Thus, $\widehat{\mu}_\alpha$ is the mean of the set S_k with $k = n(1-2\alpha)$. In the extreme case of $\alpha = 0.5 - \frac{1}{2n}$, the trimmed means estimator $\widehat{\mu}_\alpha$ is the same as the median, which is not optimal. As Figure 7.4 shows, the trimmed mean behaves like the median for large α, and decreasing α (i.e., increasing k) degrades the performance. Note that we bound the error of the k-median by bounding the range of S_k (cf. Lemma E.2.4). Therefore, the bounds for the k-median also imply bounds for $\widehat{\mu}_{\frac{n-k}{2n}}$. However, the k-median primarily allows us to define a hybrid estimator by projecting onto the set S_k, which performs better than the k-median alone.

7.4 Multivariate Case

In the following sections, we derive the main results of our paper, which generalize the theorems in Section 7.3 to d dimensions.

7.4.1 Modal Interval Estimator

The following result provides an error bound for the modal interval estimator. The proof is in Appendix E.4.1.

Theorem 7.4.1. *Recall Definitions 7.2.1, 7.2.4, and Estimator 1. Suppose* $R_r^* \geq C_{0.5}\left(\frac{(d+1)\log n}{n}\right)$. *The multidimensional modal interval estimator satisfies the error bounds*

$$\|\widehat{\mu}_{M,r}\|_2 \leq 4r\left(\frac{2}{R_r^*}\right)^{\frac{1}{d}}, \tag{7.6}$$

$$\|\widehat{\mu}_{M,r}\|_2 \leq 8\sqrt{d}\sigma_{(2Cd\log n)}\left(\frac{2}{R_r^*}\right)^{\frac{1}{d}} \leq 8\sqrt{d}\left(\frac{n}{C'd\log n}\right)^{\frac{1}{d}}\sigma_{(2Cd\log n)}, \tag{7.7}$$

with probability at least $1-2\exp(-c'd\log n)$.

As the proof of Theorem 7.4.1 reveals, inequality (7.7) could also be stated using $s_{(2k)}$ in place of $2\sqrt{d}\sigma_{(2k)}$, since it is obtained from inequality (7.6) simply by substituting the bounds of Lemma 7.2.5(v). Note that when $d=1$, the bound (7.6) in Theorem 7.4.1 reduces to the bound (7.4) in Theorem 7.3.1, up to constant factors.

Remark 7.4.2. *Our bound (7.7) may be compared with Theorem 5.1 in Chierichetti et al. [CDKL14]: note that we have removed a factor of polylog(n), although their bound depends on $\sigma_{(\log n)}$ rather than $\sigma_{(d\log n)}$. Nonetheless, we emphasize the fact that our results hold for general radially symmetric distributions, whereas the proofs in Chierichetti et al. [CDKL14] are Gaussian-specific.*

Note that by Lemma 7.2.5(iii), the bound in Theorem 7.4.1 is tighter for smaller values of r. Thus, the choice of r which optimizes the bound satisfies $R_r^* = C\left(\frac{d\log n}{n}\right)$. As discussed in Pensia et al. [PJL19b] for the univariate setting, an estimator with near-optimal performance may be obtained via Lepski's method [Lep91] even without knowledge of $\overline{P}$: Define r^* to be the interval radius satisfying $R_{r^*}^* = C_{0.5}\left(\frac{(d+1)\log n}{n}\right)$, and

suppose we have rough initial estimates $r_{\min}$ and $r_{\max}$ such that $r_{\min} \leq r^* \leq r_{\max}$. Define $r_j := r_{\min} 2^j$, and define

$$\mathcal{J} := \{j \geq 1 : r_{\min} \leq r_j < 2r_{\max}\}.$$

We then define the index j_* to be

$$\min \left\{ j \in \mathcal{J} : \forall i > j \text{ s.t. } i \in \mathcal{J}, \|\widehat{\mu}_{M,r_i} - \widehat{\mu}_{M,r_j}\|_2 \leq 8 r_i \left(\frac{2n}{C_{0.5}(d+1)\log n} \right)^{1/d} \right\},$$

which may be calculated using pairwise comparisons of the modal interval estimator computed over the gridding of $[r_{\min}, r_{\max}]$. We then have the following result, proved in Appendix E.4.2:

Theorem 7.4.3. *Recall the definition of Estimator 1. With probability at least*

$$1 - 2\left(1 + \log_2\left(\frac{2r_{\max}}{r_{\min}}\right)\right)\exp(-c'\log n),$$

we have $j_* < \infty$ *and*

$$\|\widehat{\mu}_{M,r_{j_*}}\|_2 \leq 24 r^* \left(\frac{2n}{C_{0.5}(d+1)\log n} \right)^{1/d}. \tag{7.8}$$

Note that the cost of using Lepski's method is a factor of 6 in the estimation error. Finally, the following lemma shows that the shorth estimator can be used to obtain rough initial bounds on r^*:

Lemma 7.4.4. *Recall Definition 7.2.6 of* r_k. *For* $k \geq C_{0.5} d \log n$, *with probability at least* $1 - 2\exp(-ck)$, *we have* $r_{k/2} \leq \widehat{r}_k \leq r_{2k}$.

The proof of Lemma 7.4.4 is in Appendix E.4.3, and uses Lemma 7.2.8 to control the fluctuations of $\widehat{r}_k$ from its empirical counterpart. In particular, the lemma shows that we may use $r_{\min} = \widehat{r}_{C_{0.5}(d+1)\log n/2}$ and $r_{\max} = \widehat{r}_{C_{0.5}(d+1)\log n}$.

7.4.2 Shorth Estimator

We now derive error bounds for the multidimensional shorth estimator. The proof is contained in Appendix E.4.4.

Theorem 7.4.5. *Recall Definition 7.2.6 of r_k. Suppose $k \geq C(d+1)\log n$. The multidimensional shorth estimator (Estimator 2) satisfies the error bound*

$$\|\widehat{\mu}_{S,k}\|_2 \leq 4r_{2k}\left(\frac{2n}{k}\right)^{1/d},$$

with probability at least $1 - 2\exp(-c'd\log n)$.

As in the univariate case, the estimation error guarantees for the multidimensional modal interval and shorth estimators are similar. In particular, for the "optimal" choice of r such that $R_r^* = \frac{cd\log n}{n}$, inequality (7.6) in Theorem 7.4.1 gives the bound $\|\widehat{\mu}_{M,r}\|_2 = O\left(r_{c'd\log n}\left(\frac{n}{Cd\log n}\right)^{1/d}\right)$, which is of the same form as the guarantee from Theorem 7.4.5 when $k = C(d+1)\log n$.

7.4.3 Hybrid Estimator

We now prove that the hybrid estimator produces an estimator with rates of $O(\sqrt{n}^{1/d})$, rather than the rate $O(n^{1/d})$ obtained in Theorems 7.4.1 and 7.4.5. Since the overall mixture distribution is radially symmetric, all the marginal distributions are identical and symmetric about 0. Accordingly, we denote the common marginal distribution by $\overline{P}_1$, and define $r_{k,1}$ to be the smallest interval (centered at 0) that contains $\frac{k}{n}$ mass under $\overline{P}_1$.

We then have the following result, proved in Appendix E.4.5:

Theorem 7.4.6. *Recall Definition 7.2.7, Estimator 3, and Definition 7.2.6 of the terms S_k^∞, $\widehat{\mu}_{S,k}$, and r_k. Suppose $k_1 = \sqrt{n}\log n$ and $k_2 \geq Cd\log n$. Then the error of the hybrid algorithm*

(Estimator 4) is bounded by

$$\|\widehat{\mu}_{k_1,k_2}\|_2 \leq \min\left\{\mathrm{Diam}(S_{k_1}^{\infty}), \|\widehat{\mu}_{S,k_2}\|_2\right\} \leq C' \min\left\{\sqrt{d} r_{2k_1,1}, \sqrt{n}^{1/d} r_{k_2}\right\},$$

with probability at least $1 - 2\exp(-c'k_2) - 4d\exp(-c\log^2 n)$.

Remark 7.4.7. *Similar to the univariate case, the multivariate hybrid estimator achieves good error guarantees for both slow and fast decay of* $\overline{P}$*. In particular, when data are i.i.d. Gaussian with distribution* $\mathcal{N}(0, \sigma^2 I_d)$*, as in Example 7.3.6, the error of the hybrid estimator is of the order* $O\left(\frac{\sigma\sqrt{d}\log n}{\sqrt{n}}\right)$*. This is within* $\log$ *factors of the optimal* $\frac{\sqrt{d}\sigma}{\sqrt{n}}$ *error rate. At the same time, the worst-case error guarantee is of the form* $O\left(\sqrt{d}\sqrt{n}^{1/d}\sigma_{(Cd\log n)}\right)$.

We also briefly comment on the error guarantees of the hybrid estimator on the multivariate analog of Example 7.3.8. We can show that $r_{2k_1,1} = \tilde{O}\left(n^{\alpha-0.5}\right)$ *and* $r_{k_2} = \tilde{O}\left(\sqrt{d}n^{\alpha-\frac{1}{d}}\right)$*, so Lemma E.4.1 implies a bound of* $\tilde{O}(\sqrt{d}n^{\alpha-0.5})$ *for the median estimator. On the other hand, Theorem 7.4.5 leads to a bound of* $\tilde{O}(\sqrt{d}n^{\alpha})$ *for the shorth estimator. This bound can be improved for* $\alpha \geq \frac{1}{d}$*: If* $\alpha \geq \frac{1}{d}$*, we have* $\|\widehat{\mu}_{S,k_2}\|_2 = O\sqrt{d})$ *(cf. Theorem 7.5.7). The second expression in Theorem 7.4.6 then implies that the error of the hybrid estimator is* $\tilde{O}\left(\sqrt{d}\min(n^{\alpha-0.5}, 1)\right)$ *for* $\alpha \geq \frac{1}{d}$ *and* $\tilde{O}(\sqrt{d}n^{\alpha-0.5})$ *for* $\alpha \leq \frac{1}{d}$*. This improves upon the error rates of both the median and shorth estimators.*

7.5 Bounds in Expectation

Thus far, we have focused on high-probability bounds. We now briefly discuss how to convert the upper bounds into bounds on the expected error of the estimator. We then derive lower bounds on the estimation error of any estimator, thus addressing the question of optimality in certain regimes.

7.5.1 Imposing Additional Assumptions

We first show that unlike high-probability bounds, expected error bounds of a similar order *cannot* be derived for modal interval estimator without any assumptions on the high-variance mixture components. To illustrate this point, we provide a univariate example in which it is possible to derive high-probability bounds of $O1)$ for the modal interval estimator without further assumptions, whereas bounds in expectation of a similar order provably require additional tail assumptions, since $\mathbb{E}\,|\widehat{\mu}_{M,1}| \to \infty$ as $q_n \to \infty$.

Example 7.5.1. *For any n, let the densities of the P_i's be defined as follows: For $i \leq C \log n$, let*

$$p_i(x) = \begin{cases} \frac{1}{6i}, & |x| \leq 3i, \\ 0, & \textit{otherwise}. \end{cases}$$

For $i > C \log n$ and $\alpha \in (0,1)$, let

$$p_i(x) = \begin{cases} n^{-\alpha}, & |x| \leq 1, \\ h_n, & 1 < |x| \leq q_n, \\ 0, & \textit{otherwise}, \end{cases}$$

where the $\{h_n\}$ and $\{q_n\}$ are constrained such that the total area is 1, i.e., $2n^{-\alpha} + 2(q_n - 1)h_n = 1$ and $h_n \leq \frac{n^{-\alpha}}{2}$. In particular, for an $\alpha > 0$, we can still choose q_n arbitrarily large; we will take $q_n = \Omega(n)$.

The proof of the following statement is contained in Appendix E.5.1:

Proposition 7.5.2. *For Example 7.5.1, we have $\mathbb{E}\,|\widehat{\mu}_{M,1}| \to \infty$ as $q_n \to \infty$. Moreover, $|\widehat{\mu}_{M,1}| = O1)$, w.h.p.*

As seen by the example above, additional assumptions need to be imposed to prove the bounds in expectation. Suppose the variances $\{\sigma_i\}$ are all finite. We will consider two types of assumptions: either (i) "high-noise" points do not have very large variances, or (ii) "low-noise" points have small support.

We state a result for the modal interval estimator in d dimensions; similar proofs hold for the shorth, median, and hybrid estimators. The following result is proved in Appendix E.5.2.

Theorem 7.5.3. *Recall Definitions 7.2.1, 7.2.4, and Estimator 1 for the terms $R_r^*, \sigma_{(i)}$, and $\widehat{\mu}_{M,r}$. Let $nR_r^* = \Omega(d \log n)$. The following upper bounds hold for the expected error of the modal interval estimator:*

(*i*) *Suppose*

$$\log\left(\frac{\sigma_{(n)}}{r}\right) = O\left(nR_r^*\right). \tag{7.9}$$

Then the modal interval estimator satisfies the expected error bound

$$\mathbb{E}\,\|\widehat{\mu}_{M,r}\|_2 = O\left(r\left(\frac{c}{R_r^*}\right)^{1/d}\right).$$

(*ii*) *In the case $d = 1$, suppose the support of $\Omega(nR_r^*)$ points lies in $[-r, r]$. Then*

$$\mathbb{E}\,|\widehat{\mu}_{M,r}| = O\left(\frac{r}{R_r^*}\right).$$

Remark 7.5.4. *The condition (7.9) in Theorem 7.5.3(i) can be translated into the inequality $\sigma_{(n)} \leq r\exp(CnR_r^*)$, and provides an upper bound on the variance of the worst mixture components. If we choose $r = \sigma_{(d\log n)}$, we obtain the requirement that $\sigma_{(n)}$ is at most a factor of On^{Cd}) larger than the variance $\sigma_{(d\log n)}$ of the "good" points. This can be compared to the assumption $\sigma_{(n)} = \sigma_{(1)}$poly$(n)$ imposed by Chierichetti et al. [CDKL14] when proving upper bounds on expected error in the univariate case. As the proof of Theorem 7.5.3 reveals, we could also convert*

the tighter version of the estimation error guarantee (cf. Theorem 7.3.1 in the univariate setting) into an expected error bound in a similar manner: If condition (7.9) holds in Theorem 7.5.3 and we additionally assume that $r' = \Omega(r)$, then $\mathbb{E}\,|\widehat{\mu}_{M,r}| = Or')$.

Note that the condition in Theorem 7.5.3(ii) imposes no constraints on the behavior of the large-variance mixture components. The proof proceeds by integrating the tail probability of the modal interval estimator, and showing that it must decay sufficiently quickly by considering the mass of intervals lying far from the true mean. An extension to the multivariate case is possible, but would require somewhat more refined technical analysis.

7.5.2 Minimax Bounds

We are now ready to discuss the optimality of our hybrid estimator, which we will consider in the context of expected error bounds. We state our results in the case of a general dimension $d \geq 1$. The goal of this section is to describe a general setting in which it is possible to show that the hybrid estimator is (nearly) minimax optimal.

We will consider the class of distributions $\mathcal{P}(\sigma_1, \sigma_2, p)$, containing symmetric, unimodal distributions $\{P_i\}_{i=1}^n$ with common mean μ, such that at least np distributions have marginal variance bounded by σ_2^2 and the remaining distributions have marginal variance bounded by σ_1^2. Note that σ_1, σ_2, and p may all be functions of n, e.g., $p = \frac{\log n}{n}$.

We call an algorithm agnostic if applying the algorithm does not require knowledge of the variance of individual points (e.g., the sample mean or median). We have the following minimax lower bound, proved in Appendix E.5.3:

Theorem 7.5.5. *Suppose $p \leq \frac{1}{3}$, $\sigma_2 \leq \sigma_1$, and $p = \Omega\left(\frac{\log n}{n}\right)$.*

(*i*) *The minimax error of any agnostic algorithm is*

$$\min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathcal{P}(\sigma_1, \sigma_2, p)} \mathbb{E}\left[\|\widehat{\mu} - \mu\|_2\right] \geq C_\ell \sqrt{d} \min\left\{\frac{\sigma_2}{\sqrt{np}}, \frac{\sigma_1}{\sqrt{n}}\right\}. \tag{7.10}$$

(*ii*) *In the case $d = 1$, suppose in addition we have*

$$\frac{\sigma_1}{\sigma_2} = O\left(\frac{1}{np^2}\right). \tag{7.11}$$

Then the algorithm of any agnostic algorithm satisfies that

$$\min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathcal{P}(\sigma_1, \sigma_2, p)} \mathbb{E}\left[\|\widehat{\mu} - \mu\|_2\right] \geq \frac{C'_\ell \sigma_1}{\sqrt{n}}. \tag{7.12}$$

Remark 7.5.6. *In the $d = 1$ case, the lower bound in Theorem 7.5.5 when condition (7.11) is satisfied matches the lower bound derived by Chierichetti et al. [CDKL14]. On the other hand, our proof technique is somewhat more direct and proceeds via a straightforward (albeit lengthy) calculation.*

We now state our general upper bound, achieved by the hybrid estimator. Under the specific regimes, we impose mild regularity conditions on the distributions to obtain cleaner expressions:

(i) Let $q_i(x)$ denote the marginal distribution of P_i, where $q_i : \mathbb{R} \to \mathbb{R}$ (since P_i is radially symmetric, all marginals are equal). Let ν_i^2 denote the marginal variance of P_i. Then

$$q_i(\nu_i) \geq \frac{c}{\nu_i}. \tag{7.13}$$

(ii) Let each density be written as $p_i(x) = f_i(\|x\|_2)$, where $f_i : \mathbb{R} \to \mathbb{R}$ is a decreasing

function on the positive reals. Then

$$f_i(0) \le \left(\frac{c'}{\nu_i}\right)^d, \quad \text{and} \quad \int_{B(K\sqrt{d}\nu_i, 2\sqrt{d}\nu_i)} p_i(y)dy \le C_1 \exp\left(-C_2 K^2\right), \quad \forall K \ge C_3 > 1. \tag{7.14}$$

Condition (7.13) assumes that the marginal densities do not decrease too rapidly around the mean, and implies the accuracy of the median filtering step. Condition (7.14) assumes that the joint densities do not have too much mass concentrated around any single point (e.g., the mean), from which we may derive tighter error bounds on the shorth estimator when we have sufficiently separated variances, i.e., $\frac{\sigma_1}{\sigma_2} = \Omega\left(n^{1/d}\right)$. Note that conditions (i) and (ii) hold for Gaussian distributions; furthermore, condition (ii) holds more broadly when the norm of $p_i(\cdot)$ has right $c'\nu_i\sqrt{d}$-sub-Gaussian tails around $\sqrt{d}\nu_i$. Then this expression can be upper bounded by $\mathbb{P}\{\|X\| - \sqrt{d}\nu_i \ge cK\sqrt{d}\nu_i\} \le \exp(-c'K^2)$ using the sub-Gaussian assumption.

We also define $\mathcal{Q}(\sigma_1, \sigma_2, p)$ to be the class of symmetric, unimodal distributions with $\{P_i\}_{i=1}^n$ with common mean μ, such that at least np distributions have marginal variances bounded by σ_2^2 and remaining distributions have marginal variance at least $\Omega(\sigma_1^2)$ and at most σ_1^2. Thus, $\mathcal{Q}(\sigma_1, \sigma_2, p)$ is the class of distributions with sufficient division between high-variance and low-variance points, and we clearly have $\mathcal{Q}(\sigma_1, \sigma_2, p) \subseteq \mathcal{P}(\sigma_1, \sigma_2, p)$. Finally, in order to derive bounds in expectation, we impose the additional growth condition (7.9) on the variance of the mixture components.

The following result is proved in Appendix E.5.4.

Theorem 7.5.7. *If $p = \Omega\left(\frac{d \log n}{n}\right)$ and condition (7.13) holds, then the hybrid estimator satisfies the upper bound*

$$\max_{\{P_i\} \subseteq \mathcal{P}(\sigma_1, \sigma_2, p)} \mathbb{E}\left[\|\widehat{\mu} - \mu\|_2\right] \le C_u \sqrt{d} \min\left\{\sqrt{n}^{1/d}\sigma_2, \frac{\log n}{\sqrt{n}}\sigma_1\right\}. \tag{7.15}$$

We also have the following special cases if we impose additional assumptions:

(*a*) *If* $p = \Omega\left(\frac{\sqrt{n}\log n}{n}\right)$, *we have the tighter bound*

$$\max_{\{P_i\}\subseteq\mathcal{P}(\sigma_1,\sigma_2,p)} \mathbb{E}\left[\|\widehat{\mu}-\mu\|_2\right] \leq C'_u\sqrt{d}\min\left\{\frac{\log n}{p\sqrt{n}}\sigma_2, \frac{\log n}{\sqrt{n}}\sigma_1\right\}. \tag{7.16}$$

(*b*) *If* $\frac{\sigma_1}{\sigma_2} = \Omega\left(n^{\frac{1}{d}}\right)$ *and condition* (7.14) *holds, then*

$$\max_{\{P_i\}\subseteq\mathcal{Q}(\sigma_1,\sigma_2,p)} \mathbb{E}\left[\|\widehat{\mu}-\mu\|_2\right] \leq C''_u\sqrt{d}\min\left\{\sigma_2\sqrt{\log n}, \frac{\log n}{\sqrt{n}}\sigma_1\right\}. \tag{7.17}$$

It is instructive to compare the upper bounds for the hybrid estimator in Theorem 7.5.7 with the lower bounds derived in Theorem 7.5.5. (Note that the same class of distributions used to obtain the minimax lower bounds over $\mathcal{P}$ falls into the class $\mathcal{Q}$, so the upper bounds in Theorem 7.5.7 may be directly compared with the lower bounds in inequality (7.17), as well.) In particular, we can see that the hybrid estimator is nearly minimax optimal in three somewhat different regimes of interest, which can be derived directly from the bounds in the theorems. The results are summarized in Table 7.2:

1. Large heterogeneity: When σ_1 is very large compared to σ_2 and p is very small (still satisfying $p = \Omega(\frac{d\log n}{n})$), a direct application of the median would lead to large error. However, the shorth estimator is able to focus on the low-variance points due to the sufficiently large separation in variances. As p becomes smaller, the gap between the upper and lower bounds reduces, reaching within $\log n$ factors when $p = \Theta\left(\frac{d\log n}{n}\right)$.

2. Mild heterogeneity: Since σ_1 is relatively small, the median and even mean are minimax optimal. The hybrid estimator is able to achieve these rates (including the i.i.d. case).

3. Large p: As p increases, the number of good points increase and we expect to obtain vanishing error for reasonable values of σ_1 (e.g., under condition (7.9)). Indeed, the hybrid estimator achieves vanishing error for large $p = \Omega\left(\frac{\sqrt{n}\log n}{n}\right)$ irrespective of the magnitude of σ_1. Also, the gap between the upper bound and the lower bound decreases as either $p \to 1$ or $\sigma_1 \to \sigma_2$.

	Large heterogeneity $\mathcal{Q}(\Omega(p^{-0.5} + n^{1/d}), 1, o(n^{-0.5}))$	Mild heterogeneity $\mathcal{P}(O(p^{-0.5}), 1, p)$	Large p $\mathcal{P}(\sigma_1, 1, \Omega(n^{-0.5}\log n))$
Hybrid estimator	$\sqrt{d}$	$\frac{\sigma_1\sqrt{d}}{\sqrt{n}}$	$\min\left\{\frac{\sqrt{d}}{p\sqrt{n}}, \frac{\sigma_1\sqrt{d}}{\sqrt{n}}\right\}$
Lower bound	$\frac{\sqrt{d}}{\sqrt{np}}$	$\frac{\sigma_1\sqrt{d}}{\sqrt{n}}$	$\min\left\{\frac{\sqrt{d}}{\sqrt{pn}}, \frac{\sigma_1\sqrt{d}}{\sqrt{n}}\right\}$

Table 7.2: Comparison of upper and lower bounds for estimation error, given by Theorems 7.5.5 and 7.5.7, in three regimes of interest. In all of these cases, we assume $p = \Omega\left(\frac{d\log n}{n}\right)$. For simplicity, we set $\sigma_2 = 1$ and ignore multiplicative factors which are logarithmic in n. We provide more details regarding these calculations in Appendix E.5.5.

Remark 7.5.8. *Although we have shown that the hybrid estimator is indeed optimal in several diverse regimes, the preceding discussion leaves open the question of optimality in other settings. In particular, although our general upper bounds (e.g., inequality (7.15)) suggests the presence of a $\sqrt{n}^{1/d}$ factor when using the hybrid estimator, our lower bound techniques do not show that such a factor is unavoidable for $d \geq 2$. As argued by Chierichetti et al. [CDKL14], a factor of $\sqrt{n}$ is unavoidable in $d = 1$ (cf. Theorem 7.5.5).*

7.6 Computation in High Dimensions

We now discuss how to make our estimators computationally feasible when d is large. The main idea is that both the modal interval and shorth estimators involve finding

optimal balls in $\mathbb{R}^d$. To save on computation, we will show that restricting the search to balls centered at one of the n data points leads to estimators with similar performance guarantees. This is an idea previously introduced in the literature on mode estimation in i.i.d. scenarios [ABC04; DK14; Jia17].

Concretely, the modal interval and shorth estimators are replaced by:

Estimator 5. *The computationally efficient modal interval estimator is defined by*

$$\widetilde{\mu}_{M,r} := \arg \max_{x \in \{x_1, \ldots, x_n\}} R_n(f_{x,r}). \tag{7.18}$$

Estimator 6. *The computationally efficient shorth estimator is defined by*

$$\widetilde{r}_k := \inf_r \sup_{x \in \{x_1, \ldots, x_n\}} \left\{ R_n(f_{x,r}) \geq \frac{k}{n} \right\}, \qquad \widetilde{\mu}_{S,k} := \widetilde{\mu}_{M,\widetilde{r}_k}. \tag{7.19}$$

In other words, we select the data point such that the smallest ball centered around that point containing at least k points has the minimum radius.

Note that both estimators (7.18) and (7.19) may be computed in $O(n^2 d)$ time. In contrast, computing the modal interval or shorth estimators directly would correspond to solving the circle placement problem or smallest enclosing ball problem, for which the best-known exact algorithms are $\Omega(n^d)$ [LP84; EE94; AS98].

Using a peeling argument [Van00], we can obtain a more refined concentration result than Theorem 7.2.8. The proof of the following result is contained in Appendix E.1.3. Note that the proof critically leverages radial symmetry of R, whereas the concentration inequality in Lemma 7.2.8 does not require R to be radially symmetric.

Lemma 7.6.1. *Recall Definitions 7.2.2, 7.2.4 for the terms $f_{x,r}$, $R_n(\cdot)$, and $R(\cdot)$. For any*

$t \in (0, 1]$, radii $\bar{r}, r > 0$, and $n > 1$, we have the following inequalities:

$$\mathbb{P}\left(|R_n(f_{x,r}) - R(f_{x,r})| \leq 2tR(f_{x,r}), \quad \forall x \text{ s.t. } \|x\|_2 \leq \bar{r}\right) \geq 1 - \frac{2\exp(-cnt^2 R(f_{\bar{r},r}))}{1 - \exp(-cnt^2 R(f_{\bar{r},r}))}, \tag{7.20}$$

$$\mathbb{P}\left(\sup_{\|x\|_2 \geq \bar{r}} |R_n(f_{x,r}) - R(f_{x,r})| \geq tR(f_{\bar{r},r})\right) \leq 2\exp(-cnt^2 R(f_{\bar{r},r})), \tag{7.21}$$

provided $\bar{r}$ and r are such that $R(f_{\bar{r},r}) \geq \frac{C_t d \log n}{n}$.

Using Lemma 7.6.1, we can derive the following results for the computationally efficient modal interval and shorth estimators. The proof is contained in Appendix E.4.6.

Theorem 7.6.2. *Recall Definition 7.2.6 of the term r_k. For the computationally efficient estimators, we have the following error guarantees:*

(*i*) *Suppose $r \geq 2r_{6Cd\log n}$. Then the modal interval estimator satisfies the bound $\|\widetilde{\mu}_{M,r}\|_2 \leq 4r\left(\frac{n}{Cd\log n}\right)^{1/d}$, with probability at least $1 - 6\exp(-c_3 d \log n)$.*

(*ii*) *Suppose $k \geq 2C_{0.5}(d+1)\log n$. Then the shorth estimator satisfies the bound $\|\widetilde{\mu}_{S,k}\|_2 \leq 4r_{2k}\left(\frac{2n}{k}\right)^{1/d}$, with probability at least $1 - 2\exp(-c'k)$.*

Remark 7.6.3. *Comparing Theorem 7.6.2(i) with Theorem 7.4.1, we see that the the computationally efficient modal interval essentially incurs an additional factor of 2 in the error bound, since we require $r \geq 2r_{C'd\log n}$. If we take $k = Cd\log n$, the error bound in Theorem 7.6.2(ii) is very similar to the error guarantee for the modal interval estimator (7.7) derived in Theorem 7.4.5, except for an extra factor of 2.*

Of course, the quality of the guarantee in Theorem 7.6.2(i) worsens as r increases. As discussed in Section 7.4.1, we can use Lepski's method to calibrate the modal interval radius. Note that we can again use the shorth estimator to obtain rough upper and lower bounds. Using a similar argument as in the proof of Lemma 7.4.4, we are guaranteed that

$\frac{1}{2}\widetilde{r}_{3Cd\log n} \leq r_{6Cd\log n} \leq \widetilde{r}_{6Cd\log n}$, w.h.p. Essentially the same argument as in Theorem 7.4.3 then shows that the error of the modal interval estimator with Lepski calibration is guaranteed to be upper-bounded by $12r_{6Cd\log n}\left(\frac{n}{Cd\log n}\right)^{1/d}$.

As discussed in Section 7.4.3, the projection step for the hybrid screening procedure can be computed in $O(d)$ time. The construction of the cuboid S_k^∞ itself can clearly be computed in $O(nd)$ time. Thus, one can also easily obtain the $O(\sqrt{n}^{1/d})$ rates using a computationally efficient hybrid estimator, as well.

7.7 Relaxing Radial Symmetry

We now consider the case when the population-level distribution $\overline{P} = \frac{1}{n}\sum_{i=1}^n P_i$ is not symmetric. In the case $d = 1$, we can obtain the same estimation error rates only assuming that density p_i is log-concave with a unique mode at 0. In the case $d > 1$, we can obtain weaker estimation error guarantees of the order $O(\sqrt{n})$ rather than $O(\sqrt{n}^{1/d})$ if we only assume that the mixture components are centrally symmetric. Furthermore, it is possible to obtain $O(n^{1/d})$ rates if we assume that a certain fraction of the components are radially symmetric.

Although radial symmetry is a strict assumption, it provides us an $O\sqrt{n}^{\frac{1}{d}})$ error. Whereas if we just assume central asymmetry, a union bound argument gives $O\sqrt{dn})$ error. This factor of $O\sqrt{dn})$ can not be improved in general. To see this, note that there exists a problem instance in single dimension where the lower bound is a factor of $\tilde{\Omega}(\sqrt{n})$. Central symmetry allows for having the same "hard" problem on each dimension separately, forcing an $\tilde{\Omega}(\sqrt{n})$ error in each dimension.

We can relax the radial symmetry assumptions slightly. In particular, Theorem 7.2.8 only relies on the fact that R_r^*, the mass of the interval centered around the true mode 0, is $\Omega\left(\frac{\log n}{n}\right)$ (with no additional symmetry assumptions). We do need $R(f_{x,r})$ to satisfy some additional monotonicity assumptions along rays as x moves away from 0.

7.7.1 General Theory

In place of radial symmetry, we impose the following condition (stated with respect to a fixed radius r):

(C1) The population-level quantity $R(f_{x,r})$ is maximized at $x = 0$, and otherwise monotonically decreasing along rays from the origin.

Note that condition (C1) is satisfied if the same property holds for all components p_i in the mixture. We now define the function

$$g(a, r) := \sup_{\|x\|_2 = a} R(f_{x,r}), \tag{7.22}$$

for $a, r > 0$. By Lemma 7.2.5, we can argue that under radial symmetry of R, we have $g(a, r) \leq \frac{1}{N(B_a, r)} \leq \left(\frac{r}{a}\right)^d$, which can then be plugged into the argument of Theorem 7.4.1. The proof of the following statement is contained in Appendix E.6.1.

Theorem 7.7.1. *Suppose condition (C1) holds.*

(*i*) *Recall Definition 7.2.4 of R_r^*. Suppose r is such that $R_r^* = \Omega\left(\frac{d \log n}{n}\right)$, and r' is chosen sufficiently large such that $g(r', r) < \frac{R_r^*}{2}$. Then the modal interval estimator satisfies $\|\widehat{\mu}_{M,r}\|_2 \leq r'$, w.h.p.*

(*ii*) *Recall Definition 7.2.6 of r_k. Suppose r' is chosen such that $g(r', r_{8d \log n}) \leq \frac{8d \log n}{4n}$. With high probability, the error of the shorth estimator satisfies $\|\widehat{\mu}_{S,k}\|_2 \leq r'$, and the error of the hybrid algorithm with $k_2 = r_{8d \log n}$ is bounded by $\min(r', \sqrt{d} r_{4\sqrt{n \log n}, 1})$.*

Remark 7.7.2. *For radially symmetric distributions, note that $g(r', r) \leq \left(\frac{r}{r'}\right)^d$, so we can take $r' = r\left(\frac{2}{R_r^*}\right)^{1/d}$ and $r' = r_{2k}\left(\frac{4}{R_{2k}^*}\right)^{1/d}$ to obtain the results of Theorems 7.4.1 and 7.4.5 for the modal interval and shorth estimators, respectively. Furthermore, by Lemma 7.2.5(iii), we have $r_{\sqrt{n} \log n} \leq \left(\frac{\sqrt{n}}{8d}\right)^{1/d} r_{8d \log n}$. Thus, we also recover the analog of Theorem 7.4.6 for the hybrid estimator.*

Finally, note that an analog of Theorem 7.7.1 holds when we use the computationally efficient modal interval and shorth estimators described in Section 7.6, with minor proof modifications.

7.7.2 Sufficient Conditions

Condition (C1) may be a bit difficult to interpret. We define two related conditions:

(C2) Each component density p_i is log-concave with a unique mode at 0. Recall that a distribution with density p is log-concave if $p(x) \propto e^{-\phi(x)}$ for a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$.

(C3) For all $x \in \mathbb{R}^d$ and all $1 \leq i \leq n$, we have $p_i(x) = p_i(-x)$.

Note that condition (C3) only requires symmetry of the density around 0, rather than radial symmetry; in particular, it holds for Gaussian distributions that are not necessarily isotropic.

We have the following result, proved in Appendix E.6.2:

Proposition 7.7.3. *Suppose conditions (C2) and (C3) hold. Then condition (C1) also holds. Furthermore,* $g(a, r) \leq \frac{1}{\lfloor a/2r \rfloor}$.

In fact, we can even derive a result only assuming condition (C2) in the case $d = 1$. As argued in the proof of Theorem 7.7.1, we may establish that $R(f_{\widehat{\mu}_{M,r}}, r) \geq \frac{R_r^*}{2}$, w.h.p. Thus, there exists some i such that $R_i(f_{\widehat{\mu}_{M,r}}, r) \geq \frac{R_r^*}{2}$. By properties of log-concave convolutions (cf. proof of Proposition 7.7.3), we know that $R_i(f_{x,r})$ is decreasing along rays originating from some point x_i^*, and also $\|\widehat{\mu}_{M,r} - x_i^*\|_2 \leq \frac{4r}{R_r^*}$, since we could otherwise pack too many intervals into the ray between x_i^* and $\widehat{\mu}_{M,r}$, thus contradicting the inequality $R_i(f_{\widehat{\mu}_{M,r}}, r) \geq \frac{R_r^*}{2}$. Finally, note that due to the unimodality of p_i at 0, we clearly have

$\|x_i^*\|_2 \leq r$. Altogether, we obtain the error bound

$$\|\widehat{\mu}_{M,r}\|_2 \leq \frac{4r}{R_r^*} + r,$$

which is of the same order as the guarantees in Theorem 7.3.1. A similar conclusion could be reached if we replaced condition (C2) by the condition that each p_i has a unique median and mode at 0, since $R_i(f_{x,r})$ is decreasing along rays originating from r $(-r)$ in the positive (negative) direction.

7.7.3 Examples

We now describe two examples to illustrate concrete use cases of our more general theory.

Example 7.7.4 (Elliptically symmetric distributions)**.** *We now consider the case where the components of the mixture are not spherical, but have the same axes of symmetry. Concretely, suppose that for a fixed matrix $\Sigma \succ 0$, the density of each X_i is of the form $f_i\left((x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$, where $f_i : \mathbb{R} \to \mathbb{R}$ is a decreasing function defined on the positive reals. The goal is to estimate the common parameter $\mu \in \mathbb{R}^d$. As a specific example, we might have a mixture of nonisotropic Gaussian distributions where the covariance matrices are all scalar multiples of Σ. This strictly generalizes the case of radially symmetric distributions, which corresponds to the case $\Sigma = I$.*

Suppose we employ the modal interval, shorth, or hybrid estimators described above. Note that these estimators do not require knowledge of the matrix Σ. We wish to analyze the behavior of the quantity $g(a, r)$ defined in equation (7.22), which is relevant for Theorem 7.7.1. Indeed, we can derive an analog of Lemma 7.2.5 that applies in this setting. The main step is to understand bound the quantity $g(r_2, r_1)$ when $r_1 < r_2$. We have the following result, proved in Appendix E.6.3:

Proposition 7.7.5. *Let $r_1 < r_2$. For an elliptically symmetric distribution, we have*

$$g(r_2, r_1) \leq C \left(\frac{r_1 \lambda_{\max}(\Sigma)}{r_2 \lambda_{\min}(\Sigma)} \right)^d.$$

Clearly, taking $C = 1$ and $\Sigma = I$ in Proposition 7.7.5 recovers the result for radially symmetric distributions.

Remark 7.7.6. *Similar arguments as in Example 7.7.4 could be applied in the case when the probability density functions of the distributions are proportional to $\exp(-\|x - \mu\|/\sigma)$, for a different norm $\|\cdot\|$ besides the squared ℓ_2-norm or the Mahalanobis norm. Also note that if the matrix Σ (accordingly, the norm $\|\cdot\|$) were known a priori, it might be possible to obtain better rates by using a modal interval/shorth estimator based on the level sets of the norm rather than spheres of varying radii.*

Example 7.7.7 (Mixture of radially and centrally symmetric distributions)**.** *For another interesting special case, suppose we have s points drawn from radially symmetric distributions, and $n - s$ points drawn from centrally symmetric distributions. Suppose we have $f(n)$ points which are well-behaved in the sense that the interquartile range of the corresponding distributions is small. (These distributions need not coincide with the radially symmetric distributions.) We have the following result, proved in Appendix E.6.4:*

Proposition 7.7.8. *For $r = q_{(f(n))}$ and $r' = 2rn^{1/d}$, we have*

$$g(r', r) \leq \frac{R_r^*}{2},$$

provided $s \geq n - 2n^{1/d}(f(n) - 4)$.

Thus, as the proportion of well-behaved points increases, the required proportion of radially symmetric distributions required to obtain a specific error guarantee becomes smaller. In particular, if $f(n) = \Omega(n^{1-1/d})$, we do not need any radially symmetric distributions; recall, how-

ever, that the coordinatewise median already performs well on a mixture of centrally symmetric distributions if $f(n) = \Omega(\sqrt{n}\log n)$.

7.8 Linear Regression

We now shift our focus to the problem of linear regression, and demonstrate how the methodology developed thus far may be adapted to parameter estimation in multivariate regression. Suppose we have observations $\{(x_i, y_i)\}_{i=1}^n$ from the linear model

$$y_i = x_i^\top \beta^* + \epsilon_i, \qquad \forall 1 \le i \le n, \tag{7.23}$$

where the pairs $\{(x_i, \epsilon_i)\}_{i=1}^n$ are independent but not necessarily identically distributed, and x_i and ϵ_i are independent for each i.

Following the theme of our paper, we assume that the probability density function of ϵ_i's are symmetric and unimodal. We want to study the behavior of the modal interval regression estimator

$$\widehat{\beta} \in \operatorname*{argmax}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n 1\left\{|y_i - x_i^\top \beta| \le r\right\}, \tag{7.24}$$

for an appropriate choice of $r > 0$.

A natural question is whether the true parameter β^* is the unique population-level maximizer in the regression setting. As the following proposition shows, this is indeed the case when the densities of the x_i's are absolutely continuous with respect to Lebesgue measure. The proof is contained in Appendix E.7.1.

Proposition 7.8.1. *Consider the linear model in equation* (7.23)*, where the distributions of* x_i*'s and* ϵ_i*'s have Lebesgue density. Then the population-level maximizer is given by*

$$\beta^* = \arg\max_{\beta} \sum_{i=1}^n \mathbb{E}\left[1\left\{|y_i - x_i^\top \beta| \le r\right\}\right], \qquad \forall r > 0. \tag{7.25}$$

Importantly, Proposition 7.8.1, and the ensuing theory, does not require specific assumptions on the form of the distribution of the x_i's. However, in order to derive easily interpretable error bounds on the modal interval regression estimator, we will assume further distributional assumptions (cf. the statement of Theorem 7.8.3 below).

7.8.1 Estimation Error

In order to obtain error bounds on $\|\widehat{\beta} - \beta^*\|_2$, we need to analyze the behavior of the quantities

$$R_\beta := \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(|y_i - x_i^\top \beta| \leq r\right),$$

for a fixed value of r, chosen sufficiently large that $R_{\beta^*} \geq \frac{Cd\log n}{n}$. In particular, we want to show that for $\|\beta - \beta^*\|_2$ larger than a certain value, we will have $R_\beta < \frac{R_{\beta^*}}{2} = \frac{1}{2n}\sum_{i=1}^{n} \mathbb{P}(|\epsilon_i| \leq r)$.

As before, the key ingredient for deriving error bounds is a uniform concentration result. This is proved in the following lemma:

Lemma 7.8.2. *Let $t \in (0, 1]$, and suppose r is large enough so that $R_{\beta^*} \geq \frac{Cd\log n}{n}$. Then*

$$\mathbb{P}\left(\sup_{\beta\in\mathbb{R}^d, r'\leq r}\left|\frac{1}{n}\sum_{i=1}^{n} 1\left\{|y_i - x_i^\top\beta| \leq r'\right\} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[1\left\{|y_i - x_i^\top\beta| \leq r'\right\}\right]\right| \geq tR_{\beta^*}\right)$$
$$\leq 2\exp(-cnR_{\beta^*}t^2). \quad (7.26)$$

Since the proof is directly analogous to the proof of Theorem 7.2.8, we only provide a sketch: The key point is to consider the VC dimension of the class of functions $f(x, y) = 1\{|y - x^\top\beta| \leq r\}$, indexed by the pair (β, r). Note that the subset of points in $\mathbb{R}^{d+1}$ associated with the indicator function $f(x, y)$ is an intersection of two halfspaces. Using results on the VC dimension of an intersection of concept classes [VW09], we see that the VC dimension of the desired hypothesis class is bounded by $C'd$. The concentration

result then follows by the same arguments used to derive Theorem 7.2.8.

It is generally difficult to state general bounds on estimation error that depend only on order statistics of quantiles, since as in the case of mean regression, the error bounds one can derive will be largely problem-dependent. In order to simplify our presentation, we will only discuss the case where the ϵ_i's and x_i's are Gaussian: $\epsilon_i \sim N(0, \sigma_i^2)$ and $x_i \sim N(\mu_i', \Sigma_i')$. We have the following result, proved in Appendix E.7.2:

Theorem 7.8.3. *Let $\lambda_{\min} := \min_i \lambda_{\min}(\Sigma_i')$, and suppose $\lambda_{\min} > 0$. Suppose $r > 0$ is chosen such that $R_{\beta^*} \geq \frac{Cd \log n}{n}$. Then the regression estimator (7.24) satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \leq \frac{c' n \sigma_{(cd \log n)}}{\sqrt{\lambda_{\min}}},$$

w.h.p.

We conjecture that it is possible to decrease this upper bound to $O\sqrt{n}\sigma_{(c \log n)})$ by an appropriate hybrid screening procedure, but we leave this to future work. Also note that in order for the bound in Theorem 7.8.3 to be useful, the quantity $\lambda_{\min}$ must either be a constant, or else not decrease too rapidly with n.

7.8.2 Computation

A natural question is whether the modal interval regression estimator (7.24) is actually computationally feasible. We claim that an estimator may be obtained in $O(n^d)$ time, using Algorithm 5. The proof is in Appendix E.7.3.

Theorem 7.8.4. *The output of Algorithm 5 is a maximizer of equation (7.24).*

Remark 7.8.5. *Correct application of Algorithm 5 would assume that r is chosen appropriately. It is less clear how this parameter might be calibrated based on the data, perhaps using an appropriate variant of Lepski's method. We leave this important open question to future work.*

Algorithm 5 Modal interval regression estimator

function MODALINTERVALREGRESSION($X_{1:n}, Y_{1:n}, r, d$)
  Construct the set of hyperplanes

$$\mathcal{S}_r = \{y_i = x_i^\top \beta + r\} \bigcup \{y_i = x_i^\top \beta - r\}.$$

  Let $\{S_1, \ldots, S_N\}$ denote the set of subsets of $\mathcal{S}_r$ of cardinality d.
  for $j = 1, \ldots, N$ **do**
    Solve the system of linear equations given by S_j. Let β_j be a solution (if one exists).
  end for
  $j^* \leftarrow \arg\max_{1 \le j \le N} \frac{1}{n} \sum_{i=1}^n 1\left\{|y_i - x_i^\top \beta_j| \le r\right\}$.
  return β_{j^*}
end function

7.9 Simulations

We now present the results of simulations on the recurring examples to validate our theoretical predictions (cf. Table 7.1). Although our theorem statements involve large constants, we empirically observe that smaller constants suffice to elicit the same behavior predicted by our theory. We run the k-shorth estimator with $k = 5d \log n$ and k-median with $k = \sqrt{n} \log n$. We use these estimators for the hybrid estimator, i.e., the $(\sqrt{n} \log n, 5d \log n)$-hybrid estimator. The mean estimator corresponds to the simple average, whereas the median estimator refers to the (coordinatewise) sample median.

For each n, we run $T = 200$ simulations for univariate data and $T = 20$ simulation for multivariate data and report the average error $\frac{1}{T} \sum_{i=1}^T |\widehat{\mu} - \mu^*|$ of various estimators. Both axes in all of the plots are in a $\log$-scale. In particular, the slope of the curves indicates the power of n in the estimation error, and vertical shifts correspond to constant prefactors.

7.9.1 Univariate Data

We first present simulation results when $d = 1$. We use $r = 1$ for the simulations involving r-modal interval estimators, since $R_1^* = \Omega\left(\frac{\log n}{n}\right)$ in each of the recurring

examples, although the constant prefactors do not exactly align with our theory.

In the case of Example 7.3.6 (i.i.d. observations), we generate $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. As seen in Figure 7.7(a), the mean and median estimator perform optimally in this setting, giving an error rate of $Θn^{-0.5}$). In contrast, the shorth estimator (with $k = 5\log n$) has a flat trend line indicative of constant error, as suggested by Remark 7.3.4 and the phase transition arguments in Section 7.3.1.2. On the other hand, the error of the hybrid estimator decays at a rate more comparable to the mean and median. As discussed in Remark 7.4.7, the hybrid estimator is indeed optimal up to log factors. We see that the performance of the modal interval estimator is better than the shorth but worse than the hybrid estimator, and exhibits the cube-root asymptotic decay encountered in classical statistics [KP90]. Furthermore, the estimation error of the hybrid estimator behaves more like the error of the median estimator as n increases. Note that although our bounds for the shorth and modal interval estimators are tighter for smaller values of k and r, choosing larger values results in better performance when the data are homogeneous, which is not a valid assumption in our general use case.

For Example 7.3.7 (quadratic variance), we generate $x_i \sim \mathcal{N}(0, i^2)$. In Figure 7.10(a), we see that the both the median and mean have similar slopes: Proposition 7.3.13 predicts that the median would have $\tilde{O}(\sqrt{n})$ error, compared to the $\Theta\left(\sqrt{\frac{1}{n^2}\sum_{i=1}^{n} i^2}\right) = \Theta\left(\sqrt{n}\right)$ error of the mean; indeed, the curves are roughly parallel. However, the error rate of the modal interval, shorth, and hybrid estimators is significantly smaller. As stated in Propositions 7.3.10 and 7.3.14, the error of these estimators is upper-bounded by $Θn^{\epsilon}$), for $\epsilon > 0$.

For Example 7.3.8 (α-mixture distributions), we generate $\lceil 10\log n \rceil$ samples from a $\mathcal{N}(0, 4\times 10^{-4})$ distribution and the remaining samples from a $\mathcal{N}(0, n^{\alpha})$ distribution, with $\alpha = 0.9$ and 1.3. The plots in Figure 7.15 add additional curves to the phase transition plots in Figure 7.3. As suggested by Propositions 7.3.10 and 7.3.14, the modal, shorth,

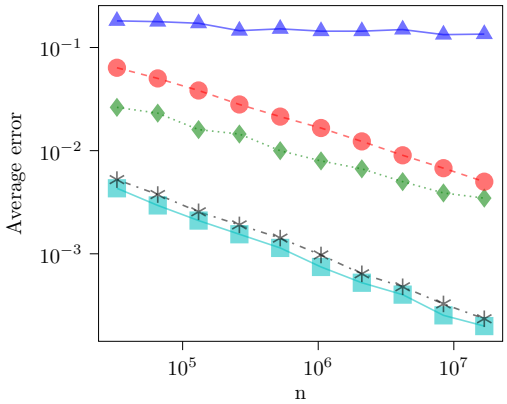

Figure 7.5: *

(a) $d = 1$

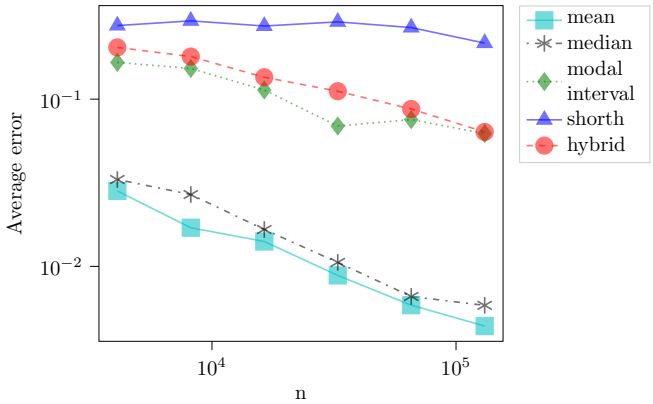

Figure 7.6: *

(b) $d = 3$

Figure 7.7: Plot comparing average error of various estimators on Example 7.3.6. Both the mean and median exhibit the familiar $On^{-0.5}$) error rate. The modal interval has errors of order $n^{-1/3}$. As suggested by our theoretical bounds, the $(\log n)$-shorth has constant error. The hybrid estimator improves the rate of the shorth estimator, with a similar error decay as the median estimator as n increases.

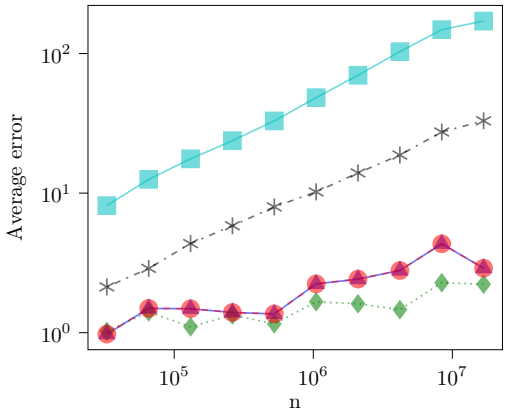

Figure 7.8: *

(a) $d = 1$

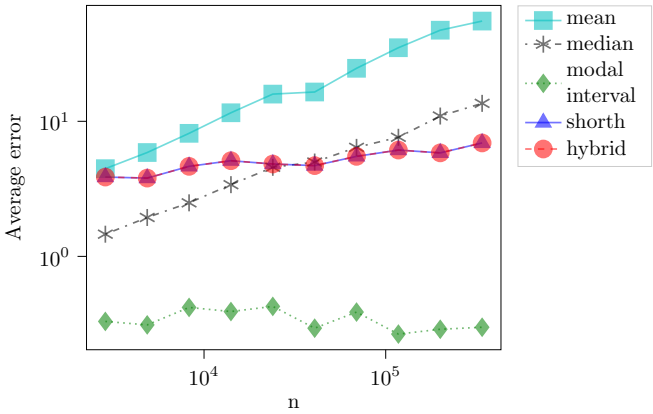

Figure 7.9: *

(b) $d = 3$

Figure 7.10: Plot comparing average error of various estimators on Example 7.3.7. As mentioned in Table 7.1, both the mean and median have $\sqrt{n}$ error rate. The error rates of the modal interval, shorth (with $k = 5d \log n$), and hybrid estimators are superior to the median in the univariate case, and the hybrid estimator is clearly superior when $d = 3$.

and hybrid estimators have constant error for $\alpha > 1$, whereas the error increases with n when $\alpha < 1$. Furthermore, the hybrid estimator performs better than the shorth

estimator when $\alpha < 1$, with an error rate of $On^{\alpha-0.5})$ rather than $On^{\alpha})$, while the modal interval estimator seems to perform comparably to the hybrid. Finally, note that the behavior of the hybrid estimator is similar to the behavior of the median estimator when $\alpha < 1$ and to the modal interval/shorth estimator when $\alpha > 1$, showing that it indeed enjoys the better of the two rates in different regimes.

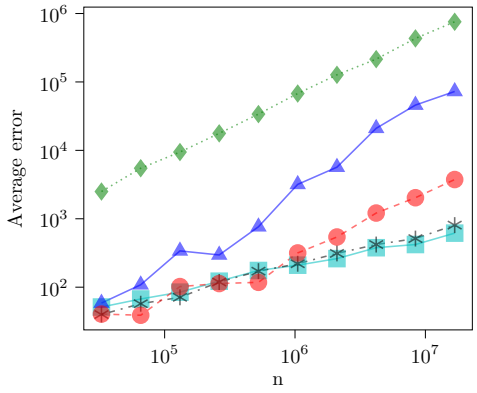

Figure 7.11: *

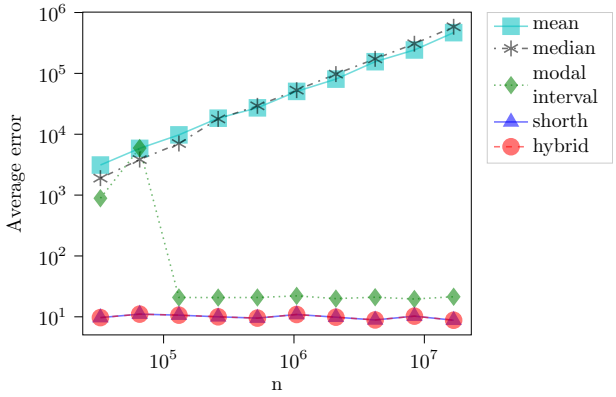

Figure 7.12: *

(a) $d = 1, \alpha = 0.9$

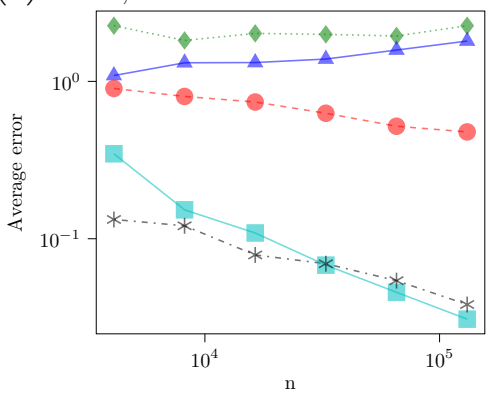

Figure 7.13: *

(b) $d = 1, \alpha = 1.3$

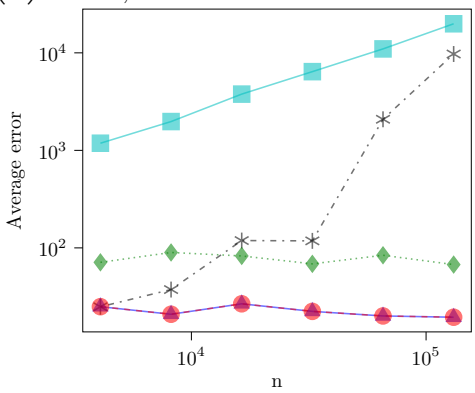

Figure 7.14: *

(c) $d = 3, \alpha = \frac{1}{2d}$

(d) $d = 3, \alpha = 1.3$

Figure 7.15: Plots comparing average error of various estimators on Example 7.3.8 for different values of α. As suggested by Proposition 7.3.13, the median and mean have superior performance to the modal interval and shorth estimators for $\alpha < 1$. Moreover, the hybrid estimator exhibits similar behavior to the median when $\alpha < 1$ and to the shorth when $\alpha > 1$.

7.9.2 Multivariate

We now present simulation results for multivariate data, using $d = 3$. The data for all three recurring examples are generated with the same parameters as in the univariate case, except with isotropic distributions. We run the computationally efficient versions of the shorth and modal interval estimators described in Section 7.6, with $k = 5d \log n$ and $r = \sqrt{d}$.

The trends for i.i.d. data, shown in Figure 7.7(b), are analogous to the univariate case. Similarly, the plots in Figure 7.10(b) for the quadratic variance example resemble the plots in Figure 7.10(a), with the hybrid, shorth, and modal interval estimators performing noticeably better than the mean or median. Note that for these experiments, the modal interval estimator appears to behave better than either the shorth or hybrid estimators by a constant factor. For the multivariate version of the α-mixture distribution, we run simulations with $\alpha = \frac{1}{2d} < 1$ and $\alpha = 1.3$, where we have chosen the first value of α so that the upper bound in Theorem 7.5.7 gives $O\left(n^{\alpha - \frac{1}{2}}\right) = On^{\frac{1}{2d} - \frac{1}{2}})$ error for the hybrid estimator, whereas the derived bounds for the modal interval and shorth are $On^{\alpha}) = On^{\frac{1}{2d}})$ (cf. Remark 7.4.7). Indeed, we see in Figure 7.15(c) that the estimation error of the hybrid estimator decreases with n, like the mean and median estimators, whereas the shorth estimator has an increasing trend line. The curve for the modal interval estimator appears to be roughly constant (or possibly slightly increasing). The curves in Figure 7.15(d) are very similar to the curves in Figure 7.15(b), suggesting the existence of a phase transition for $\alpha \in \left(\frac{1}{2d}, 1\right]$ in the multivariate case, as well.

7.10 Conclusion

We have studied the problem of mean estimation of a heterogeneous mixture when the fraction of clean points tends to 0. We have shown that the modal interval and shorth

estimator, which perform suboptimally in i.i.d. settings, are superior to the sample mean in such settings. We have also shown that these estimators and the k-median have complementary strengths that may be combined into a single hybrid estimator, which adapts to the given problem and is nearly optimal in certain settings. An important question for further study is whether the proposed hybrid estimator is always near-optimal, or optimal, for more general collections of variances.

Our discussion of linear regression estimators has been fairly brief. Some issues that we have not addressed include derivations for non-Gaussian error distributions and regression estimators in the case of a fixed design matrix. We leave these questions, and a derivation of optimal error rates in the linear regression setting, for future work.

Funding

This work was supported by the National Science Foundation [DMS-1749857 to AP & PL, CCF-1841190 to VJ, CCF-1740707 to AP].

Acknowledgments

The authors thank the reviewers for their detailed feedback, which helped improve the manuscript. PL thanks Gabor Lugosi for introducing her to the entangled mean estimation problem at the 2017 probability and combinatorics workshop in Barbados.

Part II

Constraints on Computational Resources:

Communication, Memory, and Privacy

8 OVERVIEW OF RESULTS: CONSTRAINTS ON COMPUTATIONAL RESOURCES

और भी दुख हैं ज़माने में मोहब्बत के सिवा
राहतें और भी हैं वस्ल की राहत के सिवा

— फ़ैज़ अहमद फ़ैज़

In the previous part, we presented polynomial-time algorithms that could handle sampling constraints resulting in low-quality training data. In this part, we shift our attention to the constraints on computational resources.

While runtime has historically been the primary computational resource of concern, other factors such as communication and memory have become significant bottlenecks in many applications. Our focus in this part is on designing statistical inference algorithms that meet these computational constraints. As in the previous part, our goal will be to understand fundamental inference tasks in presence of these computational constraints. For each inference task, we will describe the problem statement technically, give a brief history of the problem, and conclude with our contributions.

We divide this chapter into two sections: Section 8.1 focuses on memory constraints and Section 8.2 focuses on communication and privacy constraints.

8.1 Memory Constraints

The first constraint that we will consider is that of memory. In many applications, data is generated at such a large volume and velocity that storing the dataset is impossible. In such applications, the algorithms are required to process the dataset one example at a time while using limited memory (in addition to having small runtime). Algorithms

satisfying these desiderata are called streaming algorithms and have a long history in theoretical computer science and statistics.

Given the importance of robust algorithms outlined earlier, it is an important question whether there are efficient streaming algorithms for high-dimensional robust statistics. We will focus on parameter estimation problems (mean, covariance, linear regression). We want robust algorithms that use memory nearly-linear in the dimensionality of the parameter of interest.

Inference Task 6 (Streaming Algorithms for High-Dimensional Robust Statistics)**.** *Let $\mathcal{D}$ be a set of distributions over $\mathbb{R}^d$. Let $D \in \mathcal{D}$ be unknown with the (unknown) parameter $\theta_D \in \mathbb{R}^{d'}$ (for example, mean, covariance). Let P be an unknown arbitrary distribution satisfying that $d_{\mathrm{TV}}(P, D) \leq \epsilon$. Given a set of i.i.d. samples from P in the streaming model, compute an estimate $\widehat{\theta} \in \mathbb{R}^{d'}$ using small memory (and runtime) such that with high probability, $\|\widehat{\theta} - \theta_D\|$ is small for an an appropriate norm $\|\cdot\|$.*

Existing robust algorithms, mentioned in the previous section, required memory quadratic in the dimensionality of the parameter (even though the runtime was polynomial). For example, robust algorithms for mean estimation required memory $\Omega(d^2)$ despite being a parameter of size d; Similarly, robust covariance estimation algorithm required memory $\Omega(d^4)$ despite being a parameter of size d^2. This excessive use of memory has also been highlighted as one of the key challenges in the existing experiments of these algorithms [DKKLMS17]. This leads to the following question:

Question 12. *Is there an efficient streaming algorithm for high-dimensional robust statistics that uses nearly-linear memory?*

This is a challenging question since all the existing computationally-efficient robust algorithms rely on higher moments of the training data to curb the effect of outliers (for example, using the covariance for robust mean estimation). As these higher moments

have a much larger memory footprint, a naive implementation of this recipe requires a much larger memory. Although more efficient ways exist to extract information about the higher moments (for example, power iteration), existing algorithms still require super-linear memory (in the dimension) even after implementing these techniques. In fact, no algorithm with sub-quadratic memory was known (that used sub-exponential runtime/samples).

Our Contributions In Chapter 9, we answer Question 12 in the affirmative by developing the first streaming algorithms for high-dimensional robust statistics with nearly-linear memory. Our main result is a streaming algorithm for robust mean estimation with nearly-linearly memory, which we combine with existing approaches in the literature to develop streaming algorithms for a host of other tasks. Our critical technical insight is to transform a large family of nearly-linear time filter-based algorithms so that they additionally satisfy low-memory requirements.

8.2 Communication and Privacy Constraints

In recent years, machine learning has experienced a paradigm shift towards more distributed and edge-based implementations. This shift is primarily driven by the need to process the ever-growing volume of data generated by a vast array of small devices (such as mobile phones and remote sensors). Thus, the training data is distributed across these devices while a central server communicates with them to learn the underlying model. As the number of devices and the volume of data increase, it becomes increasingly important to minimize the communication overhead while maintaining the accuracy and efficiency of the learning process. In addition, privacy concerns and data protection regulations often limit the sharing of raw data, necessitating the development of techniques to learn from locally stored data without violating privacy constraints.

Hypothesis testing is one of the most fundamental problems in statistics. Recall that hypothesis testing is defined as follows: given a list of disjoint sets of distributions $\mathcal{P}_1, \ldots, \mathcal{P}_M$ and access to samples from an (unknown) distribution $P \in \cup_{i=1}^{M} \mathcal{P}_i$, identify (with high probability) the set $\mathcal{P}_{i^*}$ such that $P \in \mathcal{P}_{i^*}$. Our focus will primarily be on the variant of the problem where each $\mathcal{P}_i$ is a singleton set, termed "simple" hypothesis testing in the literature. We will study hypothesis testing in the decentralized setup, known as decentralized detection in the literature[Tsi93]. Recall that in the decentralized/distributed setup, it is expensive (or prohibitive) to send the original observations to the central server, and the observations need to be modified before being sent to the central server. These constraints will be captured by $\mathcal{T}$ below:

Inference Task 7 (Decentralized Detection)**.** *Let $p_1, \ldots, p_M$ be M distributions supported on the domain $\mathcal{X}$. Let $\mathcal{T}$ be a set of stochastic maps on $\mathcal{X}$ representing the constraints (for example, communication and privacy). There are n users $U_1, \ldots, U_n$, where each user U_i chooses (independently of each other) a stochastic map $f_i \in \mathcal{T}$. Each user then observes X_i and transmits $f_i(X_i)$ to the central server U. The goal of the central server is to identify the underlying measure based on $Y_1, \ldots, Y_n$, i.e., generate $\widehat{\phi} := \phi(Y_1, \ldots, Y_n)$ such that*

$$\sum_{i=1}^{M} \mathop{\mathbb{P}}_{X_1,\ldots,X_n \sim p_i^{\otimes n}} (\widehat{\phi} \neq p_i) \leq 0.1\,. \tag{8.1}$$

The smallest n where such f_i's and ϕ exist is called the sample complexity of the problem.

The set of maps, $\mathcal{T}$, decides the amount of "information" that the server can get from the original samples $X_1, \ldots, X_n$. The choice of the functions f_i is crucial here, as it must be chosen to convey as much information about the original sample that might be useful for estimating the true distribution.

Let n^* be the original sample complexity sans constraints (when the server directly observes $X_1, \ldots, X_n$).

Our primary focus in this section will be on the binary case of $M = 2$, called "simple binary hypothesis testing". We defer the discussion of $M > 2$ to Chapter 10.

8.2.1 Communication Constraints

We begin by considering the communication constraints, i.e., when $\mathcal{T}$ is the set of stochastic maps from $\mathcal{X} \to [\ell]$ for some $\ell \in \mathbb{N}$ much smaller than $|\mathcal{X}|$. On the first reading, we suggest the reader thinks of $\ell = 2$, which corresponds to the binary quantization of observations. Thus, we arrive at the following problem

Inference Task 8 (Simple Binary Hypothesis Testing: Communication Constraints)**.** *In Inference Task 7, consider $M = 2$ and let $\mathcal{T}$ be the set of stochastic maps from $\mathcal{X} \to [\ell]$ for some $\ell \in \mathbb{N}$. Let $n^*_{comm}(\ell)$ be the corresponding sample complexity.*

We refer to the blow-up in the sample complexity due to communication constraints as the *statistical cost of communication*, i.e., $n^*_{\text{comm}}(\ell)/n^*$. Prior literature on the characterization of $n^*_{\text{comm}}(\ell)$ was scarce. Existing results (by applying Scheffe's Test) implied that $n^*_{\text{comm}}(\ell)/n^* \lesssim n^*$, thus, requiring quadratically many more samples (In fact, this upper bound is tight for Scheffe's test in certain cases). It was unclear if this blow-up was inherent, leading to the following question:

Question 13. *What is the statistical cost of communication for simple binary hypothesis testing?*

Moving ahead from the statistical cost, we now focus on the algorithm's runtime. Recall that an algorithm needs to take as input p_1 and p_2 (since we are in the binary case of $M = 2$) and $\mathcal{T}$, and the algorithm must return the map f_is for each user. We refer to the runtime of finding a good f_is as the *computational cost of communication*. Prior algorithms for computing the optimal f_i's were exponential in the quantization size, $|\mathcal{X}|^\ell$.

Question 14. *What is the computational cost of communication for simple binary hypothesis testing? Are their polynomial time algorithms for near-optimal performance?*

We briefly discuss the setting of $M > 2$ and defer the remaining discussion to Chapter 10. Sans any constraints, it is known that the dependence on M in the sample complexity is logarithmic, $\log(M)$ [DL01]. However, the algorithm mentioned above is not communication-efficient. One approach to meet the communication constraints is to reduce the M-ary hypothesis testing into M-many binary hypothesis testing problems. However, the sample complexity then blows up and depends linearly on M instead of logarithmic in M. This leads to the question of whether this is inherent:

Question 15. *What is the statistical cost of communication for simple M-ary hypothesis testing? Are their algorithms with sample complexity $o(M)$?*

Our Contributions We summarize our main contributions here:

- (Question 13) We characterize the minimax-optimal statistical cost of privacy, showing that for any ℓ, $n^*_{\text{comm}}(\ell) \asymp n^* \left(1 + \frac{\log(n^*)}{\ell}\right)$. Thus, the existing upper bound was loose; the statistical cost of communication for simple binary hypothesis testing is at most logarithmic (and is at least a logarithmic factor in some cases).
- (Question 14) We then show that the aforementioned minimax-optimal sample complexity can be attained by polynomial-time algorithms, running in time $\text{poly}(\ell, k)$.
- (Question 15) For simple M-ary hypothesis testing, we show that the effect of communication on the statistical cost is drastic: there is an exponential increase in the sample complexity.

8.2.2 Privacy Constraints

Differential privacy has emerged as the standard way of ensuring privacy in data science applications. As our primary focus is on the decentralized setup, we will consider the local model of privacy, known as local differential privacy (LDP). In the local model, each user perturbs their own data before sending it to a central server for analysis. By setting $\mathcal{T}$ in Inference Task 7 to be the set of all ϵ-locally private channels and $M = 2$, we obtain the problem of simple binary hypothesis testing under local privacy constraints.

Inference Task 9 (Simple Binary Hypothesis Testing: Local Differential Privacy Constraints (LDP))**.** *In Inference Task 7, let $\mathcal{T}$ be the set of all stochastic maps from $\mathcal{X}$ that satisfy ϵ-LDP. Let $n^*_{priv}(\epsilon)$ be the corresponding sample complexity.*

When $\epsilon = \infty$, we may set Y_i equal to X_i with probability 1, and we recover the vanilla version of the problem with no privacy constraints.

A practically important regime of LDP is when ϵ is moderately large (say $\epsilon > 1$). Initially, the moderate ϵ regime was ignored mainly because of the lax privacy protection (recall that the smaller the ϵ, the larger the privacy protection). This regime has become practically relevant in the last five years because of the development of privacy amplification methods [CSUZZ19; BEMMRLRKTS17; FMT21]. Despite practical relevance, several fundamental questions, both statistical and computational, remain open in this regime. Following the previous subsection, we are interested in understanding the statistical cost (blow-up in the sample complexity, $n^*_{\text{priv}}(\epsilon)$ vs. n^*) and the computational cost (runtime to find a good ϵ-LDP stochastic map) of privacy for Inference Task 9.

Statistical Cost of LDP For the sample complexity, existing results have focused on the high-privacy regime of $\epsilon \in (0, 1)$ and have shown that the sample complexity is $\Theta\left(\frac{1}{\epsilon^2 d^2_{\text{TV}}}\right)$, where d_{TV} is the total variation distance between the two distributions p_1 and p_2. When $\epsilon = \infty$, we obtain the vanilla version of the problem, and it is known that

$n^* = n^*(\epsilon) = \Theta\left(\frac{1}{d_{\mathrm{h}}^2}\right)$, where d_{h}^2 is the Hellinger divergence between the two distributions p_1 and p_2.

Thus, when ϵ is a constant, the sample complexity is $\Theta\left(\frac{1}{d_{\mathrm{TV}}^2}\right)$, and when $\epsilon = \infty$ (no privacy), the sample complexity is $\Theta\left(\frac{1}{d_{\mathrm{h}}^2}\right)$. Although these two divergences satisfy $d_{\mathrm{TV}}^2 \lesssim d_{\mathrm{h}}^2 \lesssim d_{\mathrm{TV}}$, the bounds are tight in the worst case; i.e., the two sample complexities can be quadratically far apart. Existing results, therefore, do not inform sample complexity when $1 \ll \epsilon < \infty$. This is not an artifact of analysis: the optimal tests in the low and high privacy regimes are fundamentally different.

Question 16. *What is the statistical cost of local differential privacy for simple binary hypothesis testing?*

Computational Cost of LDP As with statistical rates, prior literature on finding optimal channels for $\epsilon \gg 1$ is scarce. Either the existing algorithms take time exponential in the domain size [KOV16], or their sample complexity is suboptimal by polynomial factors (depending on $\frac{1}{d_{\mathrm{TV}}^2}$, as opposed to $\frac{1}{d_{\mathrm{h}}^2}$). This raises the following natural question:

Question 17. *Is there a polynomial-time algorithm that finds the private stochastic maps f's whose sample complexity is (near)-optimal?*

The problem is computationally challenging because answering this question requires optimizing a convex function (Hellinger divergence between the distributions after transformed by ϵ-LDP map) over a convex set (the set of all ϵ-LDP maps). Recall that optimizing a convex function over a convex set, in general, could be computationally prohibitive.

Our Contributions We describe our contributions below:

- (Question 16) We show that the sample complexity for $\epsilon \gg 1$ is rather involved and no longer characterized by their total variation and Hellinger divergence.

- (Question 17) We give the first polynomial-time algorithms whose sample complexity is near-optimal (for all values of ϵ and all choices of distributions). Moreover, the proposed algorithm is also communication-efficient and uses only a single bit.

9 STREAMING ALGORITHMS FOR HIGH-DIMENSIONAL ROBUST STATISTICS

जो मिल गया उसी को मुक़द्दर समझ लिया
जो खो गया मैं उस को भुलाता चला गया
बर्बादियों का सोग मनाना फ़ुज़ूल था
बर्बादियों का जश्न मनाता चला गया
ग़म और ख़ुशी में फ़र्क़ न महसूस हो जहाँ
मैं दिल को उस मक़ाम पे लाता चला गया

— साहिर लुधियानवी

We study high-dimensional robust statistics tasks in the streaming model. A recent line of work obtained computationally efficient algorithms for a range of high-dimensional robust estimation tasks. Unfortunately, all previous algorithms require storing the entire dataset, incurring memory at least quadratic in the dimension. In this work, we develop the first efficient streaming algorithms for high-dimensional robust statistics with near-optimal memory requirements (up to logarithmic factors). Our main result is for the task of high-dimensional robust mean estimation in (a strengthening of) Huber's contamination model. We give an efficient single-pass streaming algorithm for this task with near-optimal error guarantees and space complexity nearly-linear in the dimension. As a corollary, we obtain streaming algorithms with near-optimal space complexity for several more complex tasks, including robust covariance estimation, robust regression, and more generally robust stochastic optimization.

9.1 Introduction

This work studies high-dimensional learning in the presence of a constant fraction of arbitrary outliers. Outlier-robust learning in high dimensions is motivated by pressing machine learning (ML) applications, including ML security [BNJT10; BNL12; SKL17;

TLM18; DKKLSS19] and exploratory analysis of datasets with natural outliers [RPWCKZF02; PLJD10; LATSCR+08]. This field has its roots in robust statistics, a branch of statistics initiated in the 60s with the pioneering works of Tukey and Huber [Tuk60; Hub64]. Early work developed minimax optimal estimators for various robust estimation tasks, albeit with runtimes exponential in the dimension. A recent line of work in computer science, starting with [DKKLMS16; LRV16], developed polynomial time robust estimators for a range of high-dimensional statistical tasks. Algorithmic high-dimensional robust statistics is by now a relatively mature field, see, e.g., [DK19; DKKLMS21] for surveys.

This recent progress notwithstanding, even for the basic task of mean estimation, previous robust estimators require the entire dataset in main memory. This space requirement can be a major bottleneck in large-scale applications, where an algorithm has access to a very large stream of data. Indeed, practical machine learning methods are typically simple iterative algorithms that make a single pass over the data and require a small amount of storage — with stochastic gradient descent being the prototypical example [Bot10; BCN18]. Concretely, in prior applications of robust statistics in data analysis [DKKLMS17] and data poisoning defenses [DKKLSS19], the storage requirements of the underlying algorithms were observed to significantly hinder scalability. This discussion motivates the following natural question:

Can we develop efficient robust estimators in the streaming model
with (near-) optimal space complexity?

We emphasize that this broad question is meaningful and interesting even ignoring computational considerations. While any method requires space complexity $\Omega(d)$, where d is the dimension of the problem (to store a single sample), it is not obvious that a matching upper bound exists. We note that it is relatively simple to design $O(d)$-memory streaming algorithms with sample complexity exponential in d. But it is by no means

clear whether there exists an estimator with near-linear space requirements and $\mathrm{poly}(d)$ sample complexity (independent of its runtime).

9.1.1 Our Results

In this work, we initiate a systematic investigation of high-dimensional robust statistics in the streaming model. We start by focusing on the most basic task — that of robust mean estimation. Our main result is the first space-efficient streaming algorithm for robust mean estimation under natural distributional assumptions. Our computationally efficient algorithm makes a single pass over the data, uses near-optimal space, and matches the error guarantees of previous polynomial-time algorithms for the problem.

Given this result, we leverage the fact that several robust statistics tasks can be reduced to robust mean estimation to obtain near-optimal space, single-pass streaming algorithms for more complex statistical tasks.

To formally state our contributions, we require some basic definitions. We start with the standard streaming model.

Definition 9.1.1 (Single-Pass Streaming Model)**.** *Let S be a fixed set. In the* one-pass streaming model*, the elements of S are revealed one at a time to the algorithm, and the algorithm is allowed a single pass over these points.*

Our robust estimators work in the following contamination model, where the adversary can corrupt the true distribution in total variation distance (for distributions P and Q, we use $d_{\mathrm{TV}}(P, Q)$ to denote their total variation distance).

Definition 9.1.2 (TV-contamination)**.** *Given a parameter $\epsilon < 1/2$ and a distribution class $\mathcal{D}$, the adversary specifies a distribution D' such that there exists $D \in \mathcal{D}$ with $d_{\mathrm{TV}}(D, D') \leq \epsilon$. Then the algorithm draws i.i.d. samples from D'. We say that the distribution D' is an ϵ-corrupted version of the distribution D in total variation distance.*

The distribution D' in Definition 9.1.2 can be adversarially selected (and can even depend on our learning algorithm). Since Huber's contamination model [Hub64] only allows additive errors, TV-contamination is a stronger model.

Streaming Algorithm for Robust Mean Estimation

The main result of this paper is the following (see Theorem 9.4.2 for a more general statement):

Theorem 9.1.3 (Streaming Robust Mean Estimation)**.** *Let $\mathcal{D}$ be a distribution family on $\mathbb{R}^d$ and $0 < \epsilon < \epsilon_0$ for a sufficiently small constant $\epsilon_0 > 0$. Let P be an ϵ-corrupted version of D in total variation distance for some $D \in \mathcal{D}$ with unknown mean μ_D. There is a single-pass streaming algorithm that, given ϵ and $\mathcal{D}$, reads a stream of n i.i.d. samples from P, runs in sample near-linear time, uses memory $d\,\mathrm{polylog}(d/\epsilon)$, and outputs an estimate $\widehat{\mu}$ that, with probability at least $9/10$, satisfies the following:*

1. *If $\mathcal{D}$ is the family of distributions with identity-bounded covariance, then $n = \tilde{O}\left(d^2/\epsilon\right)$ and $\|\widehat{\mu} - \mu_D\|_2 = O(\sqrt{\epsilon})$.*
2. *If $\mathcal{D}$ is the family of identity-covariance subgaussian distributions, then $n = \tilde{O}\left(d^2/\epsilon^2\right)$ and $\|\widehat{\mu} - \mu_D\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.*

We note that the above error guarantees are information-theoretically optimal, even in absence of resource constraints. While prior work had obtained efficient robust mean estimators matching these error guarantees [DKKLMS16; DKKLMS17; SCV18], all previous algorithms with dimension-independent error incurred space complexity $\Omega(d^2)$.

Beyond Robust Mean Estimation

Using the algorithm of Theorem 9.1.3 as a black-box, we obtain the first efficient single-pass streaming algorithms with near-optimal space complexity for a range of more complex statistical tasks. These contributions are presented in detail in Section 9.5. Here we highlight some of these results.

Our first application is a streaming algorithm for robust covariance estimation.

Theorem 9.1.4 (Robust Gaussian Covariance Estimation)**.** *Let Q be a distribution on $\mathbb{R}^d$ with $d_{\mathrm{TV}}(Q, \mathcal{N}(0, \mathbf{\Sigma})) \leq \epsilon$ and assume $\frac{1}{\kappa}\mathbf{I}_d \preceq \mathbf{\Sigma} \preceq \mathbf{I}_d$. There is a single-pass streaming algorithm that uses $n = (d^4/\epsilon^2)\mathrm{polylog}(d, \kappa, 1/\epsilon)$ samples from Q, runs in time $nd^2\mathrm{polylog}\,(d, \kappa, 1/\epsilon)$, uses memory $d^2\mathrm{polylog}\,(d, \kappa, 1/\epsilon)$, and outputs a matrix $\widehat{\mathbf{\Sigma}}$ such that $\|\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_d\|_F = O(\epsilon \log(1/\epsilon))$ with probability at least $9/10$.*

See Theorem 9.5.3 for a more detailed statement.

Our second application is for the general problem of robust stochastic optimization. Here we state two concrete results for robust linear and logistic regression (see Theorem 9.5.9 and Theorem 9.5.12 for more detailed statements). Both of these statements are special cases of a streaming algorithm for robust stochastic convex optimization (see Corollary 9.5.6).

Theorem 9.1.5 (Streaming Robust Linear Regression)**.** *Let D be the distribution of (X, Y) defined by $Y = X^\top\theta^* + Z$, where $X \sim \mathcal{N}(0, \mathbf{I}_d)$, $Z \sim \mathcal{N}(0, 1)$ independent of X, and $\|\theta^*\|_2 \leq r$. Let P be an ϵ-corruption of D in total variation distance. There is a single-pass streaming algorithm that uses $n = (d^2/\epsilon)\,\mathrm{polylog}\,(d(1 + r)/\epsilon)$ samples from P, runs in time $nd\,\mathrm{polylog}(d(1 + r)/\epsilon)$, uses memory $d\,\mathrm{polylog}(dr/\epsilon)$, and outputs an estimate $\widehat{\theta} \in \mathbb{R}^d$ such $\|\widehat{\theta} - \theta^*\|_2 = O(\sqrt{\epsilon})$ with high probability.*

Theorem 9.1.6 (Streaming Robust Logistic Regression)**.** *Consider the following model: Let $(X, Y) \sim D$, where $X \sim \mathcal{N}(0, \mathbf{I}_d)$, $Y \mid X \sim$ Bern(p), for $p = 1/(1 + e^{-X^\top\theta^*})$, and $\|\theta^*\|_2 =$*

$O(1)$. *Let P be an ϵ-corruption of D in total variation distance. There is a single-pass streaming algorithm that uses $n = (d^2/\epsilon)\,\mathrm{polylog}\,(d/\epsilon)$ samples from P, runs in time $nd\,\mathrm{polylog}(d/\epsilon)$, uses memory $d\,\mathrm{polylog}(d/\epsilon)$, and outputs an estimate $\widehat{\theta} \in \mathbb{R}^d$ such $\|\widehat{\theta} - \theta^*\|_2 = O(\sqrt{\epsilon})$ with high probability.*

Finally, in Section 9.5, we include an additional application to distributed non-convex optimization in the streaming setting.

Remark 9.1.7 (Bit complexity)**.** *For simplicity of presentation, in the main body of the paper, we consider the model of computation where the algorithms can store and manipulate real numbers exactly. We show in Appendix F.6 that our algorithms can tolerate errors due to finite precision. In particular, all our algorithms (including Algorithm 8) can be implemented in the word RAM model with $d\,\mathrm{polylog}(d/\epsilon)$ bits.*

9.1.2 Overview of Techniques

In this section, we provide a brief overview of our approach to establish Theorem 9.1.3. We start by recalling how robust mean estimation algorithms typically work without space constraints. A standard tool in the literature is the filtering technique of [DKKLMS16; DKKLMS17; DK19]. The idea of the filtering method is the following: Given a set S of corrupted samples, by analyzing spectral properties of the covariance of S, we can either certify that the sample mean of S is close to the true mean of the distribution, or can construct a filter. The filter is a method for selecting some elements of S to remove, with the guarantee that it will remove more outliers than inliers. If we can efficiently construct a filter, our algorithm can then remove the selected samples from S, obtaining a cleaner dataset and repeat the process. Eventually, this procedure must terminate, giving an accurate estimate of the true mean.

We proceed to explain how to implement the filtering method in a streaming model. We start with the easier case where the dataset is stored in read-only-memory, or more

generally in a *multi-pass* streaming setting. At each round of the algorithm, one has a subset S' of the original dataset S that needs to be maintained (in particular, the set of samples that has survived the filters applied thus far). To do this naïvely would require $n = |S|$ many bits of memory, which is too much for us. A more inventive strategy would be the following: instead of storing these subsets S' explicitly, store them implicitly by instead storing enough information to reconstruct the filters used to obtain S'. This seems like a productive idea, as most filters are relatively simple. For example, a commonly used filter is to remove all points $x \in S$ for which $v^\top x > t$, for some vector v and scalar t. One could store enough information to apply this filter by merely storing (v, t), which would take $O(d)$ bits of information. Unfortunately, most filtering algorithms may require $\Omega(d)$ many iterations before attaining their final answer. Consequently, the sets S' one needs to store are not just the result of applying a single filter, but instead the result of iteratively applying $\Omega(d)$ of them. In order to store all of these extra filters, one would need $\Omega(d^2)$ bits. (For the sake of this intuitive description, we focused on "hard-thresholding" filters. Our algorithm will actually use a soft-thresholding filter, assigning weights to each point.)

To circumvent this first obstacle, one requires as a starting point a filtering algorithm that is guaranteed to terminate after a small (namely, at most poly-logarithmic) number of iterations. Recent work [DHL19; DKKLT22] has obtained such algorithms. Here we generalize and simplify the filtering method of [DKKLT22]. This allows us to obtain an algorithm with space complexity $d \operatorname{polylog}(d/\epsilon)$ that works in the *multi-pass streaming model*, where $\operatorname{polylog}(d/\epsilon)$ passes over the same dataset are allowed.

To obtain a *single-pass* streaming algorithm, new ideas are required. In the single-pass setting, we cannot implicitly store a subset of the full dataset S; once we access some points from S, we will never be able to see them again. To deal with this issue, we will need to slightly alter our way of thinking about the algorithm. Instead of being given a

set S of samples, an ϵ-fraction of which have been corrupted, we instead adopt the view of having sample access to a distribution P, which is ϵ-close in total variation distance to the inlier distribution G. Given this point of view, instead of a filter defining a procedure for removing samples from S and outputting a subset S', we think of it as a rejection sampling procedure that replaces P with a cleaner distribution P'.

This shift in perspective comes with new technical challenges. In particular, when constructing the next round of filters, we will need to compute quantities pertaining to the current distribution P of the data points. In the setting of the multiple-pass model, this imposed no problem; these quantities could be calculated exactly. This is no longer possible when we merely have sample access to P. The best one can hope for is to approximate these quantities to sufficient precision for the rest of our analysis to carry over. However, the natural estimators for some required quantities (e.g., powers of the covariance matrix) would need to access the data multiple times. Circumventing this issue requires non-trivial technical work. Roughly speaking, instead of iterating over the same dataset to approximate the desired quantities, we show that it suffices to iterate over statistically identical datasets.

9.1.3 Prior and Related Work

Since the dissemination of [DKKLMS16; LRV16], there has been an explosion of research in algorithmic aspects of robust statistics. We now have efficient robust estimators for a range of more complex problems, including covariance estimation [DKKLMS16; CDGW19], sparse estimation tasks [BDLS17; DKKPS19; CDKGGS22], learning graphical models [CDKS18; DKSS21], linear regression [KKM18; DKS19; PJL20b], stochastic optimization [PSBR20; DKKLSS19], and robust clustering/learning various mixture models [HL18; KSS18; DKS18; DKKLT22; DKKLT22; BDHKKK20; LM21a; BDJKKV22]. The reader is referred to [DK19] for a detailed overview. We reiterate that all previously

developed algorithms work in the batch setting, i.e., require the entire dataset in memory.

For the problem of robust mean estimation, [DHL19; DKKLT22] gave filtering-based algorithms with a poly-logarithmic number of iterations. The former algorithm relies on the matrix multiplicative weights framework, while the latter is based on first principles. Our starting point in Section 9.3 can be viewed as a generalization and further simplification of the ideas in [DKKLT22]. Specifically, our algorithm works under the stability condition (Definition 9.2.8), which broadly generalizes the bounded covariance assumption used in [DKKLT22].

In the context of robust supervised learning (including, e.g., our robust linear regression application), low-space streaming algorithms are known in weaker contamination models that only allow *label* corruptions, see, e.g., [PF20; SWS20; DKTZ20]. We emphasize that the contamination model of Definition 9.1.2 is significantly more challenging, and no low-space streaming algorithms were previously known in this model.

Finally, we note that recent work [TPBR21] studies streaming algorithms for heavy-tailed stochastic optimization. While the goal of developing low-space streaming algorithms is qualitatively similar to the goal of our work, the algorithmic results in [TPBR21] have no implications in the corrupted setting studied in this work.

9.1.4 Organization

The structure of this paper is as follows: In Section 9.2, we record the notation and technical background that will be used throughout the paper. In Section 9.3, we design a filter-based algorithm for robust mean estimation under the stability condition with a poly-logarithmic number of iterations. In Section 9.4, we build on the algorithm from Section 9.3, to obtain our single-pass streaming algorithm for robust mean estimation under the stability condition. Finally, in Section 9.5, we obtain our streaming algorithms for more complex robust estimation tasks. To facilitate the flow of the presentation,

some proofs of intermediate lemmas are deferred to the Appendix.

9.2 Preliminaries

9.2.1 Notation and Basic Facts

Basic Notation We use $\mathbb{Z}_+$ to denote the set of positive integers. For $n \in \mathbb{Z}_+$, we denote $[n] := \{1, \ldots, n\}$ and use $\mathcal{S}^{d-1}$ for the d-dimensional unit sphere. For a vector v, we let $\|v\|_2$ denote its ℓ_2-norm. We use boldface letters for matrices. We use $\mathbf{I}_d$ to denote the $d \times d$ identity matrix. For a matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_2$ to denote the Frobenius and spectral norms respectively. For $\mathbf{A} \in \mathbb{R}^{m \times n}$, we use $\mathbf{A}^\flat$ to denote the nm-dimensional vector obtained by concatenating the rows of $\mathbf{A}$. We say that a symmetric $d \times d$ matrix $\mathbf{A}$ is PSD (positive semidefinite), and write $\mathbf{A} \succeq 0$, if for all vectors $x \in \mathbb{R}^d$ we have that $x^\top \mathbf{A} x \geq 0$. We denote $\lambda_{\max}(\mathbf{A}) := \max_{u \in \mathcal{S}^{d-1}} x^\top \mathbf{A} x$. We write $\mathbf{A} \preceq \mathbf{B}$ when $\mathbf{B} - \mathbf{A}$ is PSD. For a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\mathrm{tr}(\mathbf{A})$ denotes the trace of the matrix A. We use $\otimes$ to denote the Kronecker product. For the sake of conciseness, we sometimes use $x = a \pm b$ as a shorthand for $a - b \leq x \leq a + b$. We use $a \lesssim b$, to denote that there exists an absolute universal constant $C > 0$ (independent of the variables or parameters on which a and b depend) such that $a \leq Cb$. Similarly, we use the notation $a \gtrsim b$ to denote that $b \lesssim a$. We use c, c', C, C' to denote absolute constants that may change from line to line, whereas we use constants $C_1, C_2, C_3, \ldots$ to denote fixed absolute constants that are important for our algorithms. We use $\tilde{O}(\cdot)$ to ignore poly-logarithmic factors in all variables appearing inside the parentheses. For the sake of simplicity, we sometimes omit rounding non-integer quantities to integer ones. For example, we treat logarithmic factors as integers when they appear in the sample complexity or number of iterations of an algorithm. We use $\mathrm{poly}(\cdot)$ to indicate a quantity that is polynomial in its arguments. Similarly, $\mathrm{polylog}(\cdot)$ denotes a quantity that is polynomial in the logarithm of its arguments.

Probability Notation For a random variable X, we use $\mathbb{E}[X]$ for its expectation. For a set S, we use $\mathcal{U}(S)$ to denote the uniform distribution on S. We use $\mathcal{N}(\mu, \mathbf{\Sigma})$ to denote the Gaussian distribution with mean μ and covariance matrix $\mathbf{\Sigma}$. For a distribution D on $\mathbb{R}^d$, we denote $\mu_D = \mathbb{E}_{X\sim D}[X]$ and $\mathbf{\Sigma}_D = \mathbb{E}_{X\sim D}[(X-\mu_D)(X-\mu_D)^\top]$. Moreover, given a weight function $w : \mathbb{R}^d \to [0,1]$, we define the re-weighted distribution D_w to be $D_w(x) := D(x)w(x)/\int_{\mathbb{R}^d} w(x)D(x)\mathrm{d}x$. We use $\mu_{w,D} = \mathbb{E}_{X\sim D_w}[X]$ for its mean and $\overline{\mathbf{\Sigma}}^{\mu}_{w,D} = \mathbb{E}_{X\sim D_w}[(X-\mu)(X-\mu)^\top]$ for the second moment that is centered with respect to μ (we will often drop μ from the notation when it is clear from the context). We use $\mathbb{I}\{x \in E\}$ to denote the indicator function of the set E.

Basic Facts We will use the following two basic facts.

Fact 9.2.1. *Let $x \in \mathbb{R}^d$ and $p \geq 1$. Then $\|x\|_{p+1} \leq \|x\|_p \leq \|x\|_{p+1} d^{\frac{1}{p(p+1)}}$.*

Fact 9.2.2. *The following results hold:*

1. *If $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are symmetric $d \times d$ matrices satisfying $\mathbf{A} \succeq 0$ and $\mathbf{B} \preceq \mathbf{C}$, we have that $\mathrm{tr}(\mathbf{AB}) \leq \mathrm{tr}(\mathbf{AC})$.*
2. *([JLT20]) Let $\mathbf{A}$ and $\mathbf{B}$ be PSD matrices satisfying $0 \preceq \mathbf{B} \preceq \mathbf{A}$. Then for any positive integer p, we have that $\mathrm{tr}(\mathbf{B}^p) \leq \mathrm{tr}(\mathbf{A}^{p-1}\mathbf{B})$.*

Proof. We provide the proof of the first claim below; The second claim is proved in [JLT20, Lemma 7]. Since $\mathbf{A}$ is PSD, we can consider its spectral decomposition $\mathbf{A} = \sum_{i=1}^d \lambda_i v_i v_i^\top$, where $\lambda_i \geq 0$. Using the linearity of trace operator, we have that

$$\mathrm{tr}(\mathbf{AB}) = \sum_{i=1}^d \lambda_i \mathrm{tr}(v_i v_i^\top \mathbf{B}) = \sum_{i=1}^d \lambda_i \mathrm{tr}(v_i^\top \mathbf{B} v_i) \leq \sum_{i=1}^d \lambda_i \mathrm{tr}(v_i^\top \mathbf{C} v_i) = \sum_{i=1}^d \lambda_i \mathrm{tr}(v_i v_i^\top \mathbf{C}) = \mathrm{tr}(\mathbf{AC}) \ ,$$

where the inequality uses that $\mathbf{B} \preceq \mathbf{C}$ and $\lambda_i \geq 0$. □

We will use the notion of total variation distance, defined below.

Definition 9.2.3. *Let P, Q be two probability distributions on $\mathbb{R}^d$. The total variation distance between P and Q, denoted by $d_{\mathrm{TV}}(P, Q)$, is defined as $d_{\mathrm{TV}}(P, Q) = \sup_{A \subseteq \mathbb{R}^d} |P(A) - Q(A)|$. For continuous distributions P, Q with densities p, q, we have that $d_{\mathrm{TV}}(P, Q) = \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)| \mathrm{d}x$.*

Whenever $d_{\mathrm{TV}}(P, Q) = \epsilon$, it is sometimes helpful to consider the decomposition below.

Fact 9.2.4. *Let a domain $\mathcal{X}$. For any $\epsilon \in [0, 1]$ and for any two distributions D_1, D_2 on $\mathcal{X}$ with $d_{\mathrm{TV}}(D_1, D_2) = \epsilon$, there exist distributions D, Q_1, Q_2 such that $D_1 = (1 - \epsilon)D + \epsilon Q_1$ and $D_2 = (1 - \epsilon)D + \epsilon Q_2$.*

This decomposition can be achieved by the following choice of Q_1, Q_2, and D:

$$Q_1(x) = \begin{cases} \frac{D_1(x) - D_2(x)}{\epsilon}\,, & \text{if } D_1(x) > D_2(x) \\ 0\,, & \text{otherwise} \end{cases}, \quad Q_2(x) = \begin{cases} \frac{D_2(x) - D_1(x)}{\epsilon}\,, & \text{if } D_2(x) > D_1(x) \\ 0\,, & \text{otherwise} \end{cases},$$

and $D(x) = \min\{D_1(x), D_2(x)\}/(1-\epsilon)$. In light of Fact 9.2.4, the adversary that performs corruption in total variation distance can be thought of as "both additive and subtractive" adversary.

Concentration Inequalities We will also require following standard results regarding concentration of random variables:

Fact 9.2.5 ([Ver12])**.** *Consider a distribution D on $\mathbb{R}^d$ that has zero mean and is supported in an ℓ_2-ball of radius R from the origin. Denote by $\mathbf{\Sigma}$ its covariance matrix and denote by $\mathbf{\Sigma}_N = (1/n) \sum_{i=1}^N X_i X_i^\top$ the empirical covariance matrix using N samples $X_i \sim D$. There is a constant C such that for any $0 < \epsilon' < 1$ and $0 < \tau < 1$, if $N > C\epsilon'^{-2} \|\mathbf{\Sigma}\|_2^{-1} R^2 \log(d/\tau)$, we have that $\|\mathbf{\Sigma} - \mathbf{\Sigma}_N\|_2 \leq \epsilon' \|\mathbf{\Sigma}\|_2$, with probability at least $1 - \tau$.*

Fact 9.2.6 (Quadratic Polynomials of a Gaussian)**.** *The Gaussian random variable satisfies the following properties:*

1. *For every $\beta > 0$, $\mathbb{P}_{X\sim\mathcal{N}(0,\mathbf{I}_d)}[|\|X\|^2 - d| > \beta] \leq 2e^{-c\beta^2/d}$, where $c > 0$ is a universal constant.*

2. *If $\mathbf{A}$ is a PSD matrix, then for any $\beta > 0$, it holds $\mathbb{P}_{z\sim\mathcal{N}(0,\mathbf{I})}[z^\top \mathbf{A} z \geq \beta \mathrm{tr}(\mathbf{A})] \geq 1 - \sqrt{e\beta}$.*

Fact 9.2.7 ([Ach03])**.** *Let $0 < \gamma < 1$ and $u_1, \ldots, u_N \in \mathbb{R}^d$. Let z_j for $j \in [L]$ drawn from the uniform distribution on $\{\pm 1\}^d$. There exists a constant $C > 0$ such that, if $L > C\log(N/\gamma)$, then, with probability at least $1 - \gamma$, we have that $0.8\|u_i\|^2 \leq \frac{1}{L}\sum_{j=1}^{L}(z_j^\top u_i)^2 \leq 1.2\|u_i\|^2$ for all $i \in [N]$.*

9.2.2 Stability Condition and Its Properties

Our results will hold for every distribution satisfying the following key property [DK19].

Definition 9.2.8 ((ϵ,δ)-stable distribution)**.** *Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. A distribution G on $\mathbb{R}^d$ is (ϵ,δ)-stable with respect to $\mu \in \mathbb{R}^d$ if for any weight function $w : \mathbb{R}^d \to [0,1]$ with $\mathbb{E}_{X\sim G}[w(X)] \geq 1 - \epsilon$ we have that*

$$\|\mu_{w,G} - \mu\|_2 \leq \delta \quad \text{and} \quad \left\|\overline{\mathbf{\Sigma}}_{w,G} - \mathbf{I}_d\right\|_2 \leq \delta^2/\epsilon \,.$$

We call a set of points S (ϵ,δ)-stable when the uniform distribution on S is stable :

Definition 9.2.9 ((ϵ,δ)-stable set)**.** *Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. A finite set $S_0 \subset \mathbb{R}^d$ is (ϵ,δ)-stable with respect to $\mu \in \mathbb{R}^d$ if the empirical distribution $\mathcal{U}(S_0)$ is (ϵ,δ)-stable with respect to μ.*

We begin by stating some examples of stable distributions (see [DK19] for more details). If G is a subgaussian distribution with identity covariance, then G is (ϵ,δ)-stable with $\delta = O(\epsilon\sqrt{\log(1/\epsilon)})$. If G is a distribution with covariance at most identity,

i.e., $\mathbf{\Sigma}_G \preceq \mathbf{I}_d$, then G is (ϵ, δ)-stable with $\delta = O(\sqrt{\epsilon})$. Interpolating these two results, we have that if G is a distribution with identity covariance and bounded k-th moment for $k \geq 4$, i.e., $(\mathbb{E}_{X\sim G}[|v^\top (X-\mu)|^k])^{1/k} = O(1)$, then G is (ϵ, δ)-stable with $\delta = O(\epsilon^{1-1/k})$. Furthermore, it is known that $\mathrm{poly}(d/\epsilon)$ i.i.d. samples from these distributions also yields a set that contains a large stable subset (see, for example, [DK19; DKP20; DHL19; DKKLMS16]):

Fact 9.2.10 ([DK19])**.** *A set of $O(d/(\epsilon^2 \log(1/\epsilon)))$ i.i.d. samples from an identity covariance subgaussian distribution is $(\epsilon, O(\epsilon\sqrt{\log(1/\epsilon)}))$-stable with respect to μ with high probability. Similarly, a set of $\tilde{O}(d/\epsilon)$ i.i.d. samples from a distribution X with $\mathrm{Cov}[X] \preceq \mathbf{I}_d$ contains a large subset S, which is $O(\epsilon, O(\sqrt{\epsilon}))$-stable with respect to its mean $\mathbb{E}[X]$ with high probability.*

The basic fact regarding stability, which is the starting point of many robust estimation algorithms, is that any slight modification of a stable distribution can not perturb the mean by a large amount, unless it significantly changes its covariance (see, for example, [DKKLMS16; LRV16; DK19]). Here we require a slightly different statement than that of [DK19], and hence provide a proof in Appendix F.1 for completeness.

Lemma 9.2.11 (Certificate Lemma)**.** *Let G be an (ϵ, δ)-stable distribution with respect to $\mu \in \mathbb{R}^d$, for some $0 < \epsilon < 1/3$ and $\delta \geq \epsilon$. Let P be a distribution with $d_{\mathrm{TV}}(P, G) \leq \epsilon$. Denoting by $\mu_P, \mathbf{\Sigma}_P$ the mean and covariance of P, if $\lambda_{\max}(\mathbf{\Sigma}_P) \leq 1 + \lambda$, for some $\lambda \geq 0$, then $\|\mu_P - \mu\|_2 = O(\delta + \sqrt{\epsilon\lambda})$.*

Given Fact 9.2.4, we can essentially think of an ϵ-corrupted version of a stable distribution as a mixture of a stable distribution with a noise distribution, as shown below (see Appendix F.1 for a proof).

Lemma 9.2.12. *For any $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$, if a distribution G is $(2\epsilon, \delta)$-stable with respect to $\mu \in \mathbb{R}^d$, and P is an ϵ-corrupted version of G in total variation distance, there exist distributions G_0 and B such that $P = (1-\epsilon)G_0 + \epsilon B$ and G_0 is (ϵ, δ)-stable with respect to μ.*

We continue with some technical claims related to stability that we prove in Appendix F.1. Let G be an (ϵ, δ)-stable distribution with respect to μ and w a weight function with $\mathbb{E}_{X\sim G}[w(X)] \geq 1-\epsilon$. Denoting by G_w the re-weighted distribution $G_w(x) := G(x)w(x)/\int_{\mathbb{R}^d} w(x)G(x)\mathrm{d}x$, the stability of G directly implies that $1-\delta^2/\epsilon \leq \mathbb{E}_{X\sim G_w}[(v^\top(X-\mu))^2] \leq 1+\delta^2/\epsilon$. We require a generalization of this fact for a matrix $\mathbf{U}$ in place of v and an arbitrary vector b in place of μ:

Lemma 9.2.13. *Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. Let $w : \mathbb{R}^d \to [0,1]$ such that $\mathbb{E}_{X\sim G}[w(X)] \geq 1-\epsilon$ and let G be an (ϵ,δ)-stable distribution with respect to $\mu \in \mathbb{R}^d$. For any matrix $\mathbf{U} \in \mathbb{R}^{d\times d}$ and any vector $b \in \mathbb{R}^d$, we have that*

$$\mathop{\mathbb{E}}_{X\sim G_w}\left[\|\mathbf{U}(X-b)\|_2^2\right] = \|\mathbf{U}\|_F^2(1\pm\delta^2/\epsilon) + \|\mathbf{U}(\mu-b)\|_2^2 \pm 2\delta\,\|\mathbf{U}\|_F^2\|\mu-b\|_2\,.$$

We use this to show Corollary 9.2.14, which will be required when proving correctness of our algorithm. Although its exact role will become clearer later on, the corollary will be relevant to our analysis because we will filter out outliers using scores of the form $\|\mathbf{U}(x-b)\|_2^2$ for each point x.

Corollary 9.2.14. *Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. Let G be an (ϵ,δ)-stable distribution with respect to $\mu \in \mathbb{R}^d$. Let a matrix $\mathbf{U} \in \mathbb{R}^{d\times d}$ and a function $w : \mathbb{R}^d \to [0,1]$ with $\mathbb{E}_{X\sim G}[w(X)] \geq 1-\epsilon$. For the function $\tilde{g}(x) = \|\mathbf{U}(x-b)\|_2^2$, we have that*

$$\begin{aligned}(1-\epsilon)\|\mathbf{U}\|_F^2(1-\delta^2/\epsilon - 2\delta\|b-\mu\|_2) &\leq \mathop{\mathbb{E}}_{X\sim G}[w(X)\tilde{g}(X)] \\ &\leq \|\mathbf{U}\|_F^2\left(1+\delta^2/\epsilon + \|b-\mu\|_2^2 + 2\delta\|b-\mu\|_2\right)\,.\end{aligned}$$

9.3 Filtering Algorithm with Small Number of Iterations

In this section, we develop a filtering algorithm (in the batch setting) that terminates in $\mathrm{polylog}(d/\epsilon)$ iterations for any stable set. This leads to an algorithm that runs in near-linear time, i.e., $nd\,\mathrm{polylog}(nd/\epsilon)$, generalizing the results of [DHL19; DKKLT22]. Crucially, this algorithm will form the building block of our streaming algorithm in Section 9.4. We remark that the algorithm of this section works even against the *strong-contamination model* (Definition 9.3.1 below), where the outliers may not be i.i.d. samples from any distribution, but are allowed to be completely arbitrary.

Definition 9.3.1 (Strong Contamination Model)**.** *Given a parameter $0 < \epsilon < 1/2$ and a class of distributions $\mathcal{D}$, the strong adversary operates as follows: The algorithm specifies a number of samples n, then the adversary draws a set of n i.i.d. samples from some $D \in \mathcal{D}$ and after inspecting them, removes up to ϵn of them and replaces them with arbitrary points. The resulting set is given as input to the learning algorithm. We call a set ϵ-corrupted if it has been generated by the above process.*

The main result of this section is the following.

Theorem 9.3.2. *Let $d \in \mathbb{Z}_+$, $0 < \tau < 1$, $0 < \epsilon < \epsilon_0$ for a sufficiently small constant ϵ_0, and $\delta \geq \epsilon$. Let S_0 be a set of n points that is $(C\epsilon, \delta)$-stable with respect to the (unknown) vector $\mu \in \mathbb{R}^d$, for a sufficiently large constant $C > 0$. Let S be an ϵ-corrupted version of S_0 in the strong contamination model. There exists an algorithm that given ϵ, δ, τ, and S, runs in time $nd\,\mathrm{polylog}\,(d, n, 1/\epsilon, 1/\tau)$, and outputs a vector $\widehat{\mu}$ such that, with probability at least $1 - \tau$, it holds $\|\mu - \widehat{\mu}\|_2 = O(\delta)$.*

We note that Theorem 9.3.2 applies to any stable set. By Fact 9.2.10, we directly obtain (i) an $O(\epsilon\sqrt{\log(1/\epsilon)})$-accurate estimator given $O(d/(\epsilon^2/\log(1/\epsilon)))$ many ϵ-corrupted samples from an identity covariance subgaussian distribution; and (ii) an $O(\sqrt{\epsilon})$-accurate

estimator for any distribution $X \sim D$ with $\mathrm{Cov}[X] \preceq \mathbf{I}_d$, given $\tilde{O}(d/\epsilon)$ many ϵ-corrupted samples.

9.3.1 Setup and Algorithm Description

The pseudocode of the algorithm establishing Theorem 9.3.2 is presented in Algorithm 6. We will define the necessary notation as needed (see the pseudocode for details). First, we assume that the distribution over the input samples is of the form $P = (1-\epsilon)G + \epsilon B$, where G is the uniform distribution over the stable set of inliers and B is the uniform distribution on the outliers. Although this mixture may seem to suggest that the adversary only adds points, it is without loss of generality. Indeed, in the case that the adversary also removes points, we can think of G as the distribution of the remaining inliers (which continues to be stable with slightly worse parameters; see Lemma 9.2.12).

We begin with a high-level explanation of Algorithm 6. At each iteration t, we assign a weight $w_t(x) \in [0,1]$ to each point x. Let P_t be the distribution on S, weighted according to w_t. Let μ_t and $\mathbf{\Sigma}_t$ be the mean and covariance of P_t, respectively. We want to assign scores to each point, using spectral properties of $\mathbf{\Sigma}_t$ and the stability of inliers, so that the scores over outliers are more than those of inliers. Essentially, if a direction v has variance larger than $1 + \Omega(\delta^2/\epsilon)$, then the stability of inliers implies that this must be due to outliers. Thus, we can assign scores based on the values $(v^\top(x-\mu_t))^2$ that have provably more mass on outliers than inliers. The filters proposed in [DKKLMS16; DKKLMS17] assigned scores based on a single direction, the leading eigenvector of $\mathbf{\Sigma}_t$, and can take as many as $\Omega(d)$ iterations (see Section 9.1.2).

To reduce the number of iterations, we need to filter in all directions of large variance simultaneously. Letting $\mathbf{B}_t \approx \mathbf{\Sigma}_t - (1 - C_1\delta^2/\epsilon)\mathbf{I}_d$, we would like to filter along all directions where the eigenvalue of $\mathbf{B}_t$ is within a constant factor from $\lambda_t := \|\mathbf{B}_t\|_2$, not necessarily the leading eigenvector of $\mathbf{B}_t$. As we show in Section 9.3.4, this can

be approximately achieved by assigning scores for each point x based on $g_t(x) := \|\mathbf{M}_t(x-\mu_t)\|_2^2$, where $\mathbf{M}_t = \mathbf{B}_t^{\log d}$. At a high level, this happens because the spectrum of $\mathbf{M}_t$ is distributed across along *a*ll large eigenvectors of $\mathbf{B}_t$.

Algorithm 6 Robust Mean Estimation in polylog iterations

Input: $S = \{x_i\}_{i\in[n]}, \delta, \epsilon$
Let $C_1 \geq 22$, C be a sufficiently large constant, $C_2 = 100C$ and $C_3 = 0.1$.
Let $R = \sqrt{(d/\epsilon)(1+\delta^2/\epsilon)}$.
Let $P = (1-\epsilon)G + \epsilon B$ be the empirical distribution on the points from S. [12]
Let $K = C\log d \log(dR/\epsilon)$, $L = C\log((n+d)K/\tau)$.
Obtain a naïve estimate $\widehat{\mu}$ of μ with $\|\widehat{\mu}-\mu\|_2 \leq 4R$.
Initialize $w_1(x) \leftarrow \mathbb{I}\{\|x-\widehat{\mu}\|_2 \leq 5R\}$ for all $x \in S$.
for $t \in [K]$ **do**
    Let P_t be the distribution of P weighted by w_t.
    Let $\mu_t, \mathbf{\Sigma}_t$ be the mean and covariance of P_t.
    Let $\mathbf{B}_t = (\mathbb{E}_{X\sim P}[w_t(X)])^2\mathbf{\Sigma}_t - \left(1 - C_1\frac{\delta^2}{\epsilon}\right)\mathbf{I}_d$
    Let $\mathbf{M}_t = \mathbf{B}_t^{\log d}$. ▷ $\mathbf{M}_t$ does not need to be explicitly calculated.
    Let $\lambda_t = \|\mathbf{B}_t\|_2$
    Find $\widehat{\lambda}_t \in [0.8\lambda_t, 1.2\lambda_t]$ by power iteration. ▷ See Remark 9.3.3 for efficient implementation.
    if $\widehat{\lambda}_t > C_2\delta^2/\epsilon$ **then**
        for $j \in [L]$ **do**
            $z_{t,j} \sim \mathcal{U}(\{\pm 1\}^d)$,
            $v_{t,j} \leftarrow \mathbf{M}_t z_{t,j}$. ▷ See Remark 9.3.3 for efficient implementation.
        end for
        Denote by $\mathbf{U}_t$ the matrix having the vectors $\frac{1}{\sqrt{L}}v_{t,j}$ for $j \in [L]$ as rows.
        Let $\tilde{g}_t(x) = \|\mathbf{U}_t(x-\mu_t)\|_2^2$ and $\tilde{\tau}_t(x) = \tilde{g}_t(x)\,\mathbb{I}\{\tilde{g}_t(x) > C_3\|\mathbf{U}_t\|_F^2\widehat{\lambda}_t/\epsilon\}$,
        $\ell_{\max} \leftarrow (dR/\epsilon)^{C\log d}$, $T \leftarrow 0.01\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2$.
        $w_{t+1} \leftarrow \mathrm{DownweightingFilter}(P, w_t, \tilde{\tau}_t, R, T, \ell_{\max})$ ▷ Algorithm 7
    end if
end for
return μ_t.

Even though assigning scores based on $\mathbf{M}_t$, i.e., $g_t(x)$, reduces the number of iterations, computing $g_t(x)$ for all $x \in S$ is slow. We thus use a Johnson-Lindenstrauss (JL) sketch of $\mathbf{M}_t$, denoted by $\mathbf{U}_t$. We denote by $\tilde{g}_t(x) := \|\mathbf{U}_t(x-\mu_t)\|_2$ the resulting scores. We claim that the set $\{\tilde{g}_t(x)\}_{x\in S}$ can be calculated in time $\tilde{O}(nd)$ such that for each $x \in S$,

[12] Without loss of generality, outliers are within $O(R)$ from μ in ℓ_2-norm. This is ensured in Line 7, which removes only ϵ-fraction of inliers (Claim 9.3.12).

Algorithm 7 Downweighting Filter

1: **Input**: $P, w, \tilde{\tau}, R, T, \ell_{\max}$
2: $r \leftarrow CdR^{2+4\log d}$.
3: Let $w_\ell(x) = w(x)(1 - \tilde{\tau}(x)/r)^\ell$.
4: Find the smallest $\ell \in \{1, \ldots, \ell_{\max}\}$ satisfying $\mathbb{E}_{X\sim P}\left[w_\ell(X)\tilde{\tau}(X)\right] \leq 2T$ using binary search.
5: **return** w_ℓ.

$\tilde{g}_t(x) \approx g_t(x)$. First, we will show (see Lemma 9.3.5) that $\mathbf{U}_t$ can be as small as $L \times d$, where L is some $\mathrm{polylog}(nd/(\epsilon\tau))$, which follows from the classical JL lemma (stating that n points can be linearly embedded into a $\log n$-dimensional space). Also, each row of $\mathbf{U}_t$ can be computed by repeatedly multiplying a vector $\log d$ times by $\mathbf{B}_t$ (Line 18). By the following remark, all rows of $\mathbf{U}_t$ can be computed in time $\tilde{O}(nd)$ and thus, each iteration of Algorithm 6 runs in near-linear time.

Remark 9.3.3. (*Efficient Implementation*) *Note that for any $v \in \mathbb{R}^d$, the vector $\mathbf{B}_t v$ can be calculated in $O(nd)$ time. This is because $\mathbf{\Sigma}_t v = \sum_{x\in S} w_t(x)(v^\top(x-\mu_t))(x-\mu_t)/(\sum_{x\in S} w_t(x))$ which means that the result can be computed in $O(nd)$ time by calculating μ_t and $v^\top(x-\mu_t)$ first. Regarding Line 14, an approximate large eigenvector can be computed via power iteration, i.e., starting from a random Gaussian vector and multiplying by $\mathbf{B}_t$ iteratively $\log d$ many times (see, for example, [BHK20]). As mentioned above, each of these multiplications can be done in $O(nd)$ time.*

For the proof of correctness, we require that the JL and spectral approximations used by the algorithm are sufficiently accurate. We prove that the following event occurs with high probability.

Condition 9.3.4 (Deterministic Conditions For Algorithm 6)**.** *For all $t \in [K]$, the following hold:*

1. Spectral norm of $\mathbf{B}_t$: $\widehat{\lambda}_t \in [0.1\lambda_t, 20\lambda_t]$.

2. Frobenius norm: $\|\mathbf{U}_t\|_F^2 \in [0.8\|\mathbf{M}_t\|_F^2, 1.2\|\mathbf{M}_t\|_F^2]$.

3. Scores: *For all* $x \in S$, $\tilde{g}_t(x) \in [0.8 g_t(x), 1.2 g_t(x)]$.

9.3.2 Establishing the Deterministic Conditions

In this section, we establish that Condition 9.3.4 holds with high probability. Regarding Item 1 of this condition, an approximate large eigenvector can be computed via power iteration as described in Remark 9.3.3. This gives us an algorithm that runs in time $O(nd \log d \log(K/\tau))$ and satisfies Item 1 with probability $1 - \tau$.

We now move to the other two conditions. The claim is that instead of using the matrix $\mathbf{M}_t$ to calculate the scores, it suffices to store and use only a small set of random projections $\{\mathbf{M}_t z_{t,j}\}_{j \in [L]}$. This is exactly the Johnson-Lindenstrauss sketch that is computed in Line 16 of Algorithm 6. Using Fact 9.2.7, we get the following guarantees (see Appendix F.2.1 for the proof).

Lemma 9.3.5. *Fix a set of* n *points* $x_1, \ldots, x_n \in \mathbb{R}^d$. *For* $t \in [K]$, *define* $g_t(x) := \|\mathbf{M}_t(x - \mu_t)\|_2^2$ *and let* $\tilde{g}_t(x), v_{t,j}$ *as in Algorithm 6. If* C *is a sufficiently large constant and* $L = C \log((n + d)K/\tau)$, *with probability at least* $1 - \tau$, *for every* $t \in [K]$ *we have the following:*

1. $0.8 g_t(x_i) \le \tilde{g}_t(x_i) \le 1.2 g_t(x_i)$ *for every* $i \in [n]$,
2. $0.8 \|\mathbf{M}_t\|_F^2 \le \left(\frac{1}{L} \sum_{j=1}^{L} \|v_{t,j}\|_2^2\right) \le 1.2 \|\mathbf{M}_t\|_F^2$.

This concludes the proof that Condition 9.3.4 is satisfied with high probability.

9.3.3 Downweighting Filter

We use the following re-weighting procedure also used in [DHL19]. Recall that P denotes the empirical distribution on the samples, which we write as $P = (1 - \epsilon)G + \epsilon B$, where G and B are the contributions from the good and bad samples respectively. Roughly speaking, our filter guarantees two things when going from the weights $w(x)$ to $w'(x)$:

1. *The weight removed from the outliers is greater than the weight removed from the inliers.*

2. $\mathbb{E}_{X\sim P}[w'(X)\tilde{\tau}(X)] \le 2\,\mathbb{E}_{X\sim G}[w(X)\tilde{\tau}(X)]$, i.e., *the weighted mean of scores after filtering over both inliers and outliers is at most twice the weighted mean of scores of inliers* before *filtering*.

Regarding the first guarantee, since the fraction of outliers is at most ϵ, this ensures that the filtered distribution P_t will never be more than $O(\epsilon)$-far in total variation distance from the initial (corrupted) distribution, and thus the condition of the certificate lemma that $d_{\mathrm{TV}}(P, P_t) \le O(\epsilon)$ will always be satisfied. The second guarantee ensures that the filtering step reduces the average score significantly. We prove the following in Appendix F.2.2.

Lemma 9.3.6. *Let $P = (1-\epsilon)G + \epsilon B$ be the empirical distribution on n samples, as in Algorithm 6. If $(1-\epsilon)\,\mathbb{E}_{X\sim G}[w(X)\tilde{\tau}(X)] \le T$, $\|\tilde{\tau}\|_\infty \le r$, and $\ell_{\max} > r/T$, then Algorithm 7 modifies the weight function w to w' such that*

1. $(1-\epsilon)\,\mathbb{E}_{X\sim G}[w(X) - w'(X)] < \epsilon\,\mathbb{E}_{X\sim B}[w(X) - w'(X)]$,

2. $\mathbb{E}_{X\sim P}\left[w'(X)\tilde{\tau}(X)\right] \le 2T$,

and the algorithm terminates after $O(\log(\ell_{\max}))$ iterations, each of which takes $O(n)$ time.

We note that the two conditions $\|\tilde{\tau}\|_\infty \le r$, $\ell_{\max} > r/T$ of Lemma 9.3.6 hold by our choice of $\ell_{\max}$ and r inside Algorithm 6 and Algorithm 7 as follows. For $\|\tilde{\tau}\|_\infty$, we have the following upper bound

$$\tilde{\tau}_t(x) \le \tilde{g}_t(x) \le \|\mathbf{U}_t(x-\mu_t)\|_2^2 \lesssim R^2\|\mathbf{U}_t\|_2^2 \lesssim R^2\|\mathbf{M}_t\|_F^2 \lesssim R^2\|\mathbf{\Sigma}_t\|_2^{2\log d} = O(dR^{2+4\log d})\,, \tag{9.1}$$

where we used the guarantee of our JL sketch that $\|\mathbf{U}_t\|_2^2 \leq 1.2\|\mathbf{M}_t\|_F^2$ (Lemma 9.3.5). A crude upper bound on r/T follows from the following inequalities:

$$\frac{r}{T} \lesssim \frac{dR^{2+4\log d}}{\widehat{\lambda}_t \|\mathbf{M}_t\|_F^2} \lesssim \frac{dR^{2+4\log d}}{\lambda_t \|\mathbf{M}_t\|_F^2} \lesssim \left(\frac{dR}{\delta^2/\epsilon}\right)^{O(\log d)},$$

where the first inequality uses the values of r and T as set in the algorithm, the second inequality uses Items 1 and 2 of the deterministic conditions of Condition 9.3.4, and the last inequality uses the fact that $\|\mathbf{M}_t\|_F^2 \geq \|\mathbf{M}_t\|_2^2$ and $\|\mathbf{M}_t\|_2^2$ cannot be smaller than $(C_2\delta^2/\epsilon)^{O(\log d)}$ (otherwise Line 22 terminates the algorithm).

We now use the guarantees of Algorithm 7 as follows. We first show that the weighted mean of the inliers' scores is small.

Lemma 9.3.7. *Under the setting of Algorithm 6 and the deterministic Condition 9.3.4, we have that $\mathbb{E}_{X\sim G}[w_t(X)\tau_t(X)]$ and $\mathbb{E}_{X\sim G}[w_t(X)\tilde{\tau}_t(X)]$ are bounded from above by $c\lambda_t\|\mathbf{M}_t\|_F^2$ for some constant c of the form $c = C/C_2$, where C_2 is the constant used in Line 15 and C is some absolute constant.*

The proof is based on stability arguments from Section 9.2.2 and can be found in Appendix F.2.3.

Remark 9.3.8. *In our analysis, it will be important that the constant c in Lemma 9.3.7 can be made sufficiently smaller than 1, for example, $c < 0.01$. This can be achieved by choosing C_2 to be a large enough constant.*

Using Algorithm 7, we get that the weighted sum of scores after filtering is also small.

Lemma 9.3.9. *Under the setting of Algorithm 6 and the deterministic Condition 9.3.4, we have that $\epsilon\,\mathbb{E}_{X\sim B}[w_{t+1}(X)\tilde{\tau}_t(X)] < c\lambda_t\|\mathbf{M}_t\|_F^2$, with c being of the form $c = C/C_2$,*

where C_2 is the constant used in Line 15 and C is some absolute constant. Furthermore, $\epsilon\,\mathbb{E}_{X\sim B}[w_{t+1}(X)\tau_t(X)] < c\lambda_t\|\mathbf{M}_t\|_F^2$.

Proof. The first claim follows by the stopping condition of the algorithm, Lemma 9.3.7, and the fact that $w_{t+1} \le w_t$. We now prove the second conclusion by relating τ to $\tilde{\tau}$: Recall that we denote by S the ϵ-corrupted version of the original set of samples S_0. Since $\tilde{g}_t(x)$ is within a constant factor of $g_t(x)$ (Condition 9.3.4) for all $x \in S$, the scores $\tilde{\tau}_t(x)$ and $\tau_t(x)$ are comparable (up to an additive term) as shown below.

Claim 9.3.10. *In the setting of Algorithm 6 and under the Condition 9.3.4, if $x \in S$, we have that $\tau_t(x) \le 1.25\tilde{\tau}_t(x) + 3C_3(\lambda_t/\epsilon)\mathrm{tr}(\mathbf{M}_t^2)$, where C_3 is the constant used in Algorithm 6.*

We prove Claim 9.3.10 in Appendix F.2.5. Using Claim 9.3.10, we have the following set of inequalities:

$$\begin{aligned}
\epsilon \cdot \mathop{\mathbb{E}}_{X\sim B}[w_{t+1}(X)\tau_t(X)] &= \frac{1}{n}\sum_{x_i\in S\setminus S_0} w_{t+1}(x_i)\tau_t(x_i) \\
&\le 3C_3\lambda_t\,\|\mathbf{M}_t\|_F^2 + \frac{1}{n}1.25\sum_{i\in S\setminus S_0} w_{t+1}(x_i)\tilde{\tau}_t(x_i) \\
&\qquad\qquad \text{(using Claim 9.3.10)} \\
&= 3C_3\lambda_t\|\mathbf{M}_t\|_F^2 + 1.25\epsilon \mathop{\mathbb{E}}_{X\sim B}[w_{t+1}(X)\tilde{\tau}_t(X)] \\
&\le 3C_3\lambda_t\|\mathbf{M}_t\|_F^2 + 1.25c\lambda_t\|\mathbf{M}_t\|_F^2 \\
&\qquad\qquad \text{(using the first part of Lemma 9.3.9)} \\
&< 10(C/C_2)\lambda_t\|\mathbf{M}_t\|_F^2\,. \qquad \text{(using the value of } C_3\text{)}
\end{aligned}$$

The last inequality above uses the fact that the constant C_3 is chosen to be $C_3 = C/C_2$ in Algorithm 6, where C is a sufficiently large constant. □

9.3.4 Correctness of Algorithm 6: Proof of Theorem 9.3.2

The rest of this section is dedicated to proving Theorem 9.3.2. We first state the correctness of the naïve approximation step of Line 6, then record the invariants of the algorithm in Section 9.3.4.1, and finally show in Section 9.3.4.2 that it suffices for the number of iterations K to be bounded by some $\mathrm{polylog}(d, R, 1/\epsilon)$.

The naïve approximation step of Line 6 is based on the following folklore fact (see Appendix F.2.4 for more details).

Claim 9.3.11. *Let the fraction of outliers be $\epsilon < 1/10$ and a parameter $0 < \tau < 1$. Let the distribution $P = (1-\epsilon)G + \epsilon B$. Let $R > 0, \mu \in \mathbb{R}^d$ be such that $\mathbb{P}_{X\sim G}[\|X-\mu\|_2 > R] \le \epsilon$. There is an estimator $\widehat{\mu}$ on $k = O(\log(1/\tau))$ samples from P such that $\|\widehat{\mu}-\mu\|_2 \le 4R$ with probability at least $1-\tau$. Furthermore, $\widehat{\mu}$ can be computed in time $O(k^2 d)$ and memory $O(kd)$.*

Claim 9.3.12 below gives a valid upper bound on R using the (ϵ, δ)-stability of the good distribution.

Claim 9.3.12. *If $R = \sqrt{\frac{d}{\epsilon}(1+\delta^2/\epsilon)}$, then $\mathbb{P}_{X\sim G}[\|X-\mu\|_2 > R] \le \epsilon$.*

Proof. By Markov's inequality, we have that

$$\mathop{\mathbb{P}}_{X\sim G}\left[\|X-\mu\|_2^2 \ge \frac{d}{\epsilon}\left(1+\delta^2/\epsilon\right)\right] \le \epsilon \frac{\mathbb{E}_{X\sim G}[\|X-\mu\|_2^2]}{d(1+\delta^2/\epsilon)} \le \epsilon\,.$$

□

9.3.4.1 Invariants of Algorithm 6

Recall that the end goal is to obtain a filtered version, P_t, of P that is not too far from P in total variation distance $d_{\mathrm{TV}}(P, P_t) = O(\epsilon)$, and satisfies that $\|\mathbf{B}_t\|_2 = O(\delta^2/\epsilon)$. For the first condition to be satisfied, we ensure that the Downweighting filter removes more weight from G than B (Lemma 9.3.6). Using this, we show that $\mathbb{E}_{X\sim G}[w_t(X)] \ge 1 - O(\epsilon)$,

which implies the bound on the total variation distance (Claim 9.3.13). The proofs are deferred to Appendix F.2.5.

Claim 9.3.13. *Under Condition 9.3.4, Algorithm 6 maintains the following invariant:* $\mathbb{E}_{X\sim G}[w_t(X)] \geq 1 - 3\epsilon$. *In particular, if* $\epsilon \leq 1/8$, *then* $d_{\mathrm{TV}}(P_t, P) \leq 9\epsilon$.

The following properties of $\mathbf{B}_t$ as PSD operator will also be useful later on.

Claim 9.3.14. *Under Condition 9.3.4, if* $C_1 \geq 22$, $\mathbf{B}_t \succeq (0.5C_1\delta^2/\epsilon)\mathbf{I}_d$ *for every* $t \in [K]$.

The proof of Claim 9.3.14 is provided in Appendix F.2.5. Although just showing that $\mathbf{B}_t \succeq 0$ would suffice for this section, the slightly stronger bound of the above claim will be useful in Section 9.4. Claim 9.3.14 follows from Claim 9.3.13 and the stability of G. In particular, the stability of G implies that $\overline{\mathbf{\Sigma}}_G \succeq (1 - \delta^2/\epsilon)\mathbf{I}_d$. We now prove the following claim, which is the reason for having the multiplicative factor of $\mathbb{E}_{X\sim P}[w_t(X)]^2$ in the definition of $\mathbf{B}_t$.

Claim 9.3.15. *We have that* $\mathbf{B}_{t+1} \preceq \mathbf{B}_t$ *for every* $t \in [K]$.

Proof. We use the alternative definition of the covariance matrix: Let X, Y be i.i.d. from P, then

$$\mathbf{\Sigma}_t = \frac{1}{2(\mathbb{E}_{X\sim P}[w_t(X)])^2} \mathop{\mathbb{E}}_{X,Y\sim P}[w_t(X)w_t(Y)(X-Y)(X-Y)^\top] \,.$$

Since $w_{t+1}(x) \leq w_t(x)$ for all x, this completes the proof. □

9.3.4.2 Reducing the Potential Function

Recall that each iteration of Algorithm 6 can be implemented in near-linear time. Thus, it remains to show that the choice $K = C \log d \log(dR/\epsilon)$ suffices to guarantee correctness of our algorithm. We now sketch the proof using a potential function argument. Let Λ_t be the vector in $\mathbb{R}^d$ containing the eigenvalues of $\mathbf{B}_t$. Recall that our goal is to show

that $\|\mathbf{B}_t\|_2 = \|\Lambda_t\|_\infty = O(\delta^2/\epsilon)$ in polylog many iterations. Let $p := 2\log d$. Since $\|x\|_p = \Theta(\|x\|_\infty)$ for any $x \in \mathbb{R}^d$, we are motivated to use the potential function $\phi_t := \|\Lambda_t\|_p^p$. We now focus on showing that ϕ_t decreases rapidly. Observe that for any $i \in \mathbb{Z}_+$, $\mathrm{tr}(\mathbf{B}_t^i) = \|\Lambda_t\|_i^i$. We start with the following inequalities (and explain them directly below):

$$\begin{aligned}
\phi_{t+1} = \|\Lambda_{t+1}\|_p^p &\le \left(d^{\frac{1}{p(p+1)}}\|\Lambda_{t+1}\|_{p+1}\right)^p \\
&= d^{\frac{1}{p+1}}\left(\|\Lambda_{t+1}\|_{p+1}^{p+1}\right)^{\frac{p}{p+1}} \\
&= d^{\frac{1}{p+1}}(\mathrm{tr}(\mathbf{B}_{t+1}^{p+1}))^{\frac{p}{p+1}} \\
&\le d^{\frac{1}{p}}\left(\mathrm{tr}(\mathbf{M}_t\mathbf{B}_{t+1}\mathbf{M}_t)\right)^{\frac{p}{p+1}} \ , \qquad (9.2)
\end{aligned}$$

where the first line uses Fact 9.2.1, the third one uses $\mathrm{tr}(\mathbf{B}_{t+1}^i) = \|\Lambda_{t+1}\|_i^i$, and the last line uses Fact 9.2.2 along with the fact $\mathbf{B}_{t+1} \preceq \mathbf{B}_t$, which holds because removing points can only make their covariance smaller; see Section 9.3.4 for more details.

Then the goal becomes to bound from above the term $\mathrm{tr}(\mathbf{M}_t\mathbf{B}_{t+1}\mathbf{M}_t)$. The claim is that $\mathrm{tr}(\mathbf{M}_t\mathbf{B}_{t+1}\mathbf{M}_t)$ is related to $\mathbb{E}_{X\sim P}[w_{t+1}(X)\tau_t(X)]$, and thus can be bounded by $c\lambda_t\|\mathbf{M}_t\|_F^2$. Using the guarantees of the Downweighting filter (Lemma 9.3.9), we prove the following result:

Lemma 9.3.16. *Consider the setting of Algorithm 6 and assume that Condition 9.3.4 holds. Then $\mathrm{tr}(\mathbf{M}_t\mathbf{B}_{t+1}\mathbf{M}_t) \le c\lambda_t\|\mathbf{M}_t\|_F^2$ for some c of the form C/C_2, where C_2 is the constant used in Line 15 and C is some absolute constant.*

Before giving the details regarding Lemma 9.3.16, we first show that it suffices to prove that the potential function decreases by a multiplicative factor. In the rest of the proof, we will assume that $c < 0.1$, which can be guaranteed by taking C_2 to be a sufficiently large constant (cf. Remark 9.3.8). We continue with Equation (9.2) as

follows:

$$\begin{aligned}
\phi_{t+1} &\le d^{\frac{1}{p}} \left(\mathrm{tr}(\mathbf{M}_t \mathbf{B}_{t+1} \mathbf{M}_t)\right)^{\frac{p}{p+1}} \\
&\le d^{\frac{1}{p}} \left(c\|\Lambda_t\|_\infty \|\Lambda_t\|_p^p\right)^{\frac{p}{p+1}} && \text{(using Lemma 9.3.16)} \\
&\le d^{\frac{1}{p}} c^{\frac{p}{p+1}} \left(\|\Lambda_t\|_p \|\Lambda_t\|_p^p\right)^{\frac{p}{p+1}} && \text{(using } \|\Lambda_t\|_\infty \le \|\Lambda_t\|_i \text{ for } i \ge 1) \\
&= d^{\frac{1}{p}} c^{\frac{p}{p+1}} \|\Lambda_t\|_p^p \\
&\le 3\sqrt{c}\|\Lambda_t\|_p^p \le 0.9999\phi_t \ ,
\end{aligned}$$

where the last line uses that $d^{1/p} = \exp(\frac{\log d}{2 \log d}) \le 3$, $p/(p+1) \ge 0.5$, and $c < 1$. We thus get the desired convergence.

The final step is to bound the number of iterations needed for Lemma 9.2.11 to ensure that $\|\mu_t - \mu\|_2 = O(\delta)$. Concretely, due to our naïve pruning, at the beginning of the algorithm we have the upper bound $\phi_1 \le dR^{O(\log d)}$. After K iterations, we have that $\phi_K \le 0.99^K dR^{O(\log d)}$. Setting K as

$$K = C \log d \log\left(\frac{dR}{\delta^2/\epsilon}\right) \tag{9.3}$$

suffices to have that $\|\mathbf{B}_K\|_2 \le (dO(\delta^2/\epsilon)^{2\log d})^{\frac{1}{2\log d}} = O(\delta^2/\epsilon)$. This implies that

$$\|\mathbf{\Sigma}_K\|_2 \le \frac{1}{\mathbb{E}_{X\sim P}[w_K(X)]^2}\left(\|\mathbf{B}_K\|_2 + 1\right) \le 1 + \left(\frac{1}{(1-3\epsilon)^2} - 1\right) + O\left(\frac{\delta^2}{\epsilon}\right) \le 1 + O\left(\frac{\delta^2}{\epsilon}\right) \ , \tag{9.4}$$

where we used that $\mathbb{E}_{X\sim P}[w_K(X)])^2 \ge 1 - 3\epsilon$, $\delta \ge \epsilon$, and $\epsilon \le \epsilon_0$. An application of Lemma 9.2.11 shows that the estimate has error at most $\|\mu_t - \mu\|_2 = O(\delta)$. This completes the proof of Theorem 9.3.2. The rest of the section focuses on proving Lemma 9.3.16.

Proof Sketch of Lemma 9.3.16 Before giving the full proof of Lemma 9.3.16, we provide a brief proof sketch. By the definition of $\mathbf{B}_{t+1}$, we have the following (the full proof is deferred to the end of this section)

$$\mathrm{tr}(\mathbf{M}_t \mathbf{B}_{t+1} \mathbf{M}_t) \leq \mathop{\mathbb{E}}_{X\sim P}[w_{t+1}(X)] \mathop{\mathbb{E}}_{X\sim P}[w_{t+1}(X) g_t(X)] - \left(1 - C_1 \frac{\delta^2}{\epsilon}\right) \|\mathbf{M}_t\|_F^2 \,.$$

In order to bound from above $\mathbb{E}_{X\sim P}[w_{t+1}(X) g_t(X)]$, one can consider the contribution due to inliers (distribution G) and contribution due to outliers (distribution B). Using the stability of inliers and Corollary 9.2.14, we have that $\mathbb{E}_{X\sim G}[w(X) g_t(X)] \leq (1 + c\lambda_t)\|\mathbf{M}_t\|_F^2$, for any weight function w satisfying the conditions of Corollary 9.2.14. We know that w_{t+1} satisfies them because of our invariant in Claim 9.3.13. Turning to the contribution of outliers, we want to bound $\epsilon \cdot \mathbb{E}_{X\sim B}[w_{t+1}(X) g_t(X)]$. By definition, we have that $g_t(x) \leq \tau_t(x) + C_3 \lambda_t \|\mathbf{M}_t\|_F^2/\epsilon$, and thus we get that the desired expression is bounded from above by $\epsilon\, \mathbb{E}_{X\sim B}[w_{t+1}(X)\tau_t(X)] + C_3 \lambda_t \|\mathbf{M}_t\|_F^2$. The first expression was bounded from above in Lemma 9.3.9 by using the downweighting filter, and the second is small because of how C_3 is set in our algorithm.

This completes the proof sketch of Lemma 9.3.16. We now provide the complete proof.

Proof of Lemma 9.3.16. We will bound the contribution of inliers and outliers to the quantity $\mathbb{E}_{X\sim P}[w_{t+1}(X) g_t(X)]$ from above. Recall from our notation that the decomposition into inliers and outliers is $P = (1-\epsilon)G + \epsilon B$. For the inliers, we use Corollary 9.2.14 with $\mathbf{U} = \mathbf{M}_t$ and $b = \mu_t$ to obtain the following:

$$\mathop{\mathbb{E}}_{X\sim G}[w_{t+1}(X) g_t(X)] \leq \|\mathbf{M}_t\|_F^2 \left(1 + \frac{\delta^2}{\epsilon} + \|\mu_t - \mu\|_2^2 + 2\delta\|\mu_t - \mu\|_2\right) \leq \|\mathbf{M}_t\|_F^2 \left(1 + c\lambda_t\right) \,, \tag{9.5}$$

where the last inequality uses that, by the certificate lemma (Lemma 9.2.11), every term

except the first in the previous expression is less than a sufficiently small fraction of λ_t.

Regarding the outliers, we decompose their contribution to $\mathbb{E}_{X\sim P}[w_{t+1}(X)g_t(X)]$ into two sets: (i) the set of points with projection greater than the threshold $C_3\|\mathbf{M}_t\|_F^2\lambda_t/\epsilon$ used in Line 21 of the algorithm, and (ii) the set of points with smaller projection. Concretely, letting $L_t := \{x \in \mathbb{R}^d \ : \ g_t(x) > C_3\|\mathbf{M}_t\|_F^2\lambda_t/\epsilon\}$, we have that

$$\begin{aligned}\epsilon \mathop{\mathbb{E}}_{X\sim B}[w_{t+1}(X)g_t(X)] &= \epsilon \mathop{\mathbb{E}}_{X\sim B}[w_{t+1}(X)\tau_t(X)] + \epsilon \mathop{\mathbb{E}}_{X\sim B}[w_{t+1}(X)g_t(X)\,\mathbb{I}\{x \notin L_t\}] \\ &\le c\lambda_t\|\mathbf{M}_t\|_F^2 + \epsilon C_3\|\mathbf{M}_t\|_F^2\lambda_t/\epsilon \le c'\|\mathbf{M}_t\|_F^2\lambda_t \ , \end{aligned} \tag{9.6}$$

where the first inequality follows from Lemma 9.3.9 and the second inequality follows from the choice of C_3 in Algorithm 6.

We also use the following relation on $\mathbf{\Sigma}_{t+1}$:

$$\begin{aligned}\mathbf{\Sigma}_{t+1} &= \mathop{\mathbb{E}}_{X\sim P_{t+1}}\left[(X-\mu_{t+1})(X-\mu_{t+1})^\top\right] \\ &\preceq \mathop{\mathbb{E}}_{X\sim P_{t+1}}\left[(X-\mu_t)(X-\mu_t)^\top\right] \\ &= \frac{1}{\mathbb{E}_{X\sim P}[w_{t+1}(X)]}\mathop{\mathbb{E}}_{X\sim P}[w_{t+1}(X)(X-\mu_t)(X-\mu_t)^\top] \ . \end{aligned} \tag{9.7}$$

Recalling the definition $\mathbf{B}_{t+1} = (\mathbb{E}_{X\sim P}[w_{t+1}(X)])^2\mathbf{\Sigma}_{t+1} - \left(1 - C_1\frac{\delta^2}{\epsilon}\right)\mathbf{I}_d$, Equation (9.7) implies that $\mathbf{B}_{t+1} \preceq \mathbf{F}_{t+1}$, where $\mathbf{F}_{t+1} := (\mathbb{E}_{X\sim P}[w_{t+1}(X)])\,\mathbb{E}_{X\sim P}[w_{t+1}(X)(X-\mu_t)(X-\mu_t)^\top] - \left(1 - C_1\frac{\delta^2}{\epsilon}\right)\mathbf{I}_d$. Using Fact 9.2.2 along with the fact that $\mathbf{B}_{t+1} \succeq 0$ (Claim 9.3.14), we get the following:

$$\begin{aligned}\mathrm{tr}(\mathbf{M}_t\mathbf{B}_{t+1}\mathbf{M}_t) &= \mathrm{tr}(\mathbf{M}_t^2\mathbf{B}_{t+1}) \le \mathrm{tr}(\mathbf{M}_t^2\mathbf{F}_{t+1}) \qquad (\text{using } \mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{CAB})) \\ &= \mathrm{tr}\Bigg(\mathbf{M}_t\Bigg(\mathop{\mathbb{E}}_{X\sim P}[w_{t+1}(X)]\mathop{\mathbb{E}}_{X\sim P}[w_{t+1}(X)(X-\mu_t)(X-\mu_t)^\top] \\ &\qquad - \left(1 - C_1\frac{\delta^2}{\epsilon}\right)\mathbf{I}_d\Bigg)\mathbf{M}_t\Bigg) \\ &= \mathop{\mathbb{E}}_{X\sim P}[w_{t+1}(X)]\mathop{\mathbb{E}}_{X\sim P}[w_{t+1}(X)\mathrm{tr}((X-\mu_t)^\top\mathbf{M}_t^2(X-\mu_t))]\end{aligned}$$

$$
\begin{aligned}
&\quad - \left(1 - C_1 \frac{\delta^2}{\epsilon}\right) \|\mathbf{M}_t\|_F^2 \\
&= \mathop{\mathbb{E}}_{X \sim P}[w_{t+1}(X)] \mathop{\mathbb{E}}_{X \sim P}[w_{t+1}(X) g_t(X)] - \left(1 - C_1 \frac{\delta^2}{\epsilon}\right) \|\mathbf{M}_t\|_F^2 \\
&\leq \left(1 + c\lambda_t + c'\lambda_t - (1 - C_1\delta^2/\epsilon)\right) \|\mathbf{M}_t\|_F^2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(using Equations (9.5) and (9.6))} \\
&\leq c''\lambda_t \|\mathbf{M}_t\|_F^2 \,.
\end{aligned}
$$

This concludes the proof. □

9.4 Efficient Streaming Algorithm for Robust Mean Estimation

We now turn to the main focus of this paper and present a low-memory algorithm for robust mean estimation. Our algorithm works in two setups: (i) the single-pass streaming setting, where a set of i.i.d. samples from an ϵ-corrupted distribution in total variation distance (Definition 9.1.2) arrive *one at a time* (Definition 9.1.1), and (ii) the strong-contamination model (Definition 9.3.1), where the algorithm is allowed poly-logarithmically many passes over the input stream (defined below).

Definition 9.4.1 (Streaming Model in k Passes)**.** *For a fixed set S, the elements of S are revealed to the algorithm one at a time. This process is repeated k times. The sequence of elements in S within each pass can be arbitrary.*

Our main result is the following theorem for the single-pass streaming model, which is a generalized version of Theorem 9.1.3:

Theorem 9.4.2 (Robust Mean Estimation in Single-Pass Streaming Model)**.** *Let $d \in \mathbb{Z}_+$, $0 < \tau < 1$, $0 < \epsilon < \epsilon_0$ for a sufficiently small constant ϵ_0, and $\delta \geq \epsilon$. Let D be a distribution*

which is $(C\epsilon, \delta)$-stable with respect to the (unknown) vector $\mu \in \mathbb{R}^d$, for a sufficiently large constant $C > 0$. Let R be any radius such that $\mathbb{P}_{X\sim D}[\|X-\mu\|_2 > R] \leq \epsilon$. Let P be a distribution with $d_{\mathrm{TV}}(P, D) \leq \epsilon$. There exists an algorithm that given ϵ, δ, τ and

$$n = O\left(R^2 \max\left(d, \frac{\epsilon}{\delta^2}, \frac{(1+\delta^2/\epsilon)d}{\delta^2 R^2}, \frac{\epsilon^2 d}{\delta^4}, \frac{R^2\epsilon^2}{\delta^2}, \frac{R^2\epsilon^4}{\delta^6}\right) \mathrm{polylog}\left(d, \frac{1}{\epsilon}, \frac{1}{\tau}, R\right)\right) \tag{9.8}$$

i.i.d. samples from P in a stream according to the model of Definition 9.1.1, runs in time $nd\,\mathrm{polylog}\,(d, 1/\epsilon, 1/\tau, R)$, uses memory $d\,\mathrm{polylog}\,(d, 1/\epsilon, 1/\tau, R)$, and returns a vector $\widehat{\mu}$ such that, with probability at least $1-\tau$, it holds that $\|\mu - \widehat{\mu}\|_2 = O(\delta)$.

Note that Theorem 9.1.3 in Section 9.1.1 is a special case of Theorem 9.4.2 for the two important families of distributions: (i) subgaussian distributions with identity covariance, and (ii) distributions with bounded covariance.

1. For subgaussian distributions with identity covariance, we have that $R = \Theta(\sqrt{d\log(1/\epsilon)})$, $\delta = O(\epsilon\sqrt{\log(1/\epsilon)})$, and thus $n = \tilde{O}\left(d^2/\epsilon^2\right)$.

2. For distributions with covariance at most identity, we have that $R = \Theta(\sqrt{d/\epsilon})$, $\delta = O(\sqrt{\epsilon})$, and thus $n = \tilde{O}\left(d^2/\epsilon\right)$.

In order to obtain a low-memory algorithm for the robust mean estimation problem, we start with an obstacle that one faces when trying to modify the existing Algorithm 6 to that setting. The issue is that, since n can be much larger than d, we cannot even store the weight function w_t. Fortunately, this can be handled by freshly computing the scores $w_t(x)$ for any given x, whenever we need them. This requires us to store only $\{(\mathbf{U}_t, \ell_t) : t \in [K]\}$, where $\mathbf{U}_t$ is the Johnson-Lindenstrauss sketch at the iteration t, and ℓ_t is the corresponding count from the downweighting filter. This can be achieved with additional poly-logarithmic memory. Thus, Algorithm 6 can be readily extended to setting (ii), giving us Corollary 9.4.3.

Corollary 9.4.3 (Robust Mean Estimation in Multiple Passes Streaming Model)**.** *Let $d \in \mathbb{Z}_+$, $0 < \tau < 1$ and $0 < \epsilon < \epsilon_0$ for a sufficiently small constant ϵ_0, and $\delta \geq \epsilon$. Let S be an ϵ-corrupted version of a set that is $(C\epsilon, \delta)$-stable with respect to the (unknown) vector $\mu \in \mathbb{R}^d$, for a sufficiently large constant C. Denote by n the cardinality of S. There exists an algorithm that operates in the streaming model of Definition 9.4.1 with $k = \operatorname{polylog}(d, 1/\epsilon, 1/\tau)$ and, given ϵ, δ, τ and T, runs in time $nd \operatorname{polylog}(d, 1/\epsilon, 1/\tau)$, uses additional memory $d \operatorname{polylog}(d, 1/\epsilon, 1/\tau)$, and finds a vector $\widehat{\mu}$ such that, with probability at least $1 - \tau$, it holds $\|\mu - \widehat{\mu}\|_2 = O(\delta)$.*

In the main body of this section, we prove Theorem 9.4.2.

9.4.1 Setup and Algorithm Description

Moving to the single-pass streaming model and Theorem 9.4.2 requires a change in perspective: instead of having a corrupted dataset, we now have sample access to a distribution P such that $d_{\mathrm{TV}}(P, D) \leq \epsilon$, where D is a stable distribution. We will reweight this distribution using weights, $w_t(\cdot)$, that are now *functions on the whole* $\mathbb{R}^d$ instead of a fixed dataset. Thus, P_t now denotes the reweighting of the (corrupted) distribution P with the weights w_t. Similarly $\mu_t, \mathbf{\Sigma}_t, \mathbf{B}_t, \mathbf{M}_t$ denote the quantities that pertain to the distribution P_t. The goal of our algorithm remains essentially the same: obtain P_t such that $d_{\mathrm{TV}}(P_t, P) = O(\epsilon)$ and $\|\mathbf{\Sigma}_t\|_2 \leq 1 + O(\delta^2/\epsilon)$; Lemma 9.2.11 would then imply that $\|\mu_t - \mu\|_2 = O(\delta)$. Before presenting the pseudocode of Algorithm 8, we identify two problems that arise in generalizing our results from Section 9.3 and provide an overview of their solutions:

Calculating Scores Recall that the only place where $\mathbf{M}_t$ is used in Algorithm 6 is Line 18, where $\mathbf{M}_t$ is multiplied with the vectors $z_{t,j}$. Let z be an arbitrary vector. Since $\mathbf{M}_t = \mathbf{B}^{\log d}$, in the previous section we were able to compute $\mathbf{M}_t z$ by iteratively multiplying z by $\mathbf{B}_t$. Since we now do not have access to $\mathbf{B}_t$, but only sample access to P_t, we need

a sufficiently fine approximation $\widehat{\mathbf{B}}_t$ of $\mathbf{B}_t$ (obtained using i.i.d. samples). The natural approach would then be to multiply $\widehat{\mathbf{B}}_t$ with z iteratively $\log d$ many times. Even though $\widehat{\mathbf{B}}_t z$ can be computed in a streaming fashion (as outlined in the previous section), it is not possible to compute $(\widehat{\mathbf{B}}_t)^{\log d} z$ without accessing the data $\log d$ times. To circumvent this issue, we use a fresh sample approximation of $\mathbf{B}_t$ in every multiplication step. That is, we approximate $\mathbf{M}_t z$ by $\widehat{\mathbf{M}}_t z$, where $\widehat{\mathbf{M}}_t := \prod_{j=1}^{p} \widehat{\mathbf{B}}_{t,j}$ and each $\widehat{\mathbf{B}}_{t,j}$ is computed on a different set of samples. This approach crucially leverages the fact that in the contamination model of Definition 9.1.2, outliers are added in a way that is oblivious to the inliers, and therefore these datasets are statistically identical and independent of each other. We show in Section 9.4.3 that the resulting $\widehat{\mathbf{M}}_t$ is a sufficiently accurate approximation of $\mathbf{M}_t$. Similarly, we need to modify the Downweighting filter, since its implementation using binary search requires performing checks of the form $\mathbb{E}_{X\sim P}[w(X)\tilde{\tau}(x)] > 2T$ and calculating the weighted mean exactly is no longer possible. We propose a sample-efficient estimator to approximate that expectation (see Lemma 9.4.16 in Section 9.4.3.3) and run an "approximate" variant of binary search (see Section 9.4.2).

Cover Argument We now turn to the more technical issue of controlling the size of the JL-sketch, i.e., the number of rows, L, of the matrix $\mathbf{U}_t \in \mathbb{R}^{L\times d}$. For simplicity, assume $\widehat{\mathbf{M}}_t = \mathbf{M}_t$, and recall that $\tilde{\tau}(x)$ is the thresholded version of $\|\mathbf{U}_t(x-\mu_t)\|_2^2$, as defined in Line 21 and $\tau(x)$ is the same score but using $\mathbf{M}_t$. The potential-based analysis in Section 9.3 requires that $\mathbb{E}_{X\sim P}[w_{t+1}(X)\tau_t(X)]$ is small. However, the stopping condition of the Downweighting filter implies only that $\mathbb{E}_{X\sim P}[w_{t+1}(X)\tilde{\tau}_t(X)]$ is small. In Section 9.3, the bound on the former was obtained from the bound on the latter by using that $\|\mathbf{U}_t(x-\mu_t)\|_2 \approx \|\mathbf{M}_t(x-\mu_t)\|_2$ pointwise in the support of P (Claim 9.3.10).

By the classical JL lemma, the size of the JL sketch, L, needs to be at most logarithmic in the size of the set S where we require the pointwise approximation to hold. Thus, in the previous section, L scaled as $\log|S| = \log n$. However, in the streaming model where

there is no such dataset, it is far from obvious how the analysis should proceed. A näive approach would be to require the approximation to hold on a cover $\tilde{S}$ of the support of P_t. Since $|\tilde{S}|$ scales exponentially with d, the required bound on L would be $\log|\tilde{S}| = \Omega(d)$, which is too large for our purposes. Luckily, we can still find a fixed set S_{cover} such that the following holds: (i) $\log|S_{\text{cover}}| = \text{polylog}(d/\epsilon)$, and (ii) the expectation of scores over $\mathcal{U}(S_{\text{cover}})$ approximates the expectation of scores over P. That is, as far as the expectations of the scores are concerned, P can be approximated by the uniform distribution over S_{cover}. Arguing as before, if $\|\mathbf{U}_t(x-\mu_t)\| \approx \|\mathbf{M}_t(x-\mu_t)\|$ for each $x \in S_{\text{cover}}$, then the downweighting filter also ensures that $\mathbb{E}_{X\sim P}[w_{t+1}(X)\tilde{\tau}_t(X)]$ is small. Thus, S_{cover} can serve as a proxy dataset (used only in the analysis) to ensure that the size of the JL sketch is sufficiently bounded, i.e., that $\log|S_{\text{cover}}| \le C\text{polylog}(d/\epsilon)$.

Establishing the desired upper bound on the cardinality of S_{cover} requires a somewhat more sophisticated argument that relies on the VC-dimension of a family of functions corresponding to the weight update rule. This result is stated in Section 9.4.2.1.

We now present the algorithm more formally. We start by clarifying the notation used.

Notation regarding Algorithm 8: The quantities involved in the algorithm and its analysis now are based on the underlying data distribution P as well as its approximations. We note that $P_t, \mu_t, \mathbf{\Sigma}_t, \mathbf{B}_t, \mathbf{M}_t, \lambda_t$ are functionals of the distribution P and are primarily used in the analysis. The parameters $\widehat{\lambda}_t, \widehat{\mathbf{M}_t}$ are approximations for $\|\mathbf{B}_t\|_2$ and $\mathbf{M}_t$ respectively that the algorithm forms using samples from P_t. Regarding score functions, $g_t(x) = \|\mathbf{M}_t(x-\mu_t)\|_2^2$ is as before. The computations however use only the Johnson-Lindenstrauss versions $\tilde{g}_t(x) := \frac{1}{L}\sum_{i=1}^{L}(v_{t,i}^\top(x-\widehat{\mu}_t))^2$, which can also be written as $\|\mathbf{U}_t(x-\widehat{\mu}_t)\|_2^2$ in matrix form, by defining $\mathbf{U}_t$ to have the vectors $\frac{1}{\sqrt{L}}v_{t,i}$ as its rows. Note that $\tilde{g}_t(x)$ is defined using $\widehat{\mu}_t$ instead of μ_t. Finally, we denote by $\tau_t(x) = g_t(x)\,\mathbb{I}\{g_t(x) > C_3\|\mathbf{M}_t\|_F^2\lambda_t/\epsilon\}$ and $\tilde{\tau}_t(x) = \tilde{g}_t(x)\,\mathbb{I}\{\tilde{g}_t(x) > C_3\|\mathbf{U}_t\|_F^2\widehat{\lambda}_t/\epsilon\}$.

Remark 9.4.4. *Recalling Lemma 9.2.12, we may again treat the input distribution as a mixture* $P = (1-\epsilon)G + \epsilon B$*, where* G *is a distribution that is* $(C'\epsilon, \delta)$*-stable with respect to* μ*.*

As already mentioned, Algorithm 8 uses two levels of approximation: the first level is approximating the true distributional quantities by taking samples, and the second is preserving the latter quantities using the JL sketch. If both of these approximations are sufficiently accurate, the correctness of Algorithm 8 would follow similarly to Algorithm 6. Of course, the challenge is to ensure that these approximations hold over the entire distribution, while controlling the sample and memory complexity of the algorithm. As we show in Section 9.4.2.1, this can be achieved by restricting our attention to a finite set (cover) of sufficiently large cardinality. Thus, the deterministic conditions that we require now also involve the cover set, which we denote by S_{cover}. The reader may think of S_{cover} as a fixed set, which will be specified later on (Lemma 9.4.9).

Algorithm 8 Robust Mean Estimation In Single-Pass Streaming Model

function RobustMeanStreaming(δ, ϵ, τ and sample access to P)
  Let R be such that $\mathbb{P}_{X\sim G}[\|X-\mu\|_2 > R] \le \epsilon$.
  Let $P = (1-\epsilon)G+\epsilon B$. Without loss of generality, we assume that the points added by the adversary are within $O(R)$ from μ in Euclidean norm (see Section 9.3.1).
  Let C be a sufficiently large constant.
  Let $K = C\log d\log(dR/\epsilon)$.
  Let $L = C\log^3(dR/\epsilon)\log^2(1/(\tau\epsilon))$.
  Let $r = CdR^{2+4\log d}$.
  Obtain a naïve estimation $\widehat{\mu}$ of μ such that $\|\widehat{\mu}-\mu\|_2 \le 4R$.
  Let $w : \mathbb{R}^d \to [0,1]$ be the weight function.
  Initialize $w_0(x) \leftarrow \mathbb{I}\{\|x-\widehat{\mu}\|_2 \le 5R\}$ for all $x \in T$ and $\ell_1 \leftarrow 0$.
  for $t \in [K]$ **do**
    Define $w_t(x) = w_{t-1}(x)(1-\tilde{\tau}_t(x)/r)^{\ell_t}$.
    Let P_t be the distribution of P weighted by w_t, i.e., $P_t(x) = P(x)w_t(x)/\mathbb{E}_{X\sim P}[w_t(X)]$.
    Let μ_t be the mean of P_t.
    Let $\mathbf{\Sigma}_t$ be the covariance matrix of P_t.
    Let $\mathbf{B}_t = (\mathbb{E}_{X\sim P}[w_t(X)])^2\mathbf{\Sigma}_t - \left(1 - C_1\frac{\delta^2}{\epsilon}\right)\mathbf{I}_d$ and $\mathbf{M}_t = \mathbf{B}_t^{\log d}$.
    Compute an $O(\delta)$-accurate estimator $\widehat{\mu}_t$ of μ_t (see Lemma 9.4.11).
    Let $\tilde{n} = C''R^2(\log d)^2\max\left(d, \frac{\epsilon^2 d}{\delta^4}, \frac{R^2\epsilon^2}{\delta^2}, \frac{R^2\epsilon^4}{\delta^6}\right)\log\left(\frac{dK\log d}{\tau}\right)$.
    For $k \in [\log d]$, denote by $\widehat{\mathbf{B}}_{t,k}$ the empirical version of $\mathbf{B}_t$ over $\tilde{n}$ fresh i.i.d. samples (see Section 9.4.3 for more details).
    Define $\widehat{\mathbf{M}}_t := \prod_{k=1}^{\log d}\widehat{\mathbf{B}}_{t,k}$ ▷ $\widehat{\mathbf{M}}_t$ is not stored in memory.
    Let $\lambda_t = \|\mathbf{B}_t\|_2$ and an approximation $\widehat{\lambda}_t$ such that $\widehat{\lambda}_t/\lambda_t \in [0.8, 1.2]$.
    if $\widehat{\lambda}_t > C_2\delta^2/\epsilon$ **then**
      for $j \in [L]$ **do**
        $z_{t,j} \sim \mathcal{U}(\{\pm 1\}^d)$.
        $v_{t,j} \leftarrow \widehat{\mathbf{M}}_t z_{t,j}$. ▷ See Remark 9.4.12 for efficient implementation.
        Store $v_{t,j}$ in memory.
      end for
      Denote by $\mathbf{U}_t$ the matrix with rows $\frac{1}{\sqrt{L}}v_{t,j}$ for $j \in [L]$.
      Let $\tilde{g}_t(x) = \|\mathbf{U}_t(x-\widehat{\mu}_t)\|_2^2$ and $\tilde{\tau}_t(x) = \tilde{g}_t(x)\,\mathbb{I}\{\tilde{g}_t(x) > C_3\|\mathbf{U}_t\|_F^2\widehat{\lambda}_t/\epsilon\}$.
      $\ell_{\max} \leftarrow \left(\frac{dR}{\delta^2/\epsilon}\right)^{C\log d}$.
      $\ell_t \leftarrow \mathrm{DownweightingFilter}(P, w_t, \tilde{\tau}_t, R, c\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2, \ell_{\max})$. ▷ Algorithm 9
      Store ℓ_t in memory.
    end if
  end for
  return an $O(\delta)$ approximation $\widehat{\mu}_t$ of the mean μ_t of the distribution P_t (see Lemma 9.4.11).
end function

Condition 9.4.5 (Deterministic Conditions for Algorithm 8)**.** *Let S_{cover} denote the cover of Lemma 9.4.9 for $\epsilon' = \mathrm{poly}(d, R, 1/\epsilon)^{\log d}$. Our condition consists of the following event:*

1. Estimator $\widehat{\mu}_t$*: For all $t \in [K]$, we have that $\|\widehat{\mu}_t - \mu_t\|_2 \le \delta/100$.*
2. *For every $t \in [K]$, if $\|\mathbf{B}_t\|_2 \ge (C_1/2)\delta^2/\epsilon$ and $\mathbb{E}_{X\sim P}[w_t(X)] \ge 1 - O(\epsilon)$, we have that:*
   a) Spectral norm of $\mathbf{B}_t$*: $\widehat{\lambda}_t \in [0.1\lambda_t, 20\lambda_t]$.*
   b) Frobenius norm*: $\|\mathbf{U}_t\|_F^2 \in [0.8\|\mathbf{M}_t\|_F^2, 1.2\|\mathbf{M}_t\|_F^2]$.*
   c) Scores*: $\tilde{g}_t(x) \ge 0.2g_t(x) - 0.8(\delta^2/\epsilon^2)\|\mathbf{M}_t\|_F^2$, for all $x \in S_{cover}$.*
3. Stopping condition*: Let $T_t := c\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2$. For every $w : \mathbb{R}^d \to [0,1]$, the algorithm has access to an estimator $f(w)$ for the quantity $\mathbb{E}_{X\sim P}[w(X)\tilde{\tau}_t(X)]$, such that $\widehat{F}(P) > T_t/2$ whenever $\mathbb{E}_{X\sim P}[w(X)\tilde{\tau}_t(X)] > T_t$. This estimator is accurate when called $O(\log(d)\log(dR/\epsilon))$ times in every iteration $t \in [K]$.*

We note that the Item 3 above is needed to evaluate the stopping condition in the downweighting filter. For every $t \in [K]$, the stopping condition is evaluated at most $O(\log(\ell_{\max}))$ times, with $\ell_{\max} = O(dR^{2+\log d}/(\widehat{\lambda}_t\|\mathbf{U}_t\|^2)$ (using Lemma 9.3.6 with $r = C_4 dR^{2+4\log d}$ and $T := O(\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2)$). This means that we require the estimator in Item 3 to be accurate on $O(K\log(d)\log(dR/\epsilon))$ calls.

9.4.2 Correctness of Algorithm 8

The analysis in this section is along the same lines as that of Section 9.3.4. The naïve estimation step of Line 8 is the same as that used in Algorithm 6 (see Appendix F.2.4).

Given Condition 9.4.5, we first show the correctness of Algorithm 8 and leave the task of establishing Condition 9.4.5 for Section 9.4.3. The proof of correctness would largely follow by our work done in Section 9.3.4. There are two adjustments needed in these arguments.

The first concerns Lemma 9.3.6, since the algorithm cannot perform exact binary search. Instead, it can use the approximate oracle of Item 3 of Condition 9.4.5, resulting in a multiplicative constant in the final guarantee. For completeness, we prove correctness for this case in Appendix F.3.1.

Lemma 9.4.6. *In the context of Algorithm 8, if* $(1-\epsilon)\,\mathbb{E}_{X\sim G}[w(X)\tilde{\tau}(X)] \le T$, $\|\tilde{\tau}\|_\infty \le r$, *and* $\ell_{\max} > r/T$, *then Algorithm 9 modifies the weight function* w *to* w' *such that* (i) $(1-\epsilon)\,\mathbb{E}_{X\sim G}[w(X)-w'(X)] < \epsilon\,\mathbb{E}_{X\sim B}[w(X)-w'(X)]$, *and* (ii) *upon termination we have* $\mathbb{E}_{X\sim P}[w'(X)\tilde{\tau}(X)] \le 54T$. *Furthermore, if the estimator of Line 3 is set to be that of Lemma 9.4.16, the algorithm terminates after* $O(\log(\ell_{\max}))$ *iterations, each of which uses* $O((R^2\epsilon/\delta^2)\log(1/\tau))$ *samples, takes* $O(nd)$ *time and memory* $O(\log(1/\tau))$.

Algorithm 9 Downweighting Filter using Approximate Oracle

1: **function** DownweightingFilter$(P, w, \tilde{\tau}, R, T, \ell_{\max})$
2: $r \leftarrow CdR^{2+4\log d}$.
3: Denote by $f(\ell)$ an estimator close to $\mathbb{E}_{X\sim P}[w(X)(1-\tilde{\tau}(X)/r)^\ell\tilde{\tau}(X)]$ (see Lemma 9.4.16 for details).
4: $L \leftarrow \{1, 2\ldots, \ell_{\max}\}$
5: **while** $|L| > 2$ **do**
6: Let ℓ be the element in the middle of L.
7: **if** $f(\ell) > 9T$ **then**
8: Discard all elements smaller than ℓ from L.
9: **else**
10: Discard all elements greater than ℓ from L.
11: **end if**
12: **end while**
13: **return** any ℓ of L satisfying $4T \le f(\ell) \le 36T$.
14: **end function**

The second adjustment is regarding the analog of Lemma 9.3.9, i.e., $\epsilon\,\mathbb{E}_{X\sim B}[w_{t+1}\tau_t(X)]$ is small (the bound on $\epsilon\,\mathbb{E}_{X\sim B}[w_{t+1}\tilde{\tau}_t(X)]$ follows from the stopping condition as before). Since the support is unbounded, we use an argument based on a fixed cover to show that the downweighting filter succeeds with the JL-sketch of size L. The statement is given below.

Lemma 9.4.7. *Under the deterministic Condition 9.4.5 and the context of Algorithm 8, we have that* $\mathbb{E}_{X\sim B}[w_{t+1}(X)\tau_t(X)] \le 5\,\mathbb{E}_{X\sim B}[w_{t+1}(X)\tilde{\tau}_t(X)] + c(\lambda_t/\epsilon)\|\mathbf{M}_t\|_F^2$*, where* c *is of the form* C/C_2 *with* C *being a sufficiently large constant and* C_2 *being the constant used in Algorithm 8.*

The next section is dedicated to proving Lemma 9.4.7. Here we just show that Lemma 9.4.7 suffices to prove the analog of Lemma 9.3.9 below.

Lemma 9.4.8. *Consider the context of Algorithm 8 and assume that Condition 9.4.5 holds. Then* $\epsilon \cdot \mathbb{E}_{X\sim B}[w_{t+1}(X)\tau_t(X)] \le c'\lambda_t\|\mathbf{M}_t\|_F^2$*, for some constant* c'' *of the form* C''/C_2*, where* C_2 *is the constant used in Line 22 and* C' *is some absolute constant.*

Proof. Denoting by $c, c', c'' > 0$ constants that are all multiples of $1/C_2$, we have the following:

$$\begin{aligned}\epsilon \mathop{\mathbb{E}}_{X\sim B}[w_{t+1}(X)\tau_t(x)] &\le 5\epsilon \mathop{\mathbb{E}}_{X\sim B}[w_{t+1}(X)\tilde{\tau}_t(x)] + c\lambda_t\|\mathbf{M}_t\|_F^2 \\ &\le 5c'\lambda_t\|\mathbf{M}_t\|_F^2 + c\lambda_t\|\mathbf{M}_t\|_F^2 \le c''\lambda_t\|\mathbf{M}_t\|_F^2\,,\end{aligned}$$

where the first inequality uses Lemma 9.4.7 and the second inequality uses Lemma 9.4.6. □

Letting $\phi_t := \mathrm{tr}(\mathbf{M}_t^2)$ denote the potential function, the above result allows us to follow the same steps as in Section 9.3.4.2 to prove that $\phi_{t+1} \le 0.9999\phi_t$ exactly as in Section 9.3.4.2. Thus, we get that after K iterations, we have that $\|\mathbf{\Sigma}_t\|_2 \lesssim \delta^2/\epsilon$. Under Item 1 of Condition 9.4.5, we have that the final estimate $\widehat{\mu}_t$ satisfies that $\|\widehat{\mu}_t - \mu\|_2 = O(\delta)$. This completes the proof of correctness of Algorithm 8.

9.4.2.1 Proof of Lemma 9.4.7 via a Cover Argument

To outline the idea of proving Lemma 9.4.7, recall the proof in the setting of Section 9.3. There, we just required that $\tilde{g}_t(x)/g_t(x) \in [0.8, 1.2]$ for all samples x in our dataset, which

can be translated to some relation between $\tilde{\tau}(x)$ and $\tau(x)$. Then, since B was the empirical distribution on ϵn of these points, the desired condition followed. However, in our case we cannot use pointwise relationships, since the distribution B may be continuous and the Johnson-Lindenstrauss argument might not work for the entire $\mathbb{R}^d$ with $\mathrm{polylog}(d)$ vectors. The idea is first to relate $\mathbb{E}_{X\sim B}[w_{t+1}(X)\tau_t(X)]$ to a discrete expectation over N (not too many) points from a fixed set, then use the relationship between $\tilde{\tau}$ and τ for these points, and finally relate that discrete expectation back to $\mathbb{E}_{X\sim B}[w_{t+1}(X)\tilde{\tau}_t(X)]$. The existence of a cover of a small size is stated in the following.

Lemma 9.4.9. *Consider the setting of Algorithm 8, where B is the distribution of outliers supported in a ball of radius R around μ. Let $r' := (CdR^2 + 1 + C_1\delta^2/\epsilon)^{C\log d}$ for sufficiently large constant C. Denote by ϵ the contamination rate and let an arbitrary $\epsilon' \in (0,1)$. There exists a set S_{cover} of $N = \frac{1}{\epsilon'^3}d^4K^2L^2(dR\epsilon/\delta^2)^{O(\log d)}$ points $x_1,\dots,x_N$ lying in the ball of radius R around μ, such that for all $t\in[K]$, for all choices of the vectors $z_{t,j}$ of Line 24 of Algorithm 8 it holds*

$$\left|\mathop{\mathbb{E}}_{X\sim B}\left[\frac{1}{r'}w_{t+1}(X)\tilde{\tau}_t(X)\right] - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r'}w_{t+1}(x_i)\tilde{\tau}_t(x_i)\right| \le \epsilon'$$

$$\text{and}\quad \left|\mathop{\mathbb{E}}_{X\sim B}\left[\frac{1}{r'}w_{t+1}(X)\tau_t(X)\right] - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r'}w_{t+1}(x_i)\tau_t(x_i)\right| \le \epsilon' .$$

We prove this result in Appendix F.3.2. Here we show how it implies the desired condition, following the proof sketch from the start of this section.

Proof of Lemma 9.4.7. Let $r' := (CdR^2 + 1 + C_1\delta^2/\epsilon)^{C\log d}$ and $\epsilon' \in (0,1)$. Applying Lemma 9.4.9, let S_{cover} be the corresponding cover of cardinality N. From the guarantee of approximation of $\tilde{g}_t$ for every $x\in S_{\text{cover}}$, we get the following approximation for $\tilde{\tau}_t(x)$ for $x\in S_{\text{cover}}$ (proved in Appendix F.3.2).

Claim 9.4.10. *Let S be the cover of Lemma 9.4.9 with r' and ϵ' as defined above. Suppose that the deterministic condition Condition 9.4.5 holds. If $x \in S_{cover}$, then $\tau_t(x) \leq 5\tilde{\tau}_t(x) + (18C_3 + 12/C_2)(\lambda_t/\epsilon)\|\mathbf{M}_t\|_F^2$, where C_3 and C_2 are the constants used in Algorithm 8.*

Using Claim 9.4.10 and Lemma 9.4.9, we obtain the following series of inequalities:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim B}[w_{t+1}(X)\tau_t(X)] = r' \mathop{\mathbb{E}}_{X\sim B}\left[\frac{1}{r'}w_{t+1}(X)\tau_t(X)\right] \\
&\leq \epsilon' r' + r'\frac{1}{N}\sum_{i=1}^{N}\frac{1}{r'}w_{t+1}(x_i)\tau_t(x_i) && \text{(using Lemma 9.4.9 for } \tau_t\text{)} \\
&= \epsilon' r' + \frac{1}{N}\sum_{i=1}^{N} w_{t+1}(x_i)\tau_t(x_i) \\
&\leq \epsilon' r' + (18C_3 + 12/C_2)(\lambda_t/\epsilon)\|\mathbf{M}_t\|_F^2 + 5\frac{1}{N}\sum_{i=1}^{N} w_{t+1}(x_i)\tilde{\tau}_t(x_i) \\
& && \text{(using Claim 9.4.10 and } w_t \leq 1\text{)} \\
&\leq 6\epsilon' r' + (18C_3 + 12/C_2)(\lambda_t/\epsilon)\|\mathbf{M}_t\|_F^2 + 5\mathop{\mathbb{E}}_{X\sim B}[w_{t+1}(X)\tilde{\tau}_t(X)] \\
& && \text{(using Lemma 9.4.9 for } \tilde{\tau}_t\text{)} \\
&= 5\mathop{\mathbb{E}}_{X\sim B}[w_{t+1}(X)\tilde{\tau}_t(X)] + (19C_3 + 12/C_2)(\lambda_t/\epsilon)\|\mathbf{M}_t\|_F^2. && \text{(using the definition of } \epsilon'\text{)}
\end{aligned}
$$

For the last line above, we want to choose ϵ' such that $\epsilon' \leq \frac{C_3\lambda_t}{\epsilon r'}\|\mathbf{M}_t\|_F^2$. Since $\|\mathbf{M}_t\|_F^2 \geq (C_2\delta^2/\epsilon)^{2\log d}$ (otherwise the algorithm has already terminated), it suffices to choose an ϵ' that satisfies $\epsilon' \gtrsim \frac{(C_2\delta^2/\epsilon)^{2\log d}}{\epsilon(CdR^2+1+C_1\delta^2/\epsilon)^{C\log d}}$. This gives an upper bound on the cardinality of the set S_{cover}, which gives the upper bound on the size of the JL-sketch, i.e., L. We provide explicit calculations in Remark F.3.3. □

9.4.3 Establishing Condition 9.4.5

Throughout this section, we assume sample access to the distribution P_t. As mentioned earlier, Algorithm 7 can simulate this by drawing a sample x from P, calculating $w_t(x)$ (with poly-logarithmic cost in terms of running time and memory), and rejecting the

sample with probability $1-w_t(x)$. With high probability, rejection sampling can increase the sample complexity by at most a constant factor because $\mathbb{E}[w_t(X)] \geq 1 - O(\epsilon)$ (cf. Claim 9.3.13).

9.4.3.1 Item 1

We establish Item 1 in the following, which is proved in Appendix F.3.4.

Lemma 9.4.11. *In the setting of Algorithm 8, there exist estimators $\widehat{\mu}_t$ such that, with probability at least $1-\tau$, for all $t \in [K]$ we have that $\|\widehat{\mu}_t - \mu_t\|_2 \leq \delta/100$. Furthermore, each $\widehat{\mu}_t$ can be computed on a stream of $n = O\left(\frac{R^2}{\delta^2/\epsilon}\log(K/\tau) + \frac{d(1+\delta^2/\epsilon)}{\delta^2}\log(K/\tau)\right)$ independent samples from P_t, in time $O(nd\log(K/\tau))$ and using memory $O(d\log(K/\tau))$.*

9.4.3.2 Items 2a to 2c

Given that Item 1 holds, in this section, we show that Items 2a to 2c of Condition 9.4.5 hold with high probability if sufficiently many samples from the underlying corrupted distribution P are drawn. Let the scores $\widehat{g}_t(x) := \|\widehat{\mathbf{M}}_t(x-\mu_t)\|_2^2$, where $\widehat{\mathbf{M}}_t$ is the following sample-based estimator of $\mathbf{M}_t$:

1. Draw a batch S_0 of $\tilde{n}$ samples from P and let the estimate $\widehat{W}_t = \mathbb{E}_{X\sim\mathcal{U}(S_0)}[w_t(X)]$.
2. Let P_t' be the distribution of the differences $(X-X')/\sqrt{2}$ for two independent $X, X' \sim P_t$.
3. Draw $\log d$ batches $S_1, \ldots, S_{\log d}$ of $\tilde{n}$ samples, each from P_t'.
4. For $k \in [\log d]$,
   a) Let $\widehat{\mathbf{\Sigma}}_{t,k} = \frac{1}{\tilde{n}}\sum_{x\in S_k} xx^\top$.
   b) Let $\widehat{\mathbf{B}}_{t,k} = \widehat{W}_t^2\widehat{\mathbf{\Sigma}}_{t,k} - (1 - C_1\delta^2/\epsilon)\mathbf{I}_d$.
5. Return $\widehat{\mathbf{M}}_t = \prod_{k=1}^{\log d}\widehat{\mathbf{B}}_{t,k}$.

Remark 9.4.12. *Algorithm 8 does not need to calculate or store $\widehat{\mathbf{M}}_t$ because it requires only that we can calculate products of $\widehat{\mathbf{M}}_t$ with vectors z as in Line 25. This operation can be implemented in linear runtime and memory. Given the description of the estimator above, it suffices to show how to multiply $\widehat{\mathbf{\Sigma}}_{t,k}$ by a vector z in linear time and memory. To this end, we observe that $\widehat{\mathbf{\Sigma}}_{t,k} z = \frac{1}{\tilde{n}} \sum_{x \in S_k} x(x^\top z)$, thus by calculating the inner product $(x^\top z)$ first, the result can be found in $O(nd)$ time in a streaming fashion.*

We will show that Items 2a to 2c of Condition 9.4.5 follow if $\|\widehat{\mathbf{M}}_t - \mathbf{M}_t\|_2 \leq 0.01 \min\left(\frac{\delta/\epsilon}{R}, \frac{1}{\sqrt{d}}\right) \|\mathbf{M}_t\|_F$ (cf. Lemma 9.4.15) and $\|\widehat{\mu}_t - \mu_t\|_2 = O(\delta)$ (cf. Lemma 9.4.11).

Lemma 9.4.13. *Suppose that the estimators $\widehat{\mathbf{M}}_t$ in Algorithm 8 are defined by the procedure as in Lines 1 to 5 above. Let C be a sufficiently large constant and assume that the dimension is $d = \Omega(1)$[13]. Further assume that $\|\widehat{\mu}_t - \mu\| \leq 0.01\delta$. If $\tilde{n} \geq CR^2 (\log d)^2 \max\left(d, \frac{\epsilon^2 d}{\delta^4}, \frac{R^2\epsilon^2}{\delta^2}, \frac{R^2\epsilon^4}{\delta^6}\right) \log\left(\frac{Kd\log d}{\tau}\right)$, then Items 2a to 2c hold with probability at least $1 - \tau$.*

Proof. For now, we will assume that for all $t \in [K]$, $\|\widehat{\mathbf{M}}_t - \mathbf{M}_t\|_2 \leq 0.01 \min\left(\frac{\delta/\epsilon}{R}, \frac{1}{\sqrt{d}}\right) \|\mathbf{M}_t\|_F$ with the claimed sample complexity. This follows from Lemma 9.4.15 and a union bound. We will prove each of these conditions separately.

Proof of Item 2c: Denote $T := 0.1\delta/\epsilon$. Fix an iteration $t \in [K]$. We will prove that the conditions hold in the t-th iteration with probability at least $1 - \tau/K$ and then a union bound will conclude the proof.

Define $\widehat{g}_t(x) := \|\mathbf{M}_t(x - \widehat{\mu}_t)\|_2^2$. Since $\|\widehat{\mu}_t - \mu_t\|_2 \leq 0.01\delta$, we have the following relation between $\widehat{g}_t$ and g_t:$\|\mathbf{M}_t(x - \widehat{\mu}_t)\|_2^2 \geq 0.5\|\mathbf{M}_t(x - \mu_t)\|_2^2 - T^2\|\mathbf{M}_t\|_F^2$, i.e., $\widehat{g}_t(x) \geq 0.5 g_t(x) - T^2\|\mathbf{M}_t\|_F^2$. We have that $\|\widehat{\mathbf{M}}_t - \mathbf{M}_t\|_2 \leq 0.1(T/R)\|\mathbf{M}_t\|_F$ with probability at least $1 - \tau/K$ (see Lemma 9.4.15). This means that for any point x with $\|x - \widehat{\mu}_t\|_2 \leq 2R$, we have $\|\mathbf{M}_t(x - \widehat{\mu}_t)\|_2 \leq \|\widehat{\mathbf{M}}_t(x - \widehat{\mu}_t)\|_2 + 0.2T\|\mathbf{M}_t\|_F$, which implies that $\|\mathbf{M}_t(x - \widehat{\mu}_t)\|_2^2 \leq$

[13]This is without loss of generality as we could avoid the JL-sketch when $d = O(1)$.

$2\|\widehat{\mathbf{M}}_t(x-\widehat{\mu}_t)\|_2^2 + 0.08T^2\|\mathbf{M}_t\|_F^2$, or equivalently

$$\|\widehat{\mathbf{M}}_t(x-\widehat{\mu}_t)\|_2^2 \geq 0.5\widehat{g}_t(x) - 0.04T^2\mathrm{tr}(\mathbf{M}_t^2)\ . \tag{9.9}$$

The final step is taking the Johnson-Lindenstrauss sketch of $\widehat{\mathbf{M}}_t$, which gives the matrix $\mathbf{U}_t$ used in the definition of $\tilde{g}_t$. By repeating the proof of Lemma 9.3.5 with $\widehat{\mathbf{M}}_t$ in place of $\mathbf{M}_t$, we get that if $L = C\log((|S_{\text{cover}}|+d)K/\tau)$, then $\tilde{g}_t(x) \geq 0.8\|\widehat{\mathbf{M}}_t(x-\widehat{\mu}_t)\|_2$ for all the points in the set S_{cover} (the cover from Lemma 9.4.9). The value used for L in Algorithm 8 satisfies this condition (c.f. Remark F.3.3). Combining this with Equation (9.9) and the relation between $\widehat{g}_t$ and g_t, we get that $\tilde{g}_t(x) \geq 0.4\widehat{g}_t(x) - 0.04T^2\|\mathbf{M}_t\|_F^2 \geq 0.2g_t(x) - 0.44T^2\|\mathbf{M}_t\|_F^2$.

Proof of Item 2b: Again, fix a $t \in [K]$. We have that $\|\widehat{\mathbf{M}}_t - \mathbf{M}_t\|_2 \leq \frac{0.01}{\sqrt{d}}\|\mathbf{M}_t\|_F$ with probability $1-\tau/K$ using Lemma 9.4.15. We thus have that $\|\mathbf{M}_t - \widehat{\mathbf{M}}_t\|_F \leq \sqrt{d}\|\mathbf{M}_t - \widehat{\mathbf{M}}_t\|_2 \leq 0.01\|\mathbf{M}_t\|_F$, which implies

$$\left|\|\mathbf{M}_t\|_F - \|\widehat{\mathbf{M}}_t\|_F\right| \leq 0.01\|\mathbf{M}_t\|_F\ . \tag{9.10}$$

It is easy to check that this is stronger than what we initially wanted. Indeed, squaring Equation (9.10) and using $\|\widehat{\mathbf{M}}_t\|_F \leq 1.01\|\mathbf{M}_t\|_F$ gives

$$\|\widehat{\mathbf{M}}_t\|_F^2 \leq 2\|\mathbf{M}_t\|_F\|\widehat{\mathbf{M}}_t\|_F - \|\mathbf{M}_t\|_F^2 + (0.01)^2\|\mathbf{M}_t\|_F^2 < 1.1\|\mathbf{M}_t\|_F^2\ ,$$

which means that $\|\widehat{\mathbf{M}}_t\|_F^2 - \|\mathbf{M}_t\|_F^2 \leq 0.1\|\mathbf{M}_t\|_F^2$. For the other bound, Equation (9.10) implies

$$\begin{aligned}\|\mathbf{M}_t\|_F^2 &\leq 2\|\mathbf{M}_t\|_F\|\widehat{\mathbf{M}}_t\|_F - \|\widehat{\mathbf{M}}_t\|_F^2 + (0.01)^2\|\mathbf{M}_t\|_F^2 \\ &\leq 2\|\widehat{\mathbf{M}}_t\|_F^2 + 2(0.01)\|\mathbf{M}_t\|_F\|\widehat{\mathbf{M}}_t\|_F - \|\widehat{\mathbf{M}}_t\|_F^2 + (0.01)^2\|\mathbf{M}_t\|_F^2\end{aligned}$$

$$< 1.1\|\mathbf{M}_t\|_F^2 \ ,$$

which means that $\|\mathbf{M}_t\|_F^2 - \|\widehat{\mathbf{M}}_t\|_F^2 \leq 0.1\|\mathbf{M}_t\|_F^2$. Therefore we obtain the following

$$\left|\|\mathbf{M}_t\|_F^2 - \|\widehat{\mathbf{M}}_t\|_F^2\right| \leq 0.1\|\mathbf{M}_t\|_F^2 \ . \tag{9.11}$$

Finally, the Johnson-Lindenstrauss step is exactly as described in the proof of Item 2c.

Proof of Item 2a: Since we cannot access the same samples twice, the power-iteration algorithm now uses a different dataset in every step. Let the matrix $\widehat{\mathbf{M}}_t$ as in the beginning of Section 9.4.3. We have already shown in the previous paragraph that, with probability $1-\tau$, for all $t \in [K]$, $\widehat{\mathbf{M}}_t$ has Frobenius norm close to that of $\mathbf{M}_t$ (Equation (9.11)). For the rest of the proof, we condition on this event. Consider the algorithm that calculates $v = \widehat{\mathbf{M}}_t w$, where $w \sim \mathcal{N}(0, \mathbf{I}_d)$ (this can be done in the streaming model by multiplying with $\widehat{\mathbf{B}}_{t,k}$ iteratively; moreover this multiplication can be implemented in time $O(\tilde{n}d)$). We claim that the value $\widehat{\lambda}_t = \|v\|_2^{1/\log d}$ satisfies the desired relation. First, we note that with non-zero constant probability, we have that

$$0.2\mathrm{tr}(\widehat{\mathbf{M}}_t^2) \leq \|v\|_2^2 \leq 10\mathrm{tr}(\widehat{\mathbf{M}}_t^2) \ . \tag{9.12}$$

where one direction follows by Markov's inequality and the other by Fact 9.2.6.

Equations (9.11) and (9.12) imply that $0.1\mathrm{tr}(\mathbf{M}_t^2) \leq \|v\|_2^2 \leq 11\mathrm{tr}(\mathbf{M}_t^2)$. Furthermore, we have that

$$\|v\|_2^{\frac{1}{\log d}} \leq (11\mathrm{tr}(\mathbf{M}_t^2))^{\frac{1}{2\log d}} \leq \left(11d\|\mathbf{B}_t\|_2^{2\log d}\right)^{\frac{1}{2\log d}} \leq 20\|\mathbf{B}_t\|_2 \ . \tag{9.13}$$

Similarly, for the lower bound, we have that

$$\|v\|_2^{\frac{1}{\log d}} \geq (0.1\mathrm{tr}(\mathbf{M}_t^2))^{\frac{1}{2\log d}} \geq (0.1)^{\frac{1}{2\log d}}\|\mathbf{B}_t\|_2 \geq (0.1)\|\mathbf{B}_t\|_2 \ , \tag{9.14}$$

where in the last inequality we assumed that the dimension is sufficiently large. Putting Equations (9.11) to (9.13) together, with at least constant non-zero probability, we have that $0.1\|\mathbf{B}_t\|_2 \leq \|v\|_2^{1/\log d} \leq 20\|\mathbf{B}_t\|_2$. By repeating the procedure $O(\log(1/\tau'))$ times and taking the median, we boost the probability of failure to τ'. By union bound, choosing $\tau' = \tau/K$ makes the event hold for all iterations $t \in [K]$ simultaneously with probability $1-\tau$. □

The remainder of this section is dedicated to showing that $\|\widehat{\mathbf{M}}_t - \mathbf{M}_t\|_2 \leq \min\left(\frac{\delta/\epsilon}{R}, \frac{0.01}{\sqrt{d}}\right)\|\mathbf{M}_t\|_F$. We require the following lemma, which we prove in Appendix F.3.3. We use $\prod_{i=1}^p \mathbf{B}_i$ is to denote the matrix product $\mathbf{B}_1\mathbf{B}_2\cdots\mathbf{B}_p$.

Lemma 9.4.14. *Let $\mathbf{A}, \mathbf{B}, \mathbf{B}_1, \ldots, \mathbf{B}_p$ be symmetric $d \times d$ matrices and define $\mathbf{M} = \mathbf{B}^p, \mathbf{M}_S = \prod_{i=1}^p \mathbf{B}_i$. If $\|\mathbf{B}_i - \mathbf{B}\|_2 \leq \delta\|\mathbf{B}\|_2$, then $\|\mathbf{M}_S - \mathbf{B}^p\|_2 \leq p\delta(1+\delta)^p\|\mathbf{B}\|_2^p$.*

We are now ready to prove our main technical result.

Lemma 9.4.15. *Assume that $\|\mathbf{B}_t\|_2 \geq (C_1/2)\delta^2/\epsilon$ and $\mathbb{E}_{X\sim P}[w_t(X)] \geq 1 - O(\epsilon)$ hold in the t-th iteration of Algorithm 8. If $\widehat{W}$ and every $\widehat{\mathbf{B}}_{t,k}$ in the product $\widehat{\mathbf{M}}_t = \prod_{k=1}^{\log d}\widehat{\mathbf{B}}_{t,k}$ is calculated using*

$$\tilde{n} \geq CR^2(\log d)^2 \max\left(d, \frac{\epsilon^2 d}{\delta^4}, \frac{R^2\epsilon^2}{\delta^2}, \frac{R^2\epsilon^4}{\delta^6}\right)\log\left(\frac{d\log d}{\tau}\right)$$

samples, where C is a sufficiently large constant, we have that

$$\|\widehat{\mathbf{M}}_t - \mathbf{M}_t\|_2 \leq 0.01\min\left(\frac{\delta/\epsilon}{R}, \frac{1}{\sqrt{d}}\right)\|\mathbf{M}_t\|_F \ ,$$

with probability at least $1 - \tau$.

Proof. Let $T := \delta/\epsilon$ and $p := \log d$ for brevity. Using Lemma 9.4.14 we have that $\|\widehat{\mathbf{M}}_t - \mathbf{M}_t\|_2 \leq p\gamma e^{\gamma p}\|\mathbf{M}_t\|_2$, where $\gamma > 0$ is such that

$$\|\widehat{\mathbf{B}}_{t,k} - \mathbf{B}_t\|_2 \leq \gamma \|\mathbf{B}_t\|_2 \tag{9.15}$$

for all $k \in [p]$. Therefore, for the lemma to hold, it suffices that $p\gamma e^{\gamma p} \leq 0.01 \min\left(\frac{T\|\mathbf{M}_t\|_F}{R\|\mathbf{M}_t\|_2}, \frac{1}{\sqrt{d}}\right)$. For that, it suffices to choose $\gamma = \frac{0.01}{3p}\min\left(\frac{1}{\sqrt{d}}, \frac{T\|\mathbf{M}_t\|_F}{R\|\mathbf{M}_t\|_2}\right)$. At this point, we also assume two things: First, that the estimate $\widehat{\mathbf{\Sigma}}_{t,k}$ (defined in step 4a) is such that $\|\widehat{\mathbf{\Sigma}}_{t,k} - \mathbf{\Sigma}_t\|_2 \leq \epsilon' \|\mathbf{\Sigma}_t\|_2$ for some ϵ' to be specified later on. Second, that we have an estimate $\widehat{W}_t$ for $\mathbb{E}_{X\sim P}[w_t(X)]$ such that

$$\widehat{W}_t = \mathop{\mathbb{E}}_{X\sim P}[w_t(X)] + \eta \,, \tag{9.16}$$

with $|\eta| \leq \xi$ for some $\xi \leq 1$ to be decided later. By Hoeffding's inequality, if we compute $\widehat{W}$ as shown in Step 1, then $\log(2/\tau)/\xi^2$ samples suffice to guarantee that Equation (9.16) holds with probability $1 - \tau/2$. We now focus on Equation (9.15). We note that

$$\begin{aligned}
\|\widehat{\mathbf{B}}_{t,k} - \mathbf{B}_t\|_2 &= \left\| \left(\mathop{\mathbb{E}}_{X\sim P_t}[w_t(X)] + \eta \right)^2 \widehat{\mathbf{\Sigma}}_{t,k} - \mathop{\mathbb{E}}_{X\sim P_t}[w_t(X)]^2 \mathbf{\Sigma}_t \right\|_2 \\
&\leq \mathop{\mathbb{E}}_{X\sim P_t}[w_t(X)]^2 \|\widehat{\mathbf{\Sigma}}_{t,k} - \mathbf{\Sigma}_t\|_2 + (\eta^2 + 2\eta \mathop{\mathbb{E}}_{X\sim P_t}[w_t(X)]) \|\widehat{\mathbf{\Sigma}}_{t,k}\|_2 \\
&\leq \|\widehat{\mathbf{\Sigma}}_{t,k} - \mathbf{\Sigma}_t\|_2 + 3\xi(\|\widehat{\mathbf{\Sigma}}_{t,k} - \mathbf{\Sigma}\|_2 + \|\mathbf{\Sigma}_t\|_2) \\
&= (1 + 3\xi)\|\widehat{\mathbf{\Sigma}}_{t,k} - \mathbf{\Sigma}_t\|_2 + 3\xi\|\mathbf{\Sigma}_t\|_2 \,.
\end{aligned}$$

By choosing $\xi = \min(1, \epsilon'/3)$ and $\epsilon' = \frac{1}{5}\gamma\|\mathbf{B}_t\|_2/\|\mathbf{\Sigma}_t\|_2$, the above implies that Equation (9.15) holds. Thus, it suffices to show that $\|\widehat{\mathbf{\Sigma}}_{t,k} - \mathbf{\Sigma}_t\|_2 \leq \epsilon' \|\mathbf{\Sigma}_t\|_2$ for our choice of ϵ'. Note that Fact 9.2.5 is not directly applicable to the distribution P_t since it does not

have zero mean. This is why we are working with samples of the form $(X - X')/\sqrt{2}$. By Fact 9.2.5 with ϵ' set as above and $\tau = \tau/(2p)$, we have the following upper bound on the sufficient number of samples:

$$\begin{aligned}
\tilde{n} &= C\frac{R^2}{\epsilon'^2\|\mathbf{\Sigma}_t\|_2}\log\left(\frac{2pd}{\tau}\right) \\
&\lesssim \frac{R^2}{\gamma^2}\frac{\|\mathbf{\Sigma}_t\|_2}{\|\mathbf{B}_t\|_2^2}\log\left(\frac{pd}{\tau}\right) \\
&\lesssim \frac{R^2\epsilon}{\delta^2\gamma^2}\frac{\|\mathbf{\Sigma}_t\|_2}{\|\mathbf{B}_t\|_2}\log\left(\frac{pd}{\tau}\right) && \text{(using } \|\mathbf{B}_t\|_2 \geq (C_1/2)\delta^2/\epsilon\text{)} \\
&\lesssim \frac{R^2\epsilon}{\delta^2\gamma^2}\max\left(1, \frac{\epsilon}{\delta^2}\right)\log\left(\frac{pd}{\tau}\right) && (9.17) \\
&\lesssim \frac{R^2p^2\epsilon}{\delta^2}\max\left(d, \frac{R^2\epsilon^2}{\delta^2}\frac{\|\mathbf{M}_t\|_2^2}{\|\mathbf{M}_t\|_F^2}\right)\max\left(1, \frac{\epsilon}{\delta^2}\right)\log\left(\frac{pd}{\tau}\right) \\
&\leq \frac{R^2p^2\epsilon}{\delta^2}\max\left(d, \frac{R^2\epsilon^2}{\delta^2}\right)\max\left(1, \frac{\epsilon}{\delta^2}\right)\log\left(\frac{pd}{\tau}\right) && \text{(using } \|\mathbf{M}_t\|_2 \leq \|\mathbf{M}_t\|_F\text{)} \\
&\leq \frac{R^2p^2\epsilon}{\delta^2}\max\left(d, \frac{\epsilon d}{\delta^2}, \frac{R^2\epsilon^2}{\delta^2}, \frac{R^2\epsilon^3}{\delta^4}\right)\log\left(\frac{pd}{\tau}\right)\ , && (9.18)
\end{aligned}$$

Equation (9.17) is derived as follows: First we note that $\|\mathbf{\Sigma}_t\|_2 \leq \frac{\|\mathbf{B}_t\|_2+1}{\mathbb{E}_{X\sim P}[w_t(X)]^2} \lesssim \|\mathbf{B}_t\|_2 + 1$, where the last inequality uses our assumption that $\mathbb{E}_{X\sim P}[w_t(X)] \geq 1 - O(\epsilon)$. We combine this with $\|\mathbf{B}_t\|_2 \gtrsim \delta^2/\epsilon$ as follows:

$$\frac{\|\mathbf{\Sigma}_t\|_2^2}{\|\mathbf{B}_t\|_2^2} \lesssim \frac{(\|\mathbf{B}_t\|_2+1)^2}{\|\mathbf{B}_t\|_2^2} \leq 2 + \frac{2}{\|\mathbf{B}_t\|_2^2} \lesssim 1 + \frac{1}{(\delta^2/\epsilon)^2}\ .$$

Regarding the samples required to achieve Equation (9.16), a sufficient number is

$$\begin{aligned}
\tilde{n} &= C\log\left(\frac{2}{\tau}\right)\frac{1}{\xi^2} \\
&\lesssim \log\left(\frac{1}{\tau}\right)\max\left(1, \frac{1}{\epsilon'^2}\right) \\
&\lesssim \log\left(\frac{1}{\tau}\right)\max\left(1, \frac{1}{\gamma^2}\frac{\|\mathbf{\Sigma}_t\|_2^2}{\|\mathbf{B}_t\|_2^2}\right) \\
&\lesssim \log\left(\frac{1}{\tau}\right)\max\left(1, \frac{1}{\gamma^2}\max\left(1, \frac{\epsilon^2}{\delta^4}\right)\right)
\end{aligned}$$

$$\lesssim \log\left(\frac{1}{\tau}\right) \max\left(1, p^2 \max\left(d, \frac{R^2\epsilon^2}{\delta^2}\right) \max\left(1, \frac{\epsilon^2}{\delta^4}\right)\right)$$
$$\lesssim \log\left(\frac{1}{\tau}\right) \max\left(p^2 d, \frac{p^2 R^2 \epsilon^2}{\delta^2}, \frac{p^2 d \epsilon^2}{\delta^4}, \frac{p^2 R^2 \epsilon^4}{\delta^6}\right) ,$$

which is smaller compared to the right-hand side of Equation (9.18). □

9.4.3.3 Item 3

The following lemma establishes that the estimator of Item 3 is accurate when called once. By using a union bound on the maximum number of times that it can be called, we get the sample complexity requirement of $n = O\left((R^2/(\delta^2/\epsilon))\text{polylog}\left(d, R, \frac{1}{\epsilon}, \frac{1}{\tau}\right)\right)$.

Lemma 9.4.16. *Consider the context of Algorithm 8 and denote $T_t := c\widehat{\lambda}_t \|\mathbf{U}_t\|_F^2$. Given a weight function $w : \mathbb{R}^d \to [0,1]$, there exists an estimator $f(w)$ on $n = O(\frac{R^2}{\delta^2/\epsilon}\log(1/\tau))$ samples such that, if $\mathbb{E}_{X\sim P}[w(X)\tilde{\tau}_t(X)] > T_t$, then with probability at least $1-\tau$, $f > T_t/2$. Similarly, $\mathbb{E}_{X\sim P}[w(X)\tilde{\tau}_t(X)] < T_t$ implies $f < (3/2)T_t$. Moreover, the estimator uses $O(\log(1/\tau))$ memory and runs in $O(nd)$ time.*

Proof. We show the first direction; the other one has a symmetric proof. Suppose $\mathbb{E}_{X\sim P}[w(X)\tilde{\tau}_t(X)] > c\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2$. It suffices to show that with probability at least 0.9 we have that

$$\frac{1}{n}\sum_{i=1}^{N} w(X_i)\tilde{\tau}_t(X_i) > \frac{3}{4}\mathop{\mathbb{E}}_{X\sim P}[w(X)\tilde{\tau}_t(X)] - \frac{1}{4}c\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2 , \tag{9.19}$$

as we can repeat the procedure $O(\log(1/\tau))$ times and take the majority vote to boost the probability to $1-\tau$. By Chebyshev's inequality, we have that with probability 0.9 it holds that

$$\frac{1}{n}\sum_{i=1}^{n} w(X_i)\tilde{\tau}_t(X_i) > \mathop{\mathbb{E}}_{X\sim P}[w(X)\tilde{\tau}_t(X)] - \sqrt{\frac{10\,\mathbf{Var}_{X\sim P}(w(X)\tilde{\tau}(X))}{n}} .$$

Therefore, it suffices to have $\sqrt{\frac{10\,\mathbf{Var}_{X\sim P}(w(X)\tilde{\tau}(X))}{n}} \leq \frac{1}{4}\,\mathbb{E}_{X\sim P}[w(X)\tilde{\tau}_t(X)] + \frac{1}{4}c\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2$, and thus we need n to be a sufficiently large multiple of $\mathbf{Var}_{X\sim P}(w(X)\tilde{\tau}_t(X))/(\mathbb{E}_{X\sim P}[w(X)\tilde{\tau}_t(X)] + c\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2)^2$. For that, it suffices to choose

$$n = \Theta\left(\frac{\mathbf{Var}_{X\sim P}(w(X)\tilde{\tau}_t(X))}{\mathbb{E}_{X\sim P}[w(X)\tilde{\tau}_t(X)]c\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2}\right) .$$

We now focus on bounding by above the right-hand side. Let $T_t' := C_3\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2$ be the threshold used in the definition of $\tilde{\tau}_t(x) = \tilde{g}_t(x)\,\mathbb{I}\{\tilde{g}_t(x) \geq T_t'\}$. For the variance we have that

$$\begin{aligned}
\mathbf{Var}(w(X)\tilde{\tau}_t(X)) &\leq \mathop{\mathbb{E}}_{X\sim P}[((w(X)\tilde{\tau}_t(X))^2] \\
&\leq \mathop{\mathbb{E}}_{X\sim P}[w(X)\tilde{\tau}_t^2(X)] \\
&\lesssim R^2\|\mathbf{U}_t\|_F^2 \mathop{\mathbb{E}}_{X\sim P}[w(X)\tilde{\tau}_t(X)] \,, \qquad (9.20)
\end{aligned}$$

where the last inequality uses that $\mathbb{E}_{X\sim P_t}[\tilde{\tau}_t^2(X)] = \mathbb{E}_{X\sim P_t}[\tilde{g}_t^2(X)\,\mathbb{I}\{\tilde{g}_t(X) \geq T_t'\}]$ and bounds from above the one of the two factors of $\tilde{g}_t$ as follows:

$$\tilde{g}_t(x) = \|\mathbf{U}_t(x-\mu_t)\|_2^2 \leq \|\mathbf{U}_t\|_F^2 R^2 \,, \qquad (9.21)$$

where $\mathbf{U}_t$ is the matrix used in Line 29 of the algorithm. Using Equation (9.20), the number of samples that suffice can now be bounded as follows:

$$\frac{\mathbf{Var}_{X\sim P}(\tilde{\tau}_t(X))}{\widehat{\lambda}_t\,\mathbb{E}_{X\sim P}[w(X)\tilde{\tau}_t(X)]\|\mathbf{U}_t\|_F^2} \lesssim \frac{R^2\|\mathbf{U}_t\|_F^2}{\widehat{\lambda}_t\|\mathbf{U}_t\|_F^2} \lesssim \frac{R^2}{\delta^2/\epsilon} \,,$$

where we used that $\widehat{\lambda}_t > C_2\delta^2/\epsilon$ from Line 22 of our algorithm. □

9.5 Applications: Beyond Robust Mean Estimation

In this section, we develop robust streaming algorithms with near-optimal space complexity for more complex statistical tasks, specifically for robust covariance estimation and robust stochastic optimization. The main idea enabling these applications is that these tasks can be effectively reduced to robust mean estimation.

9.5.1 Robust Covariance Estimation

In this subsection, we study the problem of estimating the covariance matrix $\mathbf{\Sigma}$ of a distribution D, having access to ϵ-corrupted samples from D in the sense of Definition 9.1.2. Let $X \sim D$ and the Kronecker product $Y = X \otimes X$. Note that $\mathbb{E}[Y] = \mathbf{\Sigma}^\flat$, where $\flat$ denotes the flattening operation. Then, using any robust mean estimation algorithm on this d^2-dimensional distribution, one efficiently compute a vector close to $\mathbf{\Sigma}^\flat$ in ℓ_2-norm, which translates to a Frobenius-norm guarantee for $\mathbf{\Sigma}$. Of course, our mean estimator works as long as the distribution of Y is stable. If $\mathrm{Cov}[Y]$ is bounded from above by a multiple of the identity matrix, then Y is $(\epsilon, O(\sqrt{\epsilon}))$-stable with respect to $\mathbf{\Sigma}^\flat$, and thus we get the following as a corollary of Theorem 9.4.2:

Theorem 9.5.1 (Robust Covariance Estimation for Distributions with Bounded Moments)**.** *Let a distribution D with $\mathrm{Cov}_{X\sim D}[X \otimes X] \preceq \mathbf{I}_{d^2}$ and denote by $\mathbf{\Sigma}$ its covariance matrix. Let $d \in \mathbb{Z}_+$, $0 < \tau < 1$ and $0 < \epsilon < \epsilon_0$ for a sufficiently small constant ϵ_0. There exists an algorithm that given ϵ, τ and a set of $n = (d^4/\epsilon)\mathrm{polylog}(d, 1/\epsilon, 1/\tau)$ samples in the single-pass streaming model of Definition 9.1.1 from a distribution Q with $d_{\mathrm{TV}}(D, Q) \leq \epsilon$, runs in time $nd^2\mathrm{polylog}\,(d, 1/\epsilon, 1/\tau)$, uses memory $d^2\mathrm{polylog}\,(d, 1/\epsilon, 1/\tau)$, and outputs a matrix $\widehat{\mathbf{\Sigma}}$ such that $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F = O(\sqrt{\epsilon})$, with probability at least $1 - \tau$.*

For the special case when D is Gaussian we have that the fourth moment tensor of D is bounded:

Fact 9.5.2 (see, e.g., [CDGW19]). *Let $X \sim \mathcal{N}(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma} \preceq \mathbf{I}_d$ and $Y = X \otimes X$. Then, $\mathrm{Cov}[Y] \preceq 2\mathbf{I}_{d^2}$.*

Using the above fact, we have that the guarantees of Theorem 9.5.1 hold in the Gaussian case, giving an algorithm for $O(\sqrt{\epsilon})$-approximation in Frobenius norm. However, the information-theoretic lower bound for covariance estimation of the Gaussian distribution is of the order of ϵ. We can plug-in our streaming robust mean estimation algorithm to the covariance estimator given in [CDGW19], and achieve the nearly-optimal error of $O(\epsilon \log(1/\epsilon))$. This algorithm creates a series of estimates $\widehat{\mathbf{\Sigma}}_i$. At the $(i+1)$-th step, all samples are multiplied by $\widehat{\mathbf{\Sigma}}_i^{-1/2}$ thus, given that $\widehat{\mathbf{\Sigma}}_i$ is a good approximation for $\mathbf{\Sigma}$, this makes the distribution of the transformed samples closer to $\mathcal{N}(0, \mathbf{I}_d)$, which in turn allows us to produce a better approximation $\widehat{\mathbf{\Sigma}}_{i+1}$ of $\mathbf{\Sigma}$. The resulting guarantees are summarized in the following theorem.

Theorem 9.5.3 (Robust Gaussian Covariance Estimation). *Let Q be a distribution on $\mathbb{R}^d$ with $d_{\mathrm{TV}}(Q, \mathcal{N}(0, \mathbf{\Sigma})) \leq \epsilon$ and assume that $\frac{1}{\kappa}\mathbf{I}_d \preceq \mathbf{\Sigma} \preceq \mathbf{I}_d$, for some $\kappa > 0$. There is a single-pass streaming algorithm that uses $n = (d^4/\epsilon^2)\mathrm{polylog}(d, \kappa, 1/\epsilon, 1/\tau)$ samples from Q, runs in time $nd^2\mathrm{polylog}\,(d, \kappa, 1/\epsilon, 1/\tau)$, uses memory $d^2\mathrm{polylog}\,(d, \kappa, 1/\epsilon, 1/\tau)$, and outputs a matrix $\widehat{\mathbf{\Sigma}}$ such that $\|\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2} - \mathbf{I}_d\|_F = O(\epsilon \log(1/\epsilon))$, with probability at least $1 - \tau$.*

The reader is referred to Appendix F.5 for more details on using Algorithm 8 to obtain Theorem 9.5.3.

9.5.2 Stochastic Convex Optimization

Here we explore the implications of Algorithm 8 in outlier-robust stochastic convex optimization. This subsection crucially leverages the prior works [PSBR20; DKKLSS19], which apply robust mean estimation algorithms to perform robust stochastic optimization. In particular, we follow the framework of [PSBR20].

Concretely, we study the following generic optimization problem: Let a parameter space Θ, sample space $\mathcal{Z}$, and a loss function $f(\theta; z) : \Theta \times \mathcal{Z} \to \mathbb{R}^+$. For an unknown distribution D over $\mathcal{Z}$, the goal is to minimize the associated *risk* $\bar{f}(\theta) = \mathbb{E}_{z\sim D}[f(\theta; z)]$, given sample access to the distribution D. We will occasionally just write $f(\theta)$ instead of $f(\theta; z)$ when no confusion arises. This setup is central in machine learning, since it captures a plethora of learning tasks. For example, f can be a negative log-likelihood function for the learning problem of interest, e.g., square loss for linear regression and logistic loss for logistic regression. In the robust version of the problem, the algorithm has access only to an ϵ-corrupted version of D in the sense of Definition 9.1.2.

We start by recalling a generic optimization algorithm that works whenever $\bar{f}$ is τ_ℓ-strongly convex and τ_u-smooth, i.e., for all $\theta_1, \theta_2 \in \Theta$, we have that

$$\frac{\tau_\ell}{2}\|\theta_1 - \theta_2\|_2^2 \le \bar{f}(\theta_1) - \bar{f}(\theta_2) - (\nabla \bar{f}(\theta_2))^\top(\theta_1 - \theta_2) \le \frac{\tau_u}{2}\|\theta_1 - \theta_2\|_2^2 \,.$$

We then give specific applications for robust linear regression and logistic regression.

The work of [PSBR20] provides an analysis of projected gradient descent assuming oracle access to approximations of the gradient:

Definition 9.5.4 ((α, β)-gradient estimator)**.** *A function $g(\theta)$ is an (α, β)-gradient estimator for $\bar{f}$ if $\|g(\theta) - \nabla \bar{f}(\theta)\|_2 \le \alpha\|\theta - \theta^*\|_2 + \beta$, for every $\theta \in \Theta$.*

Denoting by η the step size of gradient descent, define the following parameter:

$$\kappa := \sqrt{1 - \frac{2\eta\tau_\ell\tau_u}{\tau_\ell + \tau_u}} + \eta\alpha \,. \tag{9.22}$$

Theorem 9.5.5 ([PSBR20])**.** *Let the domains $\Theta, \mathcal{Z} \subset \mathbb{R}^d$, a distribution D over $\mathcal{Z}$, and a loss function $f : \Theta \times \mathcal{Z} \to \mathbb{R}^+$ such that $\bar{f}(\theta) := \mathbb{E}_{z\sim D}[f(\theta; z)]$ is τ_ℓ-strongly convex and τ_u-smooth. Let g be an (α, β)-gradient estimator with $\alpha < \tau_\ell$. Let κ from Equation (9.22) and θ^* be the*

Algorithm 10 Robust Gradient Descent

Input: $g(\cdot)$, τ
for $t = 0$ **to** $T - 1$ **do**
    $\theta^{t+1} = \arg\min_{\theta \in \Theta} \|\theta^t - \eta g(\theta)\|_2^2$
end for

minimizer of $\bar{f}$. *Then Algorithm 10, initialized at* θ^0 *with step size* $\eta = 2/(\tau_\ell + \tau_u)$, *after*

$$T = \log_{\frac{1}{\kappa}} \left(\frac{(1-\kappa)\|\theta^0 - \theta^*\|_2}{\beta} \right) \tag{9.23}$$

iterations, returns a vector $\widehat{\theta}$ *such that*

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{2}{1-\kappa}\beta \, . \tag{9.24}$$

If the distribution of the gradients has bounded covariance, then one can use the low-memory estimator of the previous sections in place of $g(\cdot)$. This bound on the covariance will not necessarily be known to the algorithm, thus we first need to strengthen the robust mean estimator so that it is adaptive to that unknown scale. This can be done using Lepski's method [Lep91; Bir01a] (the details are deferred to Appendix F.4). Having that version of the estimator at hand, we then obtain the following statement (see Appendix F.5 for the proof):

Corollary 9.5.6. *In the setting of Theorem 9.5.5, suppose that the distribution of gradients satisfies* $\mathrm{Cov}[\nabla f(\theta)] \preceq \sigma^2 \mathbf{I}_d$ *with* $\sigma^2 = \alpha^2 \|\theta - \theta^*\|_2^2 + \beta^2$ *for all* $\theta \in \Theta$, *where* $\alpha\sqrt{\epsilon} < \tau_\ell$. *Assume that the radius of the domain* Θ, $r := \max_{\theta \in \Theta} \|\theta\|_2$ *is finite. There exists a single-pass streaming algorithm that given* $O(T(d^2/\epsilon)\log(1 + \alpha r/\beta)\mathrm{polylog}(d, 1/\epsilon, T/\tau, 1 + \alpha r/\beta))$ *samples, runs in time* $Tnd\,\mathrm{polylog}(d, 1/\epsilon, T/\tau, 1 + \alpha r/\beta)$, *uses memory* $d\,\mathrm{polylog}(d, 1/\epsilon, T/\tau, 1 + \alpha r/\beta)$, *and returns a vector* $\widehat{\theta} \in \mathbb{R}^d$ *such that* $\|\widehat{\theta} - \theta^*\|_2 = O(\sqrt{\epsilon}\beta/(1-\kappa))$ *with probability at least* $1 - \tau$.

We now proceed to more specific applications, where we work out the parameters α, β for some distributions of interest.

9.5.2.1 Linear Regression

For linear regression, we assume the following generative model:

$$Y = X^\top \theta^* + Z \ , \tag{9.25}$$

where $\theta^* \in \mathbb{R}^d$ belongs in the ball $\|\theta^*\|_2 \leq r$, $X \sim D_x$, $Z \sim D_Z$ independently, and D_Z has zero mean. The loss function that we use in this case is $f(\theta) = \frac{1}{2}(Y - \theta^\top X)^2$, and the risk function is

$$\bar{f}(\theta) = \mathop{\mathbb{E}}_{(X,Y)}[f(\theta)] = \frac{1}{2}(\theta - \theta^*)^\top \mathop{\mathbb{E}}_{X \sim D_x}[XX^\top](\theta - \theta^*) + \frac{1}{2}\,\mathbf{Var}(Z) \ .$$

Letting $\lambda_{\max}(\mathbb{E}[XX^\top])$ and $\lambda_{\min}(\mathbb{E}[XX^\top])$ denote the largest and smallest eigenvalue of $\mathbb{E}[XX^\top]$ respectively, it can be checked that for any $\tau_\ell \leq \lambda_{\min}(\mathbb{E}[XX^\top])$ and $\tau_u \geq \lambda_{\max}(\mathbb{E}[XX^\top])$, $\bar{f}$ is τ_ℓ-strongly convex and τ_ℓ-smooth.

Since we want the distribution of gradients to be stable, we impose the following sufficient conditions on the distributions D_x and D_Z.

Assumption 9.5.7. *The random variables X, Z are independent and satisfy the following conditions:*

1. $\mathbb{E}_{Z \sim D_Z}[Z] = 0$
2. $\mathbf{Var}_{Z \sim D_Z}[Z] \leq \xi^2$
3. $\gamma \mathbf{I}_d \preceq \mathbb{E}_{X \sim D_x}[XX^\top] \preceq \sigma^2 \mathbf{I}_d$.
4. *For some constant $C > 0$, for every $v \in \mathcal{S}^{d-1}$, $\mathbb{E}_{X \sim D_x}[(X^\top v)^4] \leq C\sigma^4$.*

As shown below, these assumptions imply that the resulting distribution of the gradients has bounded covariance (and thus is stable with respect to its mean).

Lemma 9.5.8 (see, e.g., [DKKLSS19])**.** *For D_x, D_Z satisfying Assumption 9.5.7, for every $\theta \in \Theta$, we have that $\mathrm{Cov}[\nabla f(\theta)] \preceq (4\sigma^2\xi^2 + 4C\sigma^4\|\theta - \theta^*\|_2^2)\mathbf{I}_d$.*

Having Lemma 9.5.8 in hand, Corollary 9.5.6 gives the following.

Theorem 9.5.9 (Robust Linear Regression; full version of Theorem 9.1.5)**.** *Consider the linear regression model of Equation (9.25) and suppose that Assumption 9.5.7 holds. Let $0 < \epsilon < \epsilon_0$ for a sufficiently small constant ϵ_0. Assume that $C\sigma^2\sqrt{\epsilon} < \gamma/2$. Let κ, T as in Equations (9.22) and (9.23) with $\tau_\ell = \gamma$, $\tau_u = \sigma^2$. There is an algorithm that uses $n = T \cdot (d^2/\epsilon)\log(1 + r\sigma/\xi)\,\mathrm{polylog}\,(d, 1/\epsilon, T/\tau, 1 + r\sigma/\xi)$ samples, runs in time $Tnd\,\mathrm{polylog}(d, 1/\epsilon, T/\tau, 1+r\sigma/\xi)$, uses memory $d\,\mathrm{polylog}(d, 1/\epsilon, T/\tau, 1+r\sigma/\xi)$, and returns a vector $\widehat{\theta} \in \mathbb{R}^d$ such that $\|\widehat{\theta}-\theta^*\|_2 = O(\sigma\xi\sqrt{\epsilon}/(1-\kappa))$ with probability at least $1 - \tau$.*

Proof. In our case, we have that $\tau_\ell = \gamma$ and $\tau_u = \sigma^2$. Given the bound of Lemma 9.5.8, we use Corollary 9.5.6 with $\alpha = 2C\sigma^2$ and $\beta = 2\sigma\xi$. The requirement from that corollary that $\alpha\sqrt{\epsilon} \le \tau_\ell$ becomes $C\sigma^2\sqrt{\epsilon} < \gamma/2$. Moreover, $\alpha r/\beta = O(r\sigma/\xi)$. □

9.5.2.2 Logistic Regression

We consider the joint distribution of $X \in \mathbb{R}^d, Y \in \{0, 1\}$, where $X \sim D_x$ and Y given X is Bernoulli random variable:

$$Y|X \sim \text{Bernoulli}(p), \qquad \text{with } p = \frac{1}{1 + e^{-x^\top\theta^*}}\,. \tag{9.26}$$

The loss function we are minimizing in this case is the negative log-likelihood, which eventually can be written as $f(\theta) = -(\theta^\top x)y + \Phi(\theta^\top x)$, where $\Phi(t) := \log(1 + e^t)$. Regarding the strong convexity parameters, the Hessian of $\bar{f}$ can be shown to be

$$\nabla^2\bar{f}(\theta) = \mathop{\mathbb{E}}_{X\sim D_x}\left[\frac{e^{\theta^\top X}}{(1 + e^{\theta^\top X})^2}XX^\top\right]\,. \tag{9.27}$$

The parameter space Θ needs to be bounded in order for the eigenvalues of the Hessian to remain away from zero; we thus use $\Theta = \{\theta \in \mathbb{R}^d \ : \ \|\theta\|_2^2 \le r\}$ with $r > 0$ being a universal constant. We also impose the following assumptions on the covariates.

Assumption 9.5.10. *We assume the following for the distribution of X:*

1. $\mathbb{E}[X] = 0$.
2. (concentration) *For some constant $C > 0$, $\mathbb{E}[XX^\top] \preceq C^2\mathbf{I}_d$.*
3. (anti-concentration) *There exists constant $c_1 > 0$ and $c_2 \in (0, 1/2)$ such that for every unit vector v, $\mathbb{P}_{X\sim D_x}[(v^\top X)^2 > c_1\|v\|_2^2] \ge c_2$.*

Under these assumptions, we have the following:

Lemma 9.5.11 (Lemma 4 in [PSBR20])**.** *Supposing that Assumption 9.5.10 holds, for every $\theta \in \Theta$, we have that $\mathrm{Cov}[\nabla f(\theta)] \preceq O(1)\mathbf{I}_d$.*

The above lemma shows that the distribution of $\nabla f(\theta)$ is $(\epsilon, O(\sqrt{\epsilon}))$-stable, and thus using our robust mean estimation algorithm one can get an (α, β)-gradient estimator with $\alpha = 0$ and $\beta = O(\sqrt{\epsilon})$. This proves the following (see Appendix F.5 for a detailed proof):

Theorem 9.5.12 (Robust Logistic Regression; full version of Theorem 9.1.6)**.** *Consider the logistic regression model of Equation (9.26) with the domain Θ of the unknown regressor being the ball of radius r, for some universal constant $r > 0$, and suppose that Assumption 9.5.10 holds. Assume that $0 < \epsilon < \epsilon_0$ for a sufficiently small constant ϵ_0. There is a single-pass streaming algorithm that uses $n = (d^2/\epsilon)\,\mathrm{polylog}\,(d, 1/\epsilon, 1/\tau)$ samples, runs in time $nd\,\mathrm{polylog}(d, 1/\epsilon, 1/\tau)$, uses memory $d\,\mathrm{polylog}(d, 1/\epsilon, 1/\tau)$, and returns a vector $\widehat{\theta} \in \mathbb{R}^d$ such that $\|\widehat{\theta} - \theta^*\|_2 = O(\sqrt{\epsilon})$ with probability at least $1 - \tau$.*

9.5.3 Byzantine Adversary and Second-order Optimal Point

We now describe the application of our algorithm to the setting of robust distributed non-convex optimization. As before, for a parameter space $\Theta \subset \mathbb{R}^d$, a loss function $f : \Theta \times \mathcal{Z} \to \mathbb{R}^+$, and a distribution D over $\mathcal{Z}$, the goal is to approximately minimize $\bar{f}(\theta) = \mathbb{E}_{z\sim D}[f(\theta; z)]$. In this section, we consider the case when D is a uniform distribution over mn points $\{z_{i,j} : i \in [m], j \in [n]\}$ that are distributed over m machines (workers), with each machine having access to n samples. Furthermore, we do not impose convexity constraints on f, and thus would restrict ourselves to finding a second-order stationary point, i.e., a stationary point $\widehat{\theta}$ such that the Hessian on $\widehat{\theta}$ is not too negative in any direction.

We now explain the distributed setup in more detail. There are m *workers* who have their own private samples, and a single *master* machine which is responsible for collecting gradient estimates from the workers and updating the candidate vector iteratively. Concretely, the i-th worker has n samples $\{z_{ij}\}_{j=1}^n$. The master machine queries all workers with a parameter $\theta \in \Theta$, and each i-th worker responds with $g_i(\theta)$, where $g_i : \mathbb{R}^d \to \mathbb{R}^d$ is defined as follows: (i) if the i-th worker is honest, then $g_i(\theta)$ is the average of the gradients of f at θ of their samples, i.e., $g_i(\theta) := (1/n)\sum_{j=1}^n \nabla f(\theta; z_{ij})$, and (ii) if the i-th worker is dishonest, then $g_i(\cdot)$ is an arbitrary function. In our results, we require only that $(1-\epsilon)$-fraction of workers are honest. Recent work of [YCRB19] provided an algorithm that uses a robust mean estimation algorithm on the gradients as a black-box procedure. In particular, the algorithm of [YCRB19] requires only an access to the following oracle:

Definition 9.5.13 (Δ-inexact gradient)**.** *We call the vector $v(\theta)$ a Δ-inexact gradient of $\bar{f}$ at the point θ if $\|v(\theta) - \nabla\bar{f}(\theta)\|_2 \le \Delta$.*

We assume that each worker machine has access to its own samples throughout the optimization process, and our goal is to reduce the memory requirement of the master

machine. Thus, we will use the algorithm from Corollary 9.4.3 to calculate Δ-inexact gradient for the master machine, which requires only an oracle access to the gradient estimates $\{g_i(\theta) : i \in [m]\}$.

Assumption 9.5.14. *Let $\mathcal{I} \subseteq [m]$ be the set of honest workers with $|\mathcal{I}| \geq (1-\epsilon)m$.*

1. *There exists δ with $0 \leq \epsilon \leq \delta \leq \delta_0$, for some sufficiently small δ_0, such that for every $\theta \in \Theta$, the set $\{g_i(\theta) \mid i \in \mathcal{I}\}$ is $(C\epsilon, \delta)$-stable with respect to $\nabla \bar{f}(\theta)$ for a large enough constant C.*

2. *We assume that $\bar{f}$ is L-smooth and ρ-Hessian Lipschitz on Θ, i.e., for every $\theta_1, \theta_2 \in \Theta$ we have that $\|\nabla \bar{f}(\theta_1) - \nabla \bar{f}(\theta_2)\|_2 \leq L\|\theta_1 - \theta_2\|_2$ and $\|\nabla^2 \bar{f}(\theta_1) - \nabla^2 \bar{f}(\theta_2)\|_2 \leq \rho\|\theta_1 - \theta_2\|_2$.*

We note that if the samples of honest workers are sampled i.i.d. from a distribution P, then the set $\{g_i(\theta) : i \in \mathcal{I}\}$ for a fixed $\theta \in \Theta$ will be stable with respect to $\nabla \bar{f}(\theta)$ with high probability, provided that the distribution of $\nabla f(\theta; Z)$ satisfies mild concentration under $Z \sim P$ and m is sufficiently large. Using a standard cover argument with the smoothness properties of f, this can be extended to all $\theta \in \Theta$. We thus obtain the following theorem, under Assumption 9.5.14.

Theorem 9.5.15. *Suppose that Assumption 9.5.14 holds. Let m denote the number of workers. Assume $0 < \tau < 1$, $\Delta := C'\delta < 1$, for C' a sufficiently large constant and define*

$$Q := 2\log\left(\frac{\rho(\bar{f}(\theta_0) - \inf_{\theta \in \mathbb{R}^d} \bar{f}(\theta))}{48L\tau(\Delta^{6/5}d^{3/5} + \Delta^{7/5}d^{7/10})}\right) , \quad T_{th} := \frac{L}{384(\rho^{1/2} + L(\Delta^{2/5}d^{1/5} + \Delta^{3/5}d^{3/10})} .$$

There is an algorithm where the master, if initialized at θ_0, does $T = \frac{2(\bar{f}(\theta_0) - \inf_{\theta \in \mathbb{R}^d} \bar{f}(\theta))}{3\Delta^2} Q\, T_{th}$ iterations, each running in $md\,\mathrm{polylog}(d, 1/\epsilon, T/\tau)$ time, uses $d\,\mathrm{polylog}(d, 1/\epsilon, T/\tau)$ memory, and outputs a vector $\widehat{\theta}$ such that, with probability $1-\tau$, $\|\nabla \bar{f}(\widehat{\theta})\|_2 \leq 4\Delta$ and $\lambda_{min}(\nabla^2 \bar{f}(\widehat{\theta})) \geq -\Delta^{2/5}d^{1/5}$.

9.6 Discussion

In this work, we gave the first efficient streaming algorithm with near-optimal space complexity for outlier-robust high-dimensional mean estimation. As an application, we also obtained low-space streaming algorithms for a range of other robust estimation tasks. Our work is a first step towards understanding the space complexity of high-dimensional robust statistics in the streaming setting.

Our work suggests a number of open problems. First, the sample complexity of our mean estimation algorithm is $\tilde{O}(d^2/\epsilon^2)$, while the information-theoretic optimum (without space constraints!) is $\tilde{O}(d/\epsilon^2)$. What is the *optimal* sample-space tradeoff? A similar question can be asked for the broader tasks of covariance estimation and stochastic optimization. A more general goal is to characterize the tradeoff between space complexity, number of passes, and sample size/runtime for other robust high-dimensional statistics tasks, e.g., clustering and learning of mixture models.

Finally, another research direction concerns the considered contamination model. Throughout this paper, we focused on the TV-contamination model. One can consider an even stronger contamination model with an adaptive adversary, where the outliers can be completely arbitrary (i.e., not follow any distribution), and the adversary can additionally control the order in which the points are presented in the stream. Is it possible to obtain $\tilde{O}_\epsilon(d)$-space single-pass streaming algorithms for robust mean estimation in the presence of such an adversary? While our algorithms can be shown to work in this model with a poly-logarithmic number of passes, it is not clear whether a single-pass algorithm with sub-quadratic space complexity exists in this setting.

उस ने सुकूत-ए-शब में भी अपना पयाम रख दिया
हिज्र की रात बाम पर माह-ए-तमाम रख दिया

— अहमद फ़राज़

We study hypothesis testing under communication constraints, where each sample is quantized before being revealed to a statistician. Without communication constraints, it is well known that the sample complexity of simple binary hypothesis testing is characterized by the Hellinger distance between the distributions. We show that the sample complexity of simple binary hypothesis testing under communication constraints is at most a logarithmic factor larger than in the unconstrained setting and this bound is tight. We develop a polynomial-time algorithm that achieves the aforementioned sample complexity. Our framework extends to robust hypothesis testing, where the distributions are corrupted in the total variation distance. Our proofs rely on a new reverse data processing inequality and a reverse Markov inequality, which may be of independent interest. For simple M-ary hypothesis testing, the sample complexity in the absence of communication constraints has a logarithmic dependence on M. We show that communication constraints can cause an exponential blow-up leading to $\Omega(M)$ sample complexity even for adaptive algorithms.

10.1 Introduction

Statistical inference has been extensively studied under constraints such as memory [Cov69; HC73a; HC73b; GRT18; BOS20; DKPP22], privacy [DR13; KOV16; DJW18; CKMSU19; GKKNWZ20], communication [Tsi93; BGMNW16; HÖW21; ACT20b], or a combination thereof [SVW16; Fel17; DS18; DGKR19; DR19; ACT20a], typically designed to model physical or economic constraints. Our work focuses on communication

constraints, where the statistician does not have access to the original samples—but only their quantized versions—generated through a communication-constrained channel. For example, instead of observing a sample $x \in \mathcal{X}$, the statistician might observe a single bit $f(x) \in \{0, 1\}$ for some function $f : \mathcal{X} \to \{0, 1\}$. The choice of the channel (here, the function f) crucially affects the quality of statistical inference and is the topic of study in our paper.

Under communication constraints, a recent line of work has established minimax optimal rates for a variety of problems, including distribution estimation and identity testing [Sha14; ACT20a; ACT20b; HÖW21; CKO21; Can22]. However, under the same constraints, the problem of simple hypothesis testing has received scant attention. Recall the simple hypothesis testing framework: Let $\mathcal{P}$ be a given finite set of distributions over the domain $\mathcal{X}$. Given i.i.d. samples $X_1, \ldots, X_n$ from an unknown distribution $p \in \mathcal{P}$, the goal is to correctly identify p with high probability, with n as small as possible. We denote this problem as $\mathcal{B}(\mathcal{P})$ and use $n^*(\mathcal{P})$ to denote its sample complexity; i.e., the minimum number of samples required to solve $\mathcal{B}(\mathcal{P})$.

When $\mathcal{P} = \{p, q\}$, the problem is referred to as the simple binary hypothesis testing problem and has a rich history in statistics [NP33; Wal45; HS73; Cam86]. Given its historical and practical significance, we have a deep understanding of this problem (cf. Section 10.2 for details). In particular, it is known that $n^*(\mathcal{P}) = \Theta(1/d_{\mathrm{h}}^2(p, q))$, where $d_{\mathrm{h}}(p, q)$ denotes the Hellinger distance between p and q.

Hypothesis testing under communication constraints was studied in detail in the 1980s and 1990s under the name "decentralized detection" [Tsi93]. Briefly, the setup involves n users and a central server. Each user i observes an i.i.d. sample X_i from an unknown distribution $p \in \mathcal{P}$, generates a message $Y_i \in \{0, 1, \ldots, D-1\}$ using a channel $\mathbf{T}_i$ (chosen by the statistician), and transmits Y_i to the central server. The central server observes $(Y_1, \ldots, Y_n)$ and produces an estimate $\widehat{p} \in \mathcal{P}$. The goal is to choose $(\mathbf{T}_1, \ldots, \mathbf{T}_n)$

so that the central server can identify p correctly with high probability, while keeping n as small as possible. We call this problem "simple hypothesis testing under communication constraints" and denote it by $\mathcal{B}(\mathcal{P}, D)$. We denote the corresponding sample complexity by $n^*(\mathcal{P}, D)$.

Simple binary hypothesis testing. We begin our discussion with the fundamental setting of simple binary hypothesis testing under communication constraints, i.e., $\mathcal{P} = \{p, q\}$. It is known that the central server should perform a likelihood ratio test [Tsi93]. Furthermore, an optimal choice of channels can be achieved using deterministic threshold tests; i.e., $Y_i = f_i(X_i)$ for some $f_i : \mathcal{X} \to \{0, 1, \ldots, D-1\}$, such that f_i is characterized by D intervals that partition $\mathbb{R}_+$, and $f_i(x) = j$ if and only if $p(x)/q(x)$ lies in the j^{th} interval. The optimality of threshold tests crucially relies on the f_i's being possibly non-identical across users [Tsi88].

Nonetheless, several fundamental statistical and computational questions have remained unanswered. We begin with the following statistical question:

For $\mathcal{P} = \{p, q\}$, what is the sample complexity of $\mathcal{B}(\mathcal{P}, D)$, and what is $\frac{n^*(\mathcal{P},D)}{n^*(\mathcal{P})}$?

Let $n^* = n^*(\mathcal{P})$ and $n^*_{\text{bin}} = n^*(\mathcal{P}, 2)$ for notational convenience. A folklore result using Scheffe's test (Definition 10.2.5) implies that $n^*_{\text{bin}}/n^* \lesssim n^*$ (cf. Proposition G.1.2). One of our main results is an exponential improvement on this guarantee, showing that $n^*_{\text{bin}}/n^* \lesssim \log(n^*)$, i.e., communication constraints only lead to at most a logarithmic increase in sample complexity. More precisely, we show the following sample complexity bound:

$$n^*(\mathcal{P}, D) \lesssim n^*(\mathcal{P}) \max\left\{1, \frac{\log(n^*(\mathcal{P}))}{D}\right\}. \tag{10.1}$$

Furthermore, there exist cases where the bound (10.1) is tight (cf. Theorem 10.4.3). The bound can further be improved when the support sizes of p and q are smaller than

$\log(n^*(\mathcal{P}))$.

Turning to computational considerations, let p and q be distributions over k elements. Although the optimality of threshold tests implies that each user can search over $k^{\Omega(D)}$ possible such channels, this is prohibitive for large D. Such an exponential-time barrier has been highlighted as a major computational bottleneck in decentralized detection [Tsi93], leading to the following question:

Is there a $\mathrm{poly}(k, D)$*-time algorithm to compute channels* $(\mathbf{T}_1, \ldots, \mathbf{T}_n)$ *that achieve the sample complexity bound* (10.1)*?*

We answer this question affirmatively by showing that it suffices to consider threshold tests parametrized by a single quantity (cf. equation (10.5)). In fact, we show that it suffices to use an identical channel across the users (cf. Lemma 10.4.2).

Robustness to model misspecification. In many scenarios, it may be unreasonable to assume that the true distribution is either exactly p or q, but rather that it is close to one of them in total variation distance. Let ϵ be the amount of corruption, so that the underlying distribution p' belongs to $\mathcal{P}_1 \cup \mathcal{P}_2$, where $\mathcal{P}_1 := \{\tilde{p} : d_{\mathrm{TV}}(p, \tilde{p}) \leq \epsilon\}$ and $\mathcal{P}_2 := \{\tilde{p} : d_{\mathrm{TV}}(p, \tilde{q}) \leq \epsilon\}$, and d_{TV} denotes the total variation distance. Our goal is to design channels and a test such that, given samples from any distribution in $\mathcal{P}_1$ (respectively $\mathcal{P}_2$), we output p (respectively q) with high probability. Under the communication constraint of D messages, we denote this problem by $\mathcal{B}_{\texttt{robust}}(p, q, \epsilon, D)$.

As long as $\epsilon \lesssim d_{\mathrm{TV}}(p, q)$, we can use Scheffe's test to solve $\mathcal{B}_{\texttt{robust}}(p, q, \epsilon, 2)$ with at most $O(1/d^2_{\mathrm{TV}}(p, q))$ samples. We may hope to improve upon this by using the optimal channel $\mathbf{T}'$ for the uncontaminated hypothesis testing problem, $\mathcal{B}(\{p, q\}, D)$. It is, however, unclear if $\mathbf{T}'$ satisfies any robustness properties like the channel in Scheffe's test. We show in Proposition 10.4.8 that in the moderate contamination regime, when $\epsilon \lesssim d^2_{\mathrm{TV}}(p, q)$

(up to logarithmic factors), $\mathbf{T}'$ solves $\mathcal{B}_{\texttt{robust}}(p, q, \epsilon, D)$ with the same sample complexity as $\mathcal{B}(\{p, q\}, D)$. As a converse, we present cases where $\mathbf{T}'$ is not robust to larger ϵ.

For the large contamination setting, we combine our technical results with the framework of "least favorable distributions" pioneered by Huber [Hub65] and extended to the communication-constrained setting by Veeravalli, Basar, and Poor [VBP94]. We show that the robust sample complexity under communication constraints, $n^*_{\texttt{robust}}(p, q, \epsilon, D)$, increases by at most a logarithmic factor. Specifically, letting $n^*_{\texttt{robust}} := n^*_{\texttt{robust}}(p, q, \epsilon)$ be the robust sample complexity without any communication constraints, we obtain the following result in Theorem 10.4.6:

$$n^*_{\texttt{robust}}(p, q, \epsilon, D) \lesssim n^*_{\texttt{robust}} \max\left\{1, \frac{\log\left(n^*_{\texttt{robust}}\right)}{D}\right\}. \tag{10.2}$$

This rate can be much tighter than the one obtained using Scheffe's test. Moreover, the rate above is achieved by a computationally-efficient algorithm.

M-ary hypothesis testing. Finally, we consider the setting where $\mathcal{P}$ contains $M > 2$ distributions and allow the choice of channels to be adaptive, i.e., the channel $\mathbf{T}_i$ may depend on $Y_1, \ldots, Y_{i-1}$. For simplicity, we consider the setting where D and $\mathcal{P}$ are fixed and focus on the dependence on M. Using a standard tournament procedure, we show that there is an adaptive algorithm with sample complexity $O(M \log M)$. On the other hand, in the absence of communication constraints, it is known that the sample complexity is $O(\log M)$. We show that this exponential blow-up is necessary using the techniques from Braverman, Garg, Ma, Nguyen, and Woodruff [BGMNW16], i.e., the sample complexity under communication constraints is $\Omega(M)$. We also show $\Omega(\sqrt{M})$ lower bounds using two other techniques: (i) statistical query lower bounds [SVW16; FGRVX17], and (ii) the impossibility of ℓ_1-embedding [CS02; LMN05]. Although these bounds are weaker than the $\Omega(M)$ lower bound, they have other favorable properties:

the support size of the distributions in the hard instance is much smaller (k is linear in M as opposed to exponential in M), and the technical arguments that rely on the impossibility of ℓ_1-embeddings are elementary. Lastly, we consider the setting where all of the channels are restricted to be identical across users, which may be desirable in some applications. We provide specialized upper and lower bounds in this setting.

Our contributions. We summarize our main contributions as follows:

1. (Simple binary hypothesis testing.) We establish the minimax optimal sample complexity (cf. inequality (10.1)) of binary simple hypothesis testing under communication constraints (Theorems 10.4.1 and 10.4.3). Moreover, we provide an efficient algorithm, running in $\mathrm{poly}(k, D)$ time, to find a channel that achieves the minimax optimal sample complexity.

2. (Robust version of simple binary hypothesis testing.) Theorem 10.4.6 focuses on the robust hypothesis testing problem and shows that the robust sample complexity increases by at most a logarithmic factor, which is achievable using a computationally-efficient algorithm.

3. (M-ary hypothesis testing.) Generalizing to the setting of M-ary distributions, we show that for some cases, communication constraints can lead to an exponential increase in sample complexity, even for adaptive channels. We also derive results, both upper and lower bounds, specialized to settings where the channels are restricted (cf. Section 10.5).

4. (Technical results.) Along the way, we prove the following two technical results which may be of independent interest: (i) a reverse data processing inequality for general f-divergences and communication-constrained channels (Theorem 10.3.2), and (ii) a reverse Markov inequality for bounded random variables (Lemma 10.3.7).

The remainder of the paper is organized as follows: Section 10.2 defines notation, states the problem, and recalls useful facts. Section 10.3 contains a reverse data processing inequality for f-divergences. Section 10.4 uses these inequalities to derive our statistical and computational guarantees for binary hypothesis testing. Finally, Section 10.5 presents results for M-ary hypothesis testing. More technical proofs are deferred to the supplementary appendices.

10.2 Preliminaries

Notation: Throughout this paper, we will focus on discrete distributions. For $n \in \mathbb{N}$, we use $[n]$ to denote $\{1, \dots, n\}$ and $[0 : n]$ to denote $\{0, 1, \dots, n\}$. We use Δ_k to denote the set of distributions over k elements. For a distribution $p \in \Delta_k$ and $i \in [k]$, we use both p_i and $p(i)$ to denote the probability of element i under p. Given two distributions p and q, let $d_{\mathrm{TV}}(p, q)$ and $d_{\mathrm{h}}(p, q) := \sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}$ denote the total variation and Hellinger distances between p and q, respectively. Let $\beta_{\mathrm{h}}(p, q)$ denote the Hellinger affinity, i.e., $\beta_{\mathrm{h}}(p, q) := 1 - 0.5 d_{\mathrm{h}}^2(p, q)$. Given n distributions $p^{(1)}, \dots, p^{(n)}$, we use $\prod_{i=1}^n p^{(i)}$ to denote their product distribution. When each $p^{(i)} = p$, we use $p^{\otimes n}$ to denote the n-fold product distribution. For a set $A \subseteq \mathcal{X}$, we use $\mathbb{I}_A : \mathcal{X} \to \{0, 1\}$ to denote the indicator function of A. We consider $[a, b)$ to be an empty set when $b \leq a$. For a channel $\mathbf{T} : \mathcal{X} \to \mathcal{Y}$ and a distribution p over $\mathcal{X}$, we use $\mathbf{T}p$ to denote the distribution over $\mathcal{Y}$ when $X \sim p$ passes through the channel $\mathbf{T}$. As the channels between discrete distributions can be represented by column-stochastic matrices, we also use bold capital letters, such as $\mathbf{T}$, to denote the corresponding matrices. In particular, when p is a distribution over $[k]$, represented as a vector in $\mathbb{R}^k$, and $\mathbf{T}$ is a channel from $[k] \to [d]$, represented as a matrix $\mathbf{T} \in \mathbb{R}^{d \times k}$, the output distribution $\mathbf{T}p$ corresponds to the usual matrix-vector product. We use c, C, c', C', etc., to denote absolute positive constants, whose values might change from line to line, but with values which can be inferred by careful bookkeeping, while

c_1, C_1, c_2, C_2, etc., are used to denote absolute positive constants that remain the same throughout the proof. Finally, we use the following notations for simplicity: (i) $\lesssim$ and $\gtrsim$ to hide positive constants, (ii) the standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$, and (iii) $\mathrm{poly}(\cdot)$ to denote a quantity that is polynomial in its arguments.

10.2.1 Definitions and Basic Facts

Definition 10.2.1 (f-divergence)**.** *For a convex function $f : \mathbb{R}_+ \to \mathbb{R}$ with $f(1) = 0$, we use $I_f(p, q)$ to denote the f-divergence between p and q, defined as $I_f(p, q) := \sum_i q_i f\left(p_i/q_i\right)$.*[14]

We use the following facts:

Fact 10.2.2 (Properties of divergences [Tsy09; ZZ73])**.** *For any distributions $p, p^{(1)}, \ldots, p^{(n)}$ and $q, q^{(1)}, \ldots, q^{(n)}$ in Δ_k:*

1. (*Total variation and Hellinger distance.*) $d_{\mathrm{TV}}^2(p, q) \le d_{\mathrm{h}}^2(p, q) \le 2d_{\mathrm{TV}}(p, q)$.
2. (*Sub-additivity of total variation.*) $d_{\mathrm{TV}}\left(\prod_{i=1}^n p^{(i)}, \prod_{i=1}^n q^{(i)}\right) \le \sum_{i=1}^n d_{\mathrm{TV}}\left(p^{(i)}, q^{(i)}\right)$.
3. (*Hellinger tensorization.*) $\beta_{\mathrm{h}}\left(\prod_{i=1}^n p^{(i)}, \prod_{i=1}^n q^{(i)}\right) = \prod_{i=1}^n \beta_{\mathrm{h}}\left(p^{(i)}, q^{(i)}\right)$.
4. (*Data processing.*) *For any channel $\mathbf{T}$, f-divergence I_f, and pair of distributions (p, q), we have $I_f(\mathbf{T}p, \mathbf{T}q) \le I_f(p, q)$.*

We now define the simple hypothesis testing problem:

Problem 10.2.3 (Simple M-ary hypothesis testing)**.** *Given $\mathcal{P}$, a set of M distributions over $\mathcal{X}$, we say a function (test) $\phi : \cup_{n=1}^{\infty} \mathcal{X}^n \to \mathcal{P}$ solves the simple M-ary hypothesis testing problem with sample complexity n if*

$$\sum_{p \in \mathcal{P}} \mathop{\mathbb{P}}_{x \sim p^{\otimes n}} \{\phi(x) \neq p\} \le 0.1.$$

[14] We use the following conventions [Sas18]: $f(0) = \lim_{t \to 0^+} f(t)$, $0f(0/0) = 0$, and for $a > 0$, $0f(a/0) = a \lim_{u \to \infty} f(u)/u$.

We define the sample complexity of hypothesis testing to be the smallest n such that there exists a test ϕ which solves the hypothesis testing problem with sample complexity n. We use $\mathcal{B}(\mathcal{P})$ to denote the simple M-ary hypothesis testing problem and $n^*(\mathcal{P})$ to denote the sample complexity of $\mathcal{B}(\mathcal{P})$. When $M = 2$ and $\mathcal{P} = \{p, q\}$, we also use $\mathcal{B}(p, q)$ and $n^*(p, q)$ to denote the simple binary hypothesis testing problem and its sample complexity, respectively.

Fact 10.2.4 (Hypothesis testing and divergences [Yat85; DL01; CKMSU19; Wai19])**.** *We have the following:*

1. (*Total variation and binary hypothesis testing.*) *For any random variable Z over $\mathcal{Z}$ and test $\phi : \mathcal{Z} \to \{P, Q\}$, define the probability of error to be $\mathbb{P}_P(\phi(Z) = Q) + \mathbb{P}_Q(\phi(Z) = P)$. The minimum probability of error over all tests is $1 - d_{\mathrm{TV}}(P, Q)$ and is achieved by the following test: let $A^* \subseteq \mathcal{Z}$ be any set that maximizes $P(A) - Q(A)$ over $A \subseteq \mathcal{Z}$, and define $\phi(z) = P$ when $z \in A^*$ and $\phi(z) = Q$ otherwise.*

2. (*Hellinger distance and $\mathcal{B}(p, q)$.*) *The sample complexity for the simple binary hypothesis test between p and q is $\Theta\left(\frac{1}{d_{\mathrm{h}}^2(p,q)}\right)$, i.e., $n^*(p, q) = \Theta\left(\frac{1}{d_{\mathrm{h}}^2(p,q)}\right)$.*

3. (*Sample complexity of M-ary hypothesis testing.*) *Let $\mathcal{P}$ be a set of M distributions such that $\min_{p,q\in\mathcal{P}:p\neq q} d_{\mathrm{h}}(p, q) = \rho$. Then $\frac{1}{\rho^2} \lesssim n^*(\mathcal{P}) \lesssim \frac{\log M}{\rho^2}$.*

We now define Scheffe's test, which is commonly used for simple binary hypothesis testing.

Definition 10.2.5 (Scheffe's test)**.** *For two distributions p and q, consider the set $A = \{x : p(x) \geq q(x)\}$. Let p' and q' denote the distributions of $\mathbb{I}_A(X)$ when X is distributed as p and q, respectively. Given $(x_1, \ldots, x_n) \in \mathcal{X}^n$, Scheffe's test transforms each individual point x_i to $\mathbb{I}_A(x_i)$ and then applies the optimal test between p' and q' to the transformed points.*[15]

[15]Note that p' and q' are Bernoulli distributions with probabilities of observing 1 equal to $p(A)$ and $q(A)$, respectively. The optimal test between p' and q' corresponds to a threshold on $\sum_i \mathbb{I}_A(x_i)$.

It is easy to see that $d_{\mathrm{TV}}(p', q') = d_{\mathrm{TV}}(p, q)$, which implies that $d_{\mathrm{h}}(p', q') \geq 0.5 d_{\mathrm{h}}^2(p, q)$ (using Fact 10.2.2), leading to an $O\left(\frac{1}{d_{\mathrm{h}}^4(p,q)}\right)$ sample complexity of Scheffe's test. This dependence is tight [Sur21]. Formally, see Proposition G.1.2 in Appendix G.1.

10.2.2 Simple Hypothesis Testing under Communication Constraints

Let $\mathcal{X}$ be the domain, $\mathcal{P}$ a family of distributions over $\mathcal{X}$, and $\mathcal{T}$ a family of channels from $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{T}_D$ denote the set of all channels from $\mathcal{X}$ to $[0 : D-1]$. We first formally define the problem of simple hypothesis testing under communication constraints.

Definition 10.2.6 (Simple hypothesis testing under communication constraints)**.** *Let $\{U_i\}_{i=1}^n$ denote a set of n users who choose channels $\{\mathbf{T}_i\}_{i=1}^n \subseteq \mathcal{T}$ according to a rule $\mathcal{R} : [n] \to \mathcal{T}^n$.*[16] *Each user U_i then observes a random variable X_i i.i.d. from an (unknown) $p \in \mathcal{P}$, and generates $Y_i = \mathbf{T}_i(X_i) \in \mathcal{Y}$. The central server U_0 observes $(Y_1, \ldots, Y_n)$ and constructs an estimate $\widehat{p} = \phi(Y_1, \ldots, Y_n)$. We refer to this problem as simple hypothesis testing under communication constraints of $\mathcal{T}$ and denote it by $\mathcal{B}(\mathcal{P}, \mathcal{T})$. When $\mathcal{Y} = [0 : D-1]$ and $\mathcal{T} = \mathcal{T}_D$ for $D \geq 2$, we call $\mathcal{B}(\mathcal{P}, \mathcal{T}_D)$ the simple hypothesis testing problem under communication constraints of D-messages. When $\mathcal{P} = \{p, q\}$, we also use the notation $\mathcal{B}(p, q, \mathcal{T}_D)$.*

Definition 10.2.7 (Sample complexity of $\mathcal{B}(\mathcal{P}, \mathcal{T}_D)$)**.** *For a given test-rule pair $(\phi, \mathcal{R})$ with $\phi : \cup_{j=1}^{\infty} \mathcal{Y}^j \to \mathcal{P}$, we say that $(\phi, \mathcal{R})$ solves $\mathcal{B}(\mathcal{P}, \mathcal{T}_D)$ with sample complexity n if*

$$\sum_{p \in \mathcal{P}} \mathop{\mathbb{P}}_{(x_1, \ldots, x_n) \sim p^{\otimes n}} (\phi(y_1, \ldots, y_n) \neq p) \leq 0.1. \tag{10.3}$$

We use $n^*(\mathcal{P}, \mathcal{T}_D)$ to denote the sample complexity of this task, i.e., the smallest n so that there exists a $(\phi, \mathcal{R})$-pair that solves $\mathcal{B}(\mathcal{P}, \mathcal{T}_D)$. We use $n^*_{\texttt{identical}}(\mathcal{P}, \mathcal{T}_D)$ to denote the setting where each channel is identical, i.e., $\mathcal{R} : [n] \to \cup_{\mathbf{T} \in \mathcal{T}_D} \{\mathbf{T}\}^n$. In order to emphasize the setting

[16] It suffices to consider cases where the rule $\mathcal{R}$ is deterministic. Tsitsiklis [Tsi93, Proposition 2.1] implies that the sample complexity does not decrease even if the rule $\mathcal{R}$ is coordinated among users and randomized, such that the seed is not observed by the central server U_0.

where the channels need not be identical, we sometimes use $n^*_{\texttt{non-identical}}(\mathcal{P}, \mathcal{T}_D)$ *to denote* $n^*(\mathcal{P}, \mathcal{T}_D)$*. When* $\mathcal{P} = \{p, q\}$*, we will use the notation* $n^*(p, q, \mathcal{T}_D)$*,* $n^*_{\texttt{identical}}(p, q, \mathcal{T}_D)$*, and* $n^*_{\texttt{non-identical}}(p, q, \mathcal{T}_D)$*.*

We shall discuss the setting of adaptive channels in Section 10.5.

Special case: Binary hypothesis testing. In the rest of this section, we will focus on the special case when $\mathcal{P} = \{p, q\}$. For a fixed rule $\mathcal{R}$, an optimal ϕ corresponds to the likelihood ratio test. Thus, our focus will be on designing the rule $\mathcal{R}$, while choosing the test ϕ implicitly[17], such that the test-rule pair $(\phi, \mathcal{R})$ has minimal sample complexity.

A subset of channels called *threshold channels* plays a key role in our theory: Consider a set $\Gamma = \{\gamma_1, \ldots, \gamma_{D-1}\}$ such that $0 < \gamma_1 \leq \cdots \leq \gamma_{D-1} < \infty$. Let $\gamma_0 := 0$ and $\gamma_D := \infty$. Define the function $w_\Gamma : [k] \to [0 : D-1]$ as follows[18]: if $q(x) = 0$, then $w_\Gamma(x) = D - 1$; otherwise,

$$w_\Gamma(x) = j \text{ if and only if } p(x)/q(x) \in [\gamma_j, \gamma_{j+1}). \tag{10.4}$$

We are now ready to define a threshold test:

Definition 10.2.8 (Threshold test)**.** *We say that a channel* $\mathbf{T} \in \mathcal{T}_D$ *corresponds to a threshold test for two distributions* p *and* q *over* $[k]$ *if there exists* $\Gamma = \{\gamma_1, \ldots, \gamma_{D-1}\}$ *such that* $0 < \gamma_1 \leq \cdots \leq \gamma_{D-1} < \infty$*, and* $w_\Gamma(X) \sim \mathbf{T}\tilde{p}$ *whenever* $X \sim \tilde{p}$ *for any* $\tilde{p}$ *(cf. equation* (10.4)*). Any such* Γ *is called the set of thresholds of the test* $\mathbf{T}$*. We use* $\mathcal{T}_D^{\texttt{thresh}}$ *to denote the set of all channels* $\mathbf{T} \in \mathcal{T}_D$ *that correspond to threshold tests.*

Note that a priori, searching for an optimal channel over $\mathcal{T}_D^{\texttt{thresh}}$ seems to require $k^{\Omega(D)}$ time, as it requires searching over all possible values of Γ. By restricting our attention to a special class of thresholds parametrized by a single quantity, we will obtain a $\mathrm{poly}(k, D)$-time algorithm. In particular, we will focus on channels with thresholds in the following

[17] We will mention the test explicitly wherever required, e.g., in robust hypothesis testing.

[18] When $q(x) = 0$ for some x and $p(x) \neq 0$, we take $p(x)/q(x) = \infty$. Without loss of generality, we can assume that for each $x \in [k]$, at least one of $p(x)$ or $q(x)$ is non-zero.

set:

$$\mathcal{C} := \{\Gamma = (\gamma_1, \ldots, \gamma_{D-1}) : \forall j \in [D-2],\, \gamma_{j+1}/\gamma_j = 2\}\,. \tag{10.5}$$

A classical result states that threshold tests (cf. Definition 10.2.8) are optimal tests under communication constraints:

Theorem 10.2.9 ([Tsi93, Proposition 2.4])**.** $n^*_{\texttt{non-identical}}(p, q, \mathcal{T}_D^{\texttt{thresh}}) = n^*_{\texttt{non-identical}}(p, q, \mathcal{T}_D)$.

Our lower bounds on the sample complexity of hypothesis testing under communication constraints crucially rely on the optimality of threshold tests.

10.3 Reverse Data Processing Inequality for Quantized Channels

In this section, we state and prove a reverse data processing inequality for a class of f-divergences for communication-constrained channels. We begin by defining a suitable family of f-divergences:

Definition 10.3.1 (Well-behaved f-divergences)**.** *We say $I_f(\cdot, \cdot)$ is a well-behaved f-divergence if it satisfies the following:*

I.1 f is a convex nonnegative function with $f(1) = 0$.

I.2 $xf(y/x) = yf(x/y)$.[19]

I.3 There exist $\alpha > 0, \kappa > 0, C_1 > 0$, and $C_2 > 0$ such that for all $x \in [0, \kappa]$, we have[20]

$$C_1 x^\alpha \le f(1+x) \le C_2 x^\alpha.$$

[19] This implies $I_f(p, q) = I_f(q, p)$.

[20] The convexity and non-negativity of f means that α must be at least 1.

Some examples of well-behaved f-divergences include the total variation distance, squared Hellinger distance, symmetrized χ^2-divergence, symmetrized KL-divergence, and triangular discrimination (see Claim G.6.1 for more details). If f is differentiable at 1, $f'(1) = 0$, and the corresponding f-divergence is symmetric, then f satisfies I.2 [Gil06; Sas15]. Given an f-divergence that does not satisfy I.2, we can construct a new f-divergence with $\tilde{f}(x) := f(x) + xf(1/x)$, which is also a convex function[21] satisfying $\tilde{f}(1) = 0$ and I.2.

10.3.1 Main Result

The main result of this section is as follows:

Theorem 10.3.2 (Reverse data processing inequality)**.** *Let I_f be a well-behaved f-divergence with $(\alpha, \kappa, C_1, C_2)$ as defined in Definition 10.3.1. Let p and q be two fixed distributions over $[k]$ such that for all $i \in [k]$, we have $q_i \geq \nu p_i$ and $p_i \geq \nu q_i$, for some $\nu \in [0, 1]$. Then for any $D \geq 2$, there exists a channel $\mathbf{T}^* \in \mathcal{T}_D^{\texttt{thresh}}$ (and thus in $\mathcal{T}_D$) such that*

$$1 \leq \frac{I_f(p, q)}{I_f(\mathbf{T}^* p, \mathbf{T}^* q)} \leq 4\frac{f(\nu)}{f(1/(1+\kappa))} + \frac{52C_2}{C_1} \max\left\{1, \frac{R}{D}\right\}, \tag{10.6}$$

where $R = \min\{k, k'\}$ and $k' = 1 + \log\left(\frac{4C_2\kappa^\alpha}{I_f(p,q)}\right)$. Furthermore, given f, p, and q, there is a $\mathrm{poly}(k, D)$-time algorithm that finds a $\mathbf{T}^*$ achieving the rate in inequality (10.6).

Remark 10.3.3. *In the usual data processing inequality, the f-divergence $I_f(\mathbf{T}p, \mathbf{T}q)$ is upper-bounded by $I_f(p, q)$. Since the direction of the inequality is reversed in the second inequality in Theorem 10.3.2, we interpret it as a* reverse *data processing inequality. Another natural way to interpret this result is from the lens of quantization: Theorem 10.3.2 asserts that for any p, q, and any well-behaved f-divergence, there exist good quantization schemes to preserve the f-divergence.*

[21] This can be checked by noting that $\tilde{f}''(x) = f''(x) + \frac{1}{x^3} f''(x)$, which is non-negative, as f is convex.

We provide a brief proof sketch for the special case of the Hellinger distance and $D = 2$ in Section 10.3.2, and defer the full proof to Appendix G.2.1. As our main focus will be on the Hellinger distance, we state the following corollary, which will be used later:

Corollary 10.3.4 (Preservation of Hellinger distance)**.** *For any $p, q \in \Delta_k$ and $D \geq 2$, there exists a $\mathbf{T}^* \in \mathcal{T}_D^{\texttt{thresh}}$ such that the following holds:*

$$1 \leq \frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{h}}^2(\mathbf{T}^* p, \mathbf{T}^* q)} \leq 1800 \max\left\{1, \frac{\min\{k, k'\}}{D}\right\}, \tag{10.7}$$

where $k' = \log(4/d_{\mathrm{h}}^2(p,q))$. Given p and q, there is a $\mathrm{poly}(k, D)$-time algorithm that finds $\mathbf{T}^*$ achieving the rate in inequality (10.7).

Proof. The desired bound follows by noting that $f(x) = (\sqrt{x} - 1)^2$ for the Hellinger distance and taking $\nu = 0$. As shown in Appendix G.6 (Claim G.6.1), we can take $\kappa = 1$, $C_1 = 2^{-3.5}$, $C_2 = 1$, and $\alpha = 2$. Note that $f(0) = 1$ and $f(1/(1+\kappa)) = (\sqrt{2} - 1)^2/2 \geq 0.04$. This suffices to give a guarantee of $\frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{h}}^2(\mathbf{T}^* p, \mathbf{T}^* q)} \leq 100 + \frac{900}{D}(\min\{k, k'\})$ from Theorem 10.3.2. □

Remark 10.3.5 (Dimensionality reduction using channels)**.** *Corollary 10.3.4 can be interpreted as saying that the effective support size of p and q for the Hellinger distance is at most $k' := \log(4/d_{\mathrm{h}}^2(p,q))$, because the distributions could be mapped to a k'-sized alphabet using a channel in a manner that preserves the pairwise Hellinger distance up to constant terms. We also remark that our notion of dimensionality reduction requires the transformation to be performed using a channel, which is fundamentally different from the setting in Abdullah, Kumar, McGregor, Vassilvitskii, and Venkatasubramanian [AKMVV16].*

The following result states that the bound in Corollary 10.3.4 is tight:

Lemma 10.3.6 (Reverse data processing is tight)**.** *There exist positive constants c_1, c_2, c_3, c_4, c_5, and c_6 such that for every $\rho \in (0, c_1)$ and $D \geq 2$, there exist $k \in [c_2 \log(1/\rho), c_3 \log(1/\rho)]$ and two distributions p and q on $[k]$ such that $d_{\mathrm{h}}^2(p, q) \in [c_4\rho, c_5\rho]$ and*

$$\inf_{\mathbf{T} \in \mathcal{T}_D^{\mathtt{thresh}}} \frac{d_{\mathrm{h}}^2(p, q)}{d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)} \geq c_6 \cdot \frac{R'}{D}, \tag{10.8}$$

where $R' = \max\{k, k'\}$ and $k' = \log(1/\rho)$. Thus, $R' = \Theta(k) = \Theta(\log(1/\rho))$.

The proof of Lemma 10.3.6 is given in Appendix G.2.2.

10.3.2 Proof Sketch of Theorem 10.3.2 and Corollary 10.3.4

We will focus on the case of the Hellinger distance and $D = 2$. The first step is to establish the following result:

Lemma 10.3.7 (Reverse Markov inequality)**.** *Let X be a random variable over $[0, 1)$, supported on at most k points, with $\mathbb{E}[X] > 0$. Let $k' = 1 + \log(1/\mathbb{E}[X])$. Then*

$$\sup_{\delta \in [0,1)} \delta \, \mathbb{P}(X \geq \delta) \geq \frac{\mathbb{E}[X]}{13R}, \tag{10.9}$$

where $R = \min\{k, k'\}$.

The canonical version of Markov's inequality states that $\delta \, \mathbb{P}(X \geq \delta) \leq \mathbb{E}[X]$ for any non-negative random variable X and any δ. Since the direction of the inequality is reversed in Lemma 10.3.7 (up to a shrinkage factor of roughly $\log(1/\mathbb{E}[X])$), we call it a *reverse* Markov inequality. A generalized version of Lemma 10.3.7 for the case $D > 2$, along with its proof, is given in Lemma G.2.1.

Remark 10.3.8. *Note that Lemma 10.3.7 is tight, as shown in Claim G.2.4, which is crucially used in the proof of Lemma 10.3.6.*

Remark 10.3.9. *It is instructive to compare the guarantee of Lemma 10.3.7 with existing results in the literature:*

1. *The Paley-Zygmund inequality [dG99, Corollary 3.3.2] states that for any $\delta \in (0, \mathbb{E}[X])$, we have*
$$\mathbb{P}(X \geq \delta) \geq \left(1 - \frac{\delta}{\mathbb{E}[X]}\right)^2 \frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]}.$$
Multiplying both sides by δ and optimizing the lower bound over δ (achieved at $\delta = \mathbb{E}[X]/3$) yields
$$\sup_{\delta \geq 0} \delta\, \mathbb{P}(X \geq \delta) \gtrsim \mathbb{E}[X] \cdot \frac{1}{\mathbb{E}[X^2]/(\mathbb{E}[X])^2}.$$
Note that the shrinkage factor is $\mathbb{E}[X^2]/(\mathbb{E}[X])^2$, which is at most $1/\mathbb{E}[X]$, but could be exponentially larger than the factor $\log(1/\mathbb{E}[X])$ provided in Lemma 10.3.7. (For example, consider a random variable with $\mathbb{P}(X = 0) = 1 - p$ and $\mathbb{P}(X = 1/2) = p$: We have $\mathbb{E}[X^2]/(\mathbb{E}[X])^2 = 1/p$, whereas $\log(1/\mathbb{E}[X]) = \log(2/p)$.)

2. *A standard version of the reverse Markov inequality [SB14, Lemma B.1] for a random variable bounded in $[0, 1]$ states that $\mathbb{P}(X \geq \delta) \geq \frac{\mathbb{E}[X]-\delta}{1-\delta}$, for $\delta \in (0, \mathbb{E}[X])$. Multiplying both sides by δ and optimizing the bound over $\delta \in (0, \mathbb{E}[X])$, under the condition that $\mathbb{E}[X] \leq 0.1$, gives us the following:*
$$\sup_{\delta \geq 0} \delta\, \mathbb{P}(X \geq \delta) \gtrsim (\mathbb{E}[X])^2 = \mathbb{E}[X] \cdot \frac{1}{1/\mathbb{E}[X]},$$
i.e., the shrinkage factor is $1/\mathbb{E}[X]$, which is again exponentially larger than $\log(1/\mathbb{E}[X])$.

Using Lemma 10.3.7, we now sketch the proof that there exists a channel $\mathbf{T} \in \mathcal{T}_2^{\texttt{thresh}}$ achieving $d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \gtrsim d_{\mathrm{h}}^2(p, q)/R$. For simplicity of notation, we assume that for all

$i \in [k]$, we have $p_i > 0$ and $q_i > 0$. We first define the sets

$$A_{l,u} = \left\{ i \in [k] : \frac{p_i}{q_i} \in [l, u) \right\},$$
$$A_{l,\infty} = \left\{ i \in [k] : \frac{p_i}{q_i} \in [l, \infty] \right\}. \tag{10.10}$$

Then $d_{\mathrm{h}}^2(p, q)$ can be decomposed as follows:

$$d_{\mathrm{h}}^2(p,q) = \sum_{i \in A_{0,1/2}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 + \sum_{i \in A_{1/2,1}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 + \sum_{i \in A_{1,2}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 + \sum_{i \in A_{2,\infty}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2.$$

We note that at least one of these terms must be at least $d_{\mathrm{h}}^2(p, q)/4$. By symmetry, it suffices to consider the cases where the sum over $A_{2,\infty}$ is at least $d_{\mathrm{h}}^2(p, q)/4$, or the sum over $A_{1,2}$ is at least $d_{\mathrm{h}}^2(p, q)/4$.

Case 1**:** $\sum_{i \in A_{2,\infty}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 \geq d_{\mathrm{h}}^2(p, q)/4$**.** Let $\mathbf{T} \in \mathcal{T}_2^{\texttt{thresh}}$ be a threshold test with threshold $\Gamma = \{2\}$, i.e., $\mathbf{T}$ is a deterministic channel that corresponds to the function $i \mapsto \mathbb{I}_{p_i/q_i \geq 2}$. We note that $\mathbf{T}p$ and $\mathbf{T}q$ are binary distributions, characterized by $p' = \sum_{i \in A_{2,\infty}} p_i$ and $q' = \sum_{i \in A_{2,\infty}} q_i$, respectively. Then

$$d_{\mathrm{h}}^2(p, q) \leq 4 \sum_{i \in A_{2,\infty}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 \leq 4 \sum_{i \in A_{2,\infty}} p_i = 4p',$$

where the first inequality uses the assumption. Using the fact that $p' \geq 2q'$, we also have

$$d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \geq \left(\sqrt{p'} - \sqrt{q'}\right)^2 \geq \left(\sqrt{p'} - \sqrt{p'/2}\right)^2 \geq 0.01p'.$$

Combining the two displayed equations, we obtain $d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \geq d_{\mathrm{h}}^2(p, q)/400$. This completes the proof.

Case 2: $\sum_{i\in A_{1,2}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 \geq d_{\mathrm{h}}^2(p,q)/4$**.** For $i \in A_{1,2}$, let $\delta_i := (p_i - q_i)/q_i$, which lies in $[0,1)$. Consider the random variable X over $[0,1)$ such that for $i \in A_{1,2}$, we define $\mathbb{P}(X = \delta_i) = q_i$ and $\mathbb{P}(X = 0) = 1 - \sum_{i\in A_{1,2}} q_i$. Let $\delta \in [0,1)$ be arbitrary (to be decided later). Consider the channel $\mathbf{T}$ corresponding to the threshold $1+\delta$. Suppose for now that the following inequalities hold:

$$d_{\mathrm{h}}^2(p,q) \lesssim \mathbb{E}\, X^2 \quad \text{and} \quad d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \gtrsim \delta^2\, \mathbb{P}(X \geq \delta), \tag{10.11}$$

which we will establish shortly using a Taylor approximation. Letting $Y = X^2$ and $\delta' = \delta^2$, we obtain the following inequality using the bounds (10.11):

$$\frac{d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)}{d_{\mathrm{h}}^2(p,q)} \gtrsim \frac{\delta^2\, \mathbb{P}(X \geq \delta)}{\mathbb{E}[X^2]} = \frac{\delta'\, \mathbb{P}(Y \geq \delta')}{\mathbb{E}[Y]}. \tag{10.12}$$

Fix

$$R = \log(1/\,\mathbb{E}[Y]) = \log(1/\,\mathbb{E}[X^2]) = \log(O(1/d_{\mathrm{h}}^2(p,q)).$$

By Lemma 10.3.7, we note that there exists δ' (and therefore also δ) such that $\delta'\, \mathbb{P}(Y \geq \delta') \gtrsim \mathbb{E}[Y]/R$, which yields the desired lower bound $\frac{d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)}{d_{\mathrm{h}}^2(p,q)} \gtrsim \frac{1}{R}$ using inequality (10.12).

We now provide a brief proof sketch of the bounds (10.11). We derive the first bound using the following arguments:

$$d_{\mathrm{h}}^2(p,q) \leq 4 \sum_{i\in A_{1,2}} (\sqrt{p_i} - \sqrt{q_i})^2 = 4 \sum_{i\in A_{1,2}} q_i(\sqrt{1+\delta_i} - 1)^2 \leq 4 \sum_{i\in A_{1,2}} q_i \delta_i^2 = 4\, \mathbb{E}[X^2],$$

where the first inequality uses the assumption and the second inequality uses the fact that $\sqrt{1+x} \leq 1 + x$ for $x \geq 0$.

We now turn our attention to the second bound (10.11). Recall that $\mathbf{T}$ is a channel corresponding to the threshold $1+\delta$. Let $p' = \sum_{i:\delta_i \in [\delta,1)} p_i$ and $q' = \sum_{i:\delta_i\in[\delta,1)} q_i$. Note

that $q' = \mathbb{P}(X \geq \delta)$ and $p' - q' = \sum_{i:\delta_i \in [\delta,1)} \delta_i q_i = \mathbb{E}[X\, \mathbb{I}_{X \geq \delta}]$. Thus, we have $(p' - q')/q' = \mathbb{E}[X | X \geq \delta] \geq \delta$.

It can be shown that $d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \geq (\sqrt{p'} - \sqrt{q'})^2$ (cf. inequality (G.11) in Appendix G.2), which leads to the following inequalities:

$$d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \geq (\sqrt{p'} - \sqrt{q'})^2 = q' \left(\sqrt{1 + \frac{p' - q'}{q'}} - 1 \right)^2 \gtrsim q' \left(\frac{p' - q'}{q'} \right)^2 \geq \delta^2\, \mathbb{P}(X \geq \delta),$$

as follows from studying the function $x \mapsto (\sqrt{1+x} - 1)/x$ on $(0, 1]$, extended by continuity to $[0, 1]$.

10.4 Simple Binary Hypothesis Testing

We will now apply the results from the previous sections to simple binary hypothesis testing under communication constraints. Let $\mathbf{T}$ be a fixed channel, and suppose all users use the same channel $\mathbf{T}$. In this setting, Fact 10.2.4 implies that the sample complexity is $\Theta(1/(d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)))$. Without any communication constraints, the sample complexity of the best test is known to be $\Theta(1/(d_{\mathrm{h}}^2(p, q)))$. Thus, the additional (multiplicative) penalty of using the channel $\mathbf{T}$ is $\frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)}$, which is at least 1, by the data processing inequality. As we are allowed to choose any channel $\mathbf{T} \in \mathcal{T}_D$, we would like to choose the channel that minimizes this ratio, which was precisely studied in Section 10.3.

10.4.1 Upper Bound

We begin with an upper bound, which follows directly from Corollary 10.3.4.

Theorem 10.4.1. *There exists a positive constant c satisfying the following: for any $k \in \mathbb{N}$, let p and q be two distributions on Δ_k and define $n^* := n^*(p, q)$. For any $D \geq 2$, the sample*

complexity of simple binary hypothesis testing with identical channels satisfies

$$n^*_{\texttt{identical}}(p, q, \mathcal{T}_D) \leq c \cdot n^* \cdot \max\left\{1, \frac{\min\{k, \log n^*\}}{D}\right\}. \tag{10.13}$$

Furthermore, there is an algorithm which, given p, q, and D, finds a channel $\mathbf{T}^* \in \mathcal{T}_D^{\texttt{thresh}}$ in $\mathrm{poly}(k, D)$ *time that achieves the rate in inequality* (10.13).

Proof. As noted earlier, for a fixed $\mathbf{T}$, the sample complexity is

$$\Theta\left(\frac{1}{d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)}\right) = \Theta\left(\frac{1}{d_{\mathrm{h}}^2(p, q)} \cdot \frac{d_{\mathrm{h}}^2(p, q)}{d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)}\right) = \Theta\left(n^* \cdot g(\mathbf{T})\right),$$

where $g(\mathbf{T}) := \frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{h}}^2(\mathbf{T}p,\mathbf{T}q)}$. Our proof strategy will be to upper-bound the quantity $\inf_{\mathbf{T}\in\mathcal{T}_D} g(\mathbf{T})$. By Corollary 10.3.4, there exists a $\mathbf{T}^*$ such that $g(\mathbf{T}^*) \lesssim \max\{1, \min\{k, \log(n^*)\}/D\}$, since $n^* = \Theta(1/d_{\mathrm{h}}^2(p, q))$ by Fact 10.2.4. Thus, the proof of Theorem 10.4.1 follows from Corollary 10.3.4 by choosing the optimal $\mathbf{T}^*$ achieving the bound in Corollary 10.3.4. As mentioned in Corollary 10.3.4, the channel $\mathbf{T}^*$ can be found efficiently. □

10.4.2 Lower Bound

We now prove a lower bound, showing that there exist distributions p and q such that the upper bound in Theorem 10.4.1 is tight. As discussed in Section 10.2.2, an optimal test that minimizes the probability of error under communication constraints is a threshold test based on $\frac{p(i)}{q(i)}$ [Tsi93]. However, this notion of optimality is conditioned on the fact that the channels are potentially non-identical; examples exist where such a condition is necessary even for $D = 2, M = 2$, and $n = 2$ [Tsi88].

We will show that, up to constants in the sample complexity, it suffices to consider identical channels for simple hypothesis testing. In fact, we prove a much more general result below that does not rely on restricting the function class to threshold tests.

Lemma 10.4.2 (Equivalence between identical and non-identical channels for simple hypothesis testing)**.** *Let $\mathcal{T}$ be a collection of channels from $\mathcal{X} \to \mathcal{Y}$. Let p and q be two distributions on $\mathcal{X}$. Then*

$$n^*_{\texttt{non-identical}}(p, q, \mathcal{T}) = \Theta\left(n^*_{\texttt{identical}}(p, q, \mathcal{T})\right).$$

The proof of Lemma 10.4.2 is provided in Appendix G.3.1. We now derive the following lower bound on $n^*_{\texttt{non-identical}}(p, q, \mathcal{T})$:

Theorem 10.4.3. *There exist positive constants c_1 and c_2 such that for every $n_0 \in \mathbb{N}$ and $D \geq 2$, there exist (i) $k = \Theta(\log n_0)$ and (ii) two distributions p and q on $[k]$, such that the following hold:*

1. *$c_1 n_0 \leq n^*(p, q) \leq c_2 n_0$, and*
2. *$n^*_{\texttt{non-identical}}(p, q, \mathcal{T}_D) \geq \frac{n_0 \log(n_0)}{D}$.*

Proof. We first provide a proof sketch. Using Lemma 10.4.2 and Theorem 10.2.9, it suffices to consider the setting with identical threshold channels. With identical channels, say $\mathbf{T}$, the problem reduces to that of $\mathcal{B}(\mathbf{T}p, \mathbf{T}q)$, and thus to bounding $d_{\mathrm{h}}(\mathbf{T}p, \mathbf{T}q)$, using Fact 10.2.4. Tightness of Lemma 10.3.6 then gives the desired result.

Turning to the details, note that it suffices to consider $D \leq \log n_0$. Moreover, we can consider the setting where n_0 is sufficiently large; the result for general n_0 then follows by changing the constants c_1 and c_2 appropriately.

Now define $\rho = 1/n_0$. Since n_0 is large enough, this ρ satisfies the condition of Lemma 10.3.6. Let p, q, and $k = \Theta(\log(1/\rho))$ be from Lemma 10.3.6, such that (i) $d^2_{\mathrm{h}}(p, q) = \Theta(\rho)$ and (ii) inequality (10.8) holds. Using Fact 10.2.4, we have the following:

$$n^*(p, q) = \Theta\left(\frac{1}{d^2_{\mathrm{h}}(p, q)}\right) = \Theta\left(\frac{1}{\rho}\right) = \Theta\left(n_0\right). \tag{10.14}$$

Thus, p and q satisfy the first condition of the theorem. Furthermore, we have the following (where c' represents a positive constant which may change from line to line):

$$
\begin{aligned}
n^*_{\texttt{non-identical}}(p, q, \mathcal{T}_D) &= n^*_{\texttt{non-identical}}(p, q, \mathcal{T}_D^{\texttt{thresh}}) && \text{(using Theorem 10.2.9)} \\
&\geq c' n^*_{\texttt{identical}}(p, q, \mathcal{T}_D^{\texttt{thresh}}) && \text{(using Lemma 10.4.2)} \\
&\geq c' \inf_{\mathbf{T} \in \mathcal{T}_D^{\texttt{thresh}}} \frac{1}{d^2_{\mathrm{h}}(\mathbf{T}p, \mathbf{T}q)} && \text{(using Fact 10.2.4)} \\
&= \frac{c'}{d^2_{\mathrm{h}}(p, q)} \inf_{\mathbf{T} \in \mathcal{T}_D^{\texttt{thresh}}} \frac{d^2_{\mathrm{h}}(p, q)}{d^2_{\mathrm{h}}(\mathbf{T}p, \mathbf{T}q)} \\
&\geq \frac{c'}{d^2_{\mathrm{h}}(p, q)} \frac{\log(1/d^2_{\mathrm{h}}(p, q))}{D} && \text{(using Lemma 10.3.6)} \\
&\geq c' n^*(p, q) \frac{\log(c'' n^*(p, q))}{D} && \text{(using equation (10.14))} \\
&\geq n_0 \frac{\log(n_0)}{D} && \text{(using the bounds on } n^*(p, q)) \,.
\end{aligned}
$$

This completes the proof. □

10.4.3 Robust Tests

In this section, we study the robust version of $\mathcal{B}(p, q, \mathcal{T}_D)$. Here, the data-generating distribution may not belong to $\mathcal{P} := \{p, q\}$, but is only guaranteed to lie within a certain radius of an element of $\mathcal{P}$. Our main result (Theorem 10.4.6) shows that communication constraints increase the sample complexity of the robust version of hypothesis testing by at most logarithmic factors.

We begin by formally defining the robust version of simple hypothesis testing under communication constraints:

Definition 10.4.4 (Robust version of $\mathcal{B}(p, q, \mathcal{T}_D)$)**.** *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be defined as $\mathcal{P}_1 := \{\tilde{p} : d_{\mathrm{TV}}(p, \tilde{p}) \leq \epsilon\}$ and $\mathcal{P}_2 := \{\tilde{q} : d_{\mathrm{TV}}(q, \tilde{q}) \leq \epsilon\}$. The robust version of $\mathcal{B}(p, q, \mathcal{T})$, denoted by $\mathcal{B}_{\texttt{robust}}(p, q, \epsilon, \mathcal{T})$, is defined as in Definition 10.2.6, but with $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$. For a given test-rule pair $(\phi, \mathcal{R})$ with $\phi : \cup_{j=1}^{\infty} \mathcal{Y}^j \to \mathcal{P}$, we say that $(\phi, \mathcal{R})$ solves $\mathcal{B}_{\texttt{robust}}(p, q, \epsilon, \mathcal{T})$ with sample*

complexity n if

$$\sup_{\tilde{p}\in\mathcal{P}_1} \mathop{\mathbb{P}}_{(x_1,\ldots,x_n)\sim\tilde{p}^{\otimes n}}(\phi(y_1,\ldots,y_n)\neq p) + \sup_{\tilde{q}\in\mathcal{P}_2} \mathop{\mathbb{P}}_{(x_1,\ldots,x_n)\sim\tilde{q}^{\otimes n}}(\phi(y_1,\ldots,y_n)\neq q) \leq 0.1. \quad (10.15)$$

We use $n^*_{\texttt{robust}}(p,q,\epsilon,\mathcal{T})$ to denote the sample complexity of this task, i.e., the smallest n so that there exists a $(\phi,\mathcal{R})$-pair that solves $\mathcal{B}_{\texttt{robust}}(p,q,\epsilon,\mathcal{T})$ with sample complexity n. We use $\mathcal{B}(p,q,\epsilon)$ and $n^*_{\texttt{robust}}(p,q,\epsilon)$ to denote, respectively, the robust hypothesis testing problem and the sample complexity of robust testing in the absence of any channel constraints.

For any two distributions p and q with $d_{\mathrm{TV}}(p,q) = 3\epsilon$, it is possible to obtain an ϵ-robust test with sample complexity $O(1/\epsilon^2)$ (e.g., using Scheffe's test). However, as the following example shows, the optimal communication-efficient channel for $\mathcal{B}(p,q,\mathcal{T}_D)$ may not be robust to $\epsilon^{1+\alpha}$-contamination for any $\alpha\in[0,1)$ (see Appendix G.3.2 for more details).

Example 10.4.5 (Optimal channel may not be robust)**.** *Let $\alpha\in(0,1)$ and $\epsilon>0$ be small enough. Let $p\in\Delta_3$ be the distribution $(0.5-3\epsilon-\epsilon^{1+\alpha}, 0.5+3\epsilon, \epsilon^{1+\alpha})$, and let $q\in\Delta_3$ be the distribution $(0.5,0.5,0)$. Then $d_{\mathrm{TV}}(p,q)\geq 3\epsilon$ and $n^*_{\texttt{robust}}(p,q,\epsilon)=\Theta(1/\epsilon^2)$. However, the optimal[22] channel $\mathbf{T}^*$ for $\mathcal{B}(p,q,\mathcal{T}_2)$ is not robust to $\epsilon^{1+\alpha}$-corruption: there exists $\tilde{p}$, satisfying $d_{\mathrm{TV}}(p,\tilde{p})\leq\epsilon^{1+\alpha}$, such that $\mathbf{T}^*\tilde{p}=\mathbf{T}^*q$.*

As our main result, we show that there is an (efficient) way to choose channels such that the sample complexity increases by at most a logarithmic factor:

Theorem 10.4.6 (Sample complexity of $\mathcal{B}_{\texttt{robust}}(p,q,\mathcal{T}_D)$)**.** *There exists a constant $c>0$ such that for any $p,q\in\Delta_k$ with $\epsilon<\frac{d_{\mathrm{TV}}(p,q)}{2}$ and any $D\geq 2$, we have*

$$n^*_{\texttt{robust}}(p,q,\epsilon,\mathcal{T}_D) \leq c\cdot n^*\cdot\max\left\{1,\frac{\min\{k,\log n^*\}}{D}\right\}, \quad (10.16)$$

[22]The channel corresponds to the function $\mathbb{I}_{\{3\}}(x)$, and it transforms p and q to the distributions $(1-\epsilon^{1+\alpha},\epsilon^{1+\alpha})$ and $(1,0)$, respectively.

where $n^* := n^*_{\texttt{robust}}(p, q, \epsilon)$. *Furthermore, there is an algorithm which, given* p, q, ϵ, *and* D, *finds a channel* $\mathbf{T}^* \in \mathcal{T}_D^{\texttt{thresh}}$ *in* $\mathrm{poly}(k, D)$ *time that achieves the rate in inequality* (10.16).

Note that the optimal channel in Theorem 10.4.6 may depend on ϵ. Our proof critically uses the framework of least favorable distributions (LFDs) for binary hypothesis testing, pioneered by Huber [Hub65]. LFDs are pairs of distributions $\tilde{p} \in \mathcal{P}_1$ and $\tilde{q} \in \mathcal{P}_2$ that maximize $\inf_\phi f(\phi, \tilde{p}, \tilde{q}, n)$, where $f(\phi, \tilde{p}, \tilde{q}, n)$ is the probability of error of a test ϕ which distinguishes $\tilde{p}$ and $\tilde{q}$ based on n samples. Remarkably, LFDs do not depend on n when $\mathcal{P}_1$ and $\mathcal{P}_2$ are ϵ-balls around p and q, respectively, in the total variation distance [Hub65; HS73]. Moreover, these LFDs can be constructed algorithmically. Particularly relevant for us is the result of Veeravalli et al. [VBP94], who extended these results in the presence of communication constraints. We achieve Theorem 10.4.6 by applying Corollary 10.3.4 to $\tilde{p}$ and $\tilde{q}$, the LFDs under ϵ-contamination. See Appendix G.3.2 for more details.

As the following remark shows, the sample complexity of robust testing crucially depends on ϵ:

Remark 10.4.7 (Sample complexity without communication constraints)**.** *The sample complexity* $n^*_{\texttt{robust}}(p, q, \epsilon')$ *may have phase transitions with respect to* ϵ'. *For example,* $n^*_{\texttt{robust}}(p, q, \epsilon')$ *in Example 10.4.5 satisfies* $n^*_{\texttt{robust}}(p, q, \epsilon^{1+\beta}) = \Theta(1/\epsilon^{1+\alpha})$ *for* $\beta > \alpha$ (*small corruption*) *and* $\Theta(1/\epsilon^2)$ *for* $\beta \in [0, \alpha)$ (*large corruption*). *See Example G.3.1 for another instance.*

It is instructive to compare the guarantees of Theorem 10.4.6 with the sample complexity of Scheffe's test: Examples 10.4.5 and G.3.1 show that Scheffe's test may be strictly suboptimal in some regimes.

Finally, we present the following result, proved in Appendix G.3.2, showing that the channels in Theorem 10.4.1 are moderately robust:

Proposition 10.4.8 (Optimal channels are moderately robust)**.** *Let* p *and* q *be two distributions over* $[k]$. *Define* $\epsilon_0 := cd^2_{\mathrm{TV}}(p, q) \cdot \min\left\{1, \frac{D}{\log(1/d_{\mathrm{TV}}(p,q))}\right\}$ *for a small enough constant* c.[23]

[23] This upper bound can be generalized to $\epsilon_0 := cd^2_{\mathrm{h}}(\mathbf{T}^* p, \mathbf{T}^* q)$.

Let $\mathbf{T}^*$ be a channel that maximizes $d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)$ over $\mathbf{T} \in \mathcal{T}_D$. Let n^*_D be the sample complexity of $\mathbf{T}^*$ for p and q (recall that $n^*_D = \Theta(n^*(p, q, \mathcal{T}_D))$). Let ϕ^* be the corresponding optimal test.[24] *Then there exists a test ϕ' that uses $\mathbf{T}^*$ for each user and solves $\mathcal{B}_{\mathtt{robust}}(p, q, \epsilon_0, \mathcal{T}_D)$ with sample complexity $\Theta(n^*_D)$.*

Proposition 10.4.8 implies that optimal communication channels are already $\Theta(\epsilon^2)$-robust up to logarithmic factors. However, the result falls short of our desired goal of designing an $\Theta(\epsilon)$-robust test. (Informally, we say a channel is ϵ'-robust if it can be used to perform hypothesis testing with reasonable sample complexity despite ϵ'-corruption.) This guarantee is roughly the best possible, as can be seen by taking $\alpha \to 1$ in Example 10.4.5.

10.5 Simple M-ary Hypothesis Testing

In this section, we study the M-wise simple hypothesis testing problem, i.e., $\mathcal{P}$ is a set of $M \geq 2$ distributions. Our focus in this section will be slightly different from that of Section 10.4 in the following ways: (i) in addition to the choices of identical or non-identical channels, we will also allow channels to be selected adaptively; and (ii) our primary focus will be on studying the effect of M, the number of hypotheses, instead of the pairwise distance, i.e., $\min_{p,q \in \mathcal{P}: p \neq q} d_{\mathrm{h}}(p, q)$.

Definition 10.5.1 (Sequentially adaptive channels)**.** *Let $\mathcal{X}$ be the domain, $\mathcal{P}$ a family of distributions over $\mathcal{X}$, and $\mathcal{T}$ a family of channels from $\mathcal{X}$ to $\mathcal{Y}$. Let $(U_1, \ldots, U_n)$ denote n (ordered) users. Each user U_i observes a random variable X_i i.i.d. from an (unknown) $p' \in \mathcal{P}$. The observations are then released sequentially, as follows: for each time $i \in [n]$, user U_i first selects a channel $\mathbf{T}_i \in \mathcal{T}$ based on X_i (personal sample) and $(Y_1, \ldots, Y_{i-1})$ (public knowledge up to now), generates $Y_i = \mathbf{T}_i(X_i)$, and finally releases Y_i to everyone. The central server U_0*

[24] The optimal test corresponds to a likelihood ratio test between $\mathbf{T}^* p$ and $\mathbf{T}^* q$.

observes $Y_1, \ldots, Y_n$ and constructs an estimate $\widehat{p} = \phi(Y_1, \ldots, Y_n)$. Both the hypothesis testing task and its sample complexity are defined analogously to Definitions 10.2.6 and 10.2.7. When $\mathcal{T}$ is the set of channels that map to D alphabets, i.e., $\mathcal{T} = \mathcal{T}_D$, we denote the sample complexity by $n^*_{\texttt{adaptive}}(\mathcal{P}, \mathcal{T}_D)$.

10.5.1 Upper Bounds

In the following result, we show using a standard argument that we can use a communication-efficient binary test from Theorem 10.4.1 as a subroutine to solve the M-wise hypothesis testing problem.

Proposition 10.5.2 (Upper bounds using threshold tests)**.** *Let $\mathcal{P}$ be set of M distributions in Δ_k such that $\rho = \min_{p,q \in \mathcal{P}: p \neq q} d_{\mathrm{h}}(p, q)$. Let $k' = \log(1/\rho)$ and define the blow-up factor $R := \frac{\min\{k, \log(1/\rho)\}}{D} + 1$. Then the sample complexity of the simple M-ary hypothesis testing problem satisfies the bounds*

1. $n^*_{\texttt{non-identical}}(\mathcal{P}, \mathcal{T}_D) \lesssim \frac{M^2 \log M}{\rho^2} \cdot R$,
2. $n^*_{\texttt{adaptive}}(\mathcal{P}, \mathcal{T}_D) \lesssim \frac{M \log M}{\rho^2} \cdot R$.

The proof of Proposition 10.5.2, which is provided in Appendix G.4.1, proceeds by analyzing a standard tournament procedure, which we now briefly describe. We think of each hypothesis as a player, and each hypothesis test between any two distributions (players) as a game. The tournament procedure decides the fixtures of the games (which distributions will play against each other) and the overall winner of the tournament (the hypothesis that will be returned) based on the results of individual games. It is easy to see that the true distribution $p \in \mathcal{P}$ will never lose a game against any of the competitors, with high probability. Thus, as long as we have a unique player who has not lost a single game, we can confidently choose it to be the winner. An obvious strategy is to organize all pairwise $\Theta(M^2)$ tests and output the player who never loses

a game. Sans communication constraints, each game (hypothesis test) can be played with the same set of samples, and since the failure probability is exponentially small, it suffices to take $O(\log M)$ samples so that the results of all the games involving player p are correct. However, under communication constraints, we observe the samples only after they have passed through a channel. Furthermore, the channel that would ideally be employed for the game between p and q crucially relies on p and q, and it is unclear if the same channel provides useful information for the game between p and another player q'. We can circumvent this obstacle by using a new channel and a fresh set of samples for each game, which guarantees correctness after taking $O(M^2 \log M)$ samples. Recall that the choice of channels was non-adaptive here—when the channels can be adaptive, we can reduce the number of games (and thus the sample complexity) by organizing a "knock-out" style tournament of $M - 1$ games, where each losing player is discarded from the tournament.

Remark 10.5.3 (Dependence on M and ρ)**.** *We now comment on the dependence of these bounds on M and ρ. In particular, note that:*

1. *As shown later in Theorem 10.5.8, the dependence on M is nearly tight (up to logarithmic factors) for the case of adaptive algorithms (for constant ρ and D).*

2. *For non-adaptive algorithms and $D = 2$, Theorem 10.5.10 shows a lower bound of $\Omega(M^2)$ for the case of identical channels.*

3. *The dependence on ρ is tight for constant M (Theorem 10.4.3).*

Remark 10.5.4 (Robust M-ary hypothesis testing)**.** *One can also consider a robust version of simple M-ary hypothesis testing, which is often called hypothesis selection, i.e., the true distribution p satisfies $\min_{q \in \mathcal{P}} d_{\mathrm{TV}}(p, q) \leq \epsilon$ (analogous to Definition 10.4.4). One can use the robust test for binary hypothesis testing from Section 10.4.3 to obtain a similar dependence on M as in Proposition 10.5.2 under this setting.*

As the dependence on M is nearly tight for adaptive channels (for constant ρ), we now shift our attention to procedures that use identical channels, which might be desirable in certain practical situations. We establish the following bound for identical channels:

Theorem 10.5.5 (Upper bounds with identical channels)**.** *Let $\mathcal{P}$ be a set of M distributions in Δ_k satisfying $\min_{p,q\in\mathcal{P}:p\neq q} d_{\mathrm{TV}}(p,q) > \epsilon$. Then*

$$n^*_{\texttt{identical}}(\mathcal{P}, \mathcal{T}_D) \lesssim \frac{D\log M}{\epsilon^2} \min\left\{\log(DM^2)M^{6+\frac{4}{D-1}}, kM^{\frac{4}{D-1}}\log(Dk)\right\}. \tag{10.17}$$

In particular, for $D = \Omega(\log(M))$, we have

$$n^*_{\texttt{identical}}(\mathcal{P}, \mathcal{T}_D) \lesssim \frac{\log^2 M}{\epsilon^2} \min\left\{M^6, k\log k\right\} \tag{10.18}$$

(since $n^*_{\texttt{identical}}(\mathcal{P}, \mathcal{T}_D)$ decreases in D). Furthermore, for any p and q, the channel achieving the rates in inequalities (10.17) and (10.18) can be found efficiently using a linear program of polynomial size.

Proof. Let $\mathcal{P} = \left\{p^{(1)}, \ldots, p^{(M)}\right\}$. We prove the bound (10.17) by reducing the problem to a (decentralized) testing problem between distributions $\mathcal{P}' = \left\{q^{(1)}, \ldots, q^{(M)}\right\}$, where $q^{(i)} \in \Delta_D$ and $q^{(i)} = \mathbf{T}p^{(i)}$ for some $\mathbf{T} \in \mathcal{T}_D$. Defining $\epsilon' = \min_{i\neq j} d_{\mathrm{TV}}\left(\mathbf{T}q^{(i)}, \mathbf{T}q^{(j)}\right)$, Fact 10.2.4 shows the existence of an algorithm with sample complexity $O\left(\frac{\log M}{\epsilon'^2}\right)$. Thus, the goal is to find a channel $\mathbf{T}$ that maximizes $\min_{i\neq j} d_{\mathrm{TV}}\left(\mathbf{T}p^{(i)}, \mathbf{T}p^{(j)}\right)$, leading to the linear program

$$\max_{\mathbf{T}\in\mathcal{T}_D} \min_{i\neq j} d_{\mathrm{TV}}(\mathbf{T}p^{(i)}, \mathbf{T}p^{(j)}). \tag{10.19}$$

Let OPT be the value of the maximum in expression (10.19). The overall sample complexity of the algorithm is then $O\left(\frac{\log M}{\mathrm{OPT}^2}\right)$. We now prove each of the two bounds in inequality (10.17) by lower-bounding OPT in two different ways.

Bound II. We first prove the second bound in inequality (10.17) of Theorem 10.5.5. The following result, proved in Appendix G.4.2, provides a lower bound on the quantity (10.19) by using a Johnson-Lindenstrauss (JL) type of sketch:

Lemma 10.5.6 (JL-sketch)**.** *There exists a constant $c > 0$ such that the following holds: Let $\left\{p^{(1)}, \ldots, p^{(M)}\right\} \subseteq \Delta_k$ be M distributions such that $\min_{i \neq j} d_{\mathrm{TV}}(p^{(i)}, p^{(j)}) > \epsilon$. Then*

$$\max_{\mathbf{T} \in \mathcal{T}_D} \min_{i \neq j} d_{\mathrm{TV}}(\mathbf{T}p^{(i)}, \mathbf{T}p^{(j)}) \geq c \cdot \frac{\epsilon}{\sqrt{k} M^{\frac{2}{D-1}} \sqrt{D \log(Dk)}}.$$

Bound I. We note that the first bound in inequality (10.17) is better than the second bound when $k \gg M$. Our strategy will be to reduce the problem from a domain with k elements to a domain of (potentially) smaller size.

Claim 10.5.7 (Reduction to a domain of size M^2)**.** *Let $\mathcal{P} = \left\{p^{(1)}, \ldots, p^{(M)}\right\}$ and consider the setting of Theorem 10.5.5. There exists a channel $\mathbf{T} : [k] \to [M^2]$ such that for all $1 \leq i < j \leq M$, we have $d_{\mathrm{TV}}\left(\mathbf{T}p^{(i)}, \mathbf{T}p^{(j)}\right) \geq \frac{1}{M^2} \cdot d_{\mathrm{TV}}\left(p^{(i)}, p^{(j)}\right)$.*

Proof. Note that the result holds trivially for $k \leq M^2$, so assume that $k > M^2$. Let $d = \binom{M}{2}$. For two distributions p and q, we have $d_{\mathrm{TV}}(p, q) = \frac{1}{2}\|p - q\|_1$. Thus, we will show the existence of a column-stochastic matrix $\mathbf{T} \in \mathbb{R}^{d \times k}$, i.e., each entry of $\mathbf{T}$ is non-negative and the sum of each column is 1, satisfying the following conclusion when interpreted as an inequality concerning matrices and vectors: $\left\|\mathbf{T}(p^{(i)} - p^{(j)})\right\|_1 \geq \frac{1}{M^2} \cdot \left\|p^{(i)} - p^{(j)}\right\|_1$.

We will index rows by (i, j), for $1 \leq i < j \leq M$. We first define a matrix $\mathbf{T}' \in \mathbb{R}^{d \times k}$, such that the $(i, j)^{\text{th}}$ row is the vector $z'_{(i,j)} \in \mathbb{R}^k$ with ℓ^{th} entry equal to $\mathbb{I}_{p^{(i)}(\ell) > p^{(j)}(\ell)}$. It is easy to see that

$$\left\|\mathbf{T}'p^{(i)} - \mathbf{T}'p^{(j)}\right\|_1 = \left|\left\langle z'_{(i,j)}, p^{(i)} - p^{(j)} \right\rangle\right| = \frac{1}{2}\left\|p^{(i)} - p^{(j)}\right\|_1. \tag{10.20}$$

We now construct $\mathbf{T}$ by transforming $\mathbf{T}'$ into a column-stochastic matrix by dividing each column by the sum of its entries. Let $\left\{z_{(i,j)}\right\}$ denote the rows of $\mathbf{T}$. As each entry of $\mathbf{T}'$ is at most 1 and the number of rows is d, each entry of $\mathbf{T}$ is at least $\frac{1}{d}$ times the corresponding entry of $\mathbf{T}'$, i.e., $z_{i,j} \geq \frac{z'_{i,j}}{d}$, interpreted as an entrywise inequality.

Thus, for any $1 \leq i < j \leq M$, we have

$$\left\|\mathbf{T}p^{(i)} - \mathbf{T}p^{(j)}\right\|_1 \geq \left|\left\langle z_{i,j}, p^{(i)} - p^{(j)}\right\rangle\right| \geq \frac{1}{d}\left|\left\langle z'_{(i,j)}, p^{(i)} - p^{(j)}\right\rangle\right|,$$

noting that each entry in the sum $\langle z_{i,j}, p^{(i)} - p^{(j)}\rangle$ is nonnegative by construction. Combining with inequality (10.20), we obtain

$$\left\|\mathbf{T}p^{(i)} - \mathbf{T}p^{(j)}\right\|_1 \geq \frac{1}{2d}\left\|p^{(i)} - p^{(j)}\right\|_1 \geq \frac{1}{M^2}\left\|p^{(i)} - p^{(j)}\right\|_1.$$

This completes the proof. □

Returning to the original problem setting, let $\mathbf{T}_1$ be a channel from Claim 10.5.7 that transforms $p^{(i)} \in \Delta_k$ to $q^{(i)} \in \Delta_{M^2}$, such that $\min_{i\neq j} d_{\mathrm{TV}}(q^{(i)}, q^{(j)}) \geq \frac{\epsilon}{M^2}$. Define $\epsilon' = \frac{\epsilon}{M^2}$ and $k' = M^2$. Applying Lemma 10.5.6, there exists a channel $\mathbf{T}_2 : [k'] \to [D]$ such that for all $i \neq j$, we have

$$d_{\mathrm{TV}}\left(\mathbf{T}_2 q^{(i)}, \mathbf{T}_2 q^{(j)}\right) \gtrsim \frac{\epsilon'}{\sqrt{k'}} \frac{1}{M^{\frac{2}{D-1}}\sqrt{D\log(Dk')}} = \frac{\epsilon}{M^{3+\frac{2}{D-1}}\sqrt{D\log(DM^2)}}.$$

We define the final channel $\mathbf{T} : [k] \to [D]$ to be the concatenation of the two channels $\mathbf{T}_1$ and $\mathbf{T}_2$. In matrix notation, this corresponds to $\mathbf{T} := \mathbf{T}_2 \times \mathbf{T}_1$. Then $\mathrm{OPT} \gtrsim \frac{\epsilon}{M^{3+\frac{2}{D-1}}\sqrt{D\log(DM^2)}}$. □

10.5.2 Lower Bounds

In this section, we present lower bounds for the M-ary hypothesis testing problem. We begin by proving a lower bound of $\Omega(M)$ for adaptive algorithms, thus also for non-adaptive algorithms. Although we state our results for sequentially adaptive algorithms, these lower bounds also hold for the more general blackboard protocol; see the papers [BGMNW16; SVW16] for more details.

Theorem 10.5.8 (Adaptive lower bounds)**.** *For every $M \geq 2$ and $\epsilon < 0.1$, there exist $k \in \mathbb{N}$ and a set of M distributions $\mathcal{P}_M \subseteq \Delta_k$ such that the following hold:*

1. *(Lower bound from strong distributed data processing and direct-sum reduction [BGMNW16].) $k = O\left(2^M\right)$, $n^*(\mathcal{P}_M) \lesssim \frac{\log M}{\epsilon^2}$, and $n^*_{\texttt{adaptive}}(\mathcal{P}_M, \mathcal{T}_D) \gtrsim \frac{M}{\epsilon^2 \log D}$.*
2. *(Lower bound from SQ dimension [SVW16; Fel17].) $k = O(M)$, $n^*(\mathcal{P}_M) \lesssim \frac{\log M}{\epsilon^2}$, and $n^*_{\texttt{adaptive}}(\mathcal{P}_M, \mathcal{T}_D) \gtrsim \Omega\left(\frac{M^{1/3}}{\epsilon^{2/3} D^{2/3} (\log D)^{1/3}}\right)$ as long as $M \gtrsim \frac{\log D}{\epsilon D}$.*

The proof of Theorem 10.5.8 is given in Appendix G.5.

Remark 10.5.9 (An elementary proof of $\Omega(\sqrt{M})$)**.** *We also provide an elementary proof of an $\Omega(\sqrt{M})$ lower bound for non-adaptive channels that relies on the impossibility of ℓ_1-embedding using linear transforms [LMN05; CS02]. See Appendix G.5.3 for more details.*

Theorem 10.5.10 (Lower bounds for identical channels and $D = 2$)**.** *There exist constants $c_1, c_2 > 0$ such that the following holds for every $M \geq 2$ and $\mathcal{P}_M := \left\{p^{(1)}, \ldots, p^{(M)}\right\} \subseteq \Delta_k$: Let $\epsilon_1 := \min_{i \neq j} d_{\mathrm{h}}\left(p^{(i)}, p^{(j)}\right)$ and $\epsilon_2 := \max_{i \neq j} d_{\mathrm{h}}\left(p^{(i)}, p^{(j)}\right)$. Then $n^*(\mathcal{P}_M) \leq \frac{c_1 \log M}{\epsilon_1^2}$ and $n^*_{\texttt{identical}}(\mathcal{P}_M, \mathcal{T}_2) \geq \frac{c_2 M^2}{\epsilon_2^2}$.*

Remark 10.5.11. *Theorem 10.5.10 is a strong lower bound in the sense that it holds for every set of distributions. By a standard volumetric argument, it is possible to construct a set of M distributions $\mathcal{P}_M \subseteq \Delta_k$ such that (i) $\epsilon_2 \lesssim \epsilon_1 = c$ for a constant c and (ii) $M = 2^{\Omega(k)}$. As an algorithm for distribution estimation in d_{TV} implies a testing algorithm (see, e.g., [Tsy09]),*

a standard argument implies that any algorithm that uses an identical channel $\mathbf{T} \in \mathcal{T}_2$ and learns the underlying distribution p in d_{TV} up to a small constant requires at least $2^{\Omega(k)}$ samples. This is in stark contrast to the setting of non-identical channels, where Acharya, Canonne, and Tyagi [ACT20b] provide an algorithm with sample complexity $O(k^2)$.

Proof of Theorem 10.5.10. Observe that the upper bound on $n^*(\mathcal{P}_M)$ follows directly from Fact 10.2.4. We now focus on the lower bound. Our goal is to show that there exists a constant c_2 such that

$$\sup_{\mathbf{T}\in\mathcal{T}_2} \min_{i\neq j} d_{\mathrm{h}}(\mathbf{T}p^{(i)}, \mathbf{T}p^{(j)}) \leq \frac{c_2\epsilon_2}{M}, \tag{10.21}$$

since Fact 10.2.4 then implies a lower bound of $n^*(\mathcal{P}_M, \{\mathbf{T}\}) \gtrsim \frac{M^2}{\epsilon_2^2}$.

Fix a channel $\mathbf{T} \in \mathcal{T}_2$. Let $\mathcal{Q}_M = \left\{q^{(1)}, \ldots, q^{(M)}\right\}$ be the set of binary distributions obtained after transforming $\mathcal{P}_M$ via the channel, where $q^{(i)} = \mathbf{T}p_i$. Since $\mathcal{Q}_M$ is a set of binary distributions, each distribution $q^{(i)}$ can be represented by a single scalar parameter, which we denote by q_i. Without loss of generality, let $0 \leq q_1 < q_2 < \cdots < q_k \leq 1$. By the data processing inequality, we have $\max_{i\neq j} d_{\mathrm{h}}\left(q^{(i)}, q^{(j)}\right) \leq d_{\mathrm{h}}\left(p^{(i)}, p^{(j)}\right) \leq \epsilon_2$. We will show that in fact, there exists an index i^* such that $d_{\mathrm{h}}\left(q^{(i^*)}, q^{(i^*+1)}\right) \lesssim \frac{\epsilon_2}{M}$. Taking a supremum over all $\mathbf{T} \in \mathcal{T}_2$ would then establish the result in inequality (10.21).

For a $q \in [0, 1]$, let $\mathrm{Ber}(q)$ denote the Bernoulli distribution with parameter q. For $0 \leq q \leq q' \leq 1/2$, we have the following (cf. Claim G.6.2):

$$\sqrt{q'} - \sqrt{q} \leq d_{\mathrm{h}}\left(\mathrm{Ber}(q), \mathrm{Ber}(q')\right) \leq \sqrt{2}\left(\sqrt{q'} - \sqrt{q}\right). \tag{10.22}$$

Let r be the largest integer such that $q_r \leq 1/2$. We will assume that $r \geq \frac{M}{2}$ (otherwise, apply the following argument to $1 - q_i$). Using inequality (10.22) twice, we obtain

$$\sum_{i=1}^{r-1} d_{\mathrm{h}}\left(q^{(i)}, q^{(i+1)}\right) \leq \sqrt{2}\sum_{i=1}^{r-1}\left(\sqrt{q_{i+1}} - \sqrt{q_i}\right) = \sqrt{2}\left(\sqrt{q_r} - \sqrt{q_1}\right) \leq \sqrt{2} d_{\mathrm{h}}\left(q^{(1)}, q^{(r)}\right) \leq \sqrt{2}\epsilon_2.$$

As $r \geq \frac{M}{2}$, the preceding inequality implies that there exists some i^* such that

$$d_{\mathrm{h}}\left(q^{(i^*)}, q^{(i^*+1)}\right) \lesssim \frac{\epsilon_2}{r} \lesssim \frac{\epsilon_2}{M}.$$

□

10.6 Discussion

We have studied the problem of simple hypothesis testing under communication constraints. Taking a cue from past work on decentralized detection, we have focused on threshold channels and analyzed the sample complexity for a (near-optimal) threshold channel. For simple binary hypothesis testing, we showed that this choice leads to an at most logarithmic increase in the sample complexity of the test. We extended this result to the robust setting, where distributions may be contaminated in total variation. Importantly, our algorithms for hypothesis testing in the simple and robust settings were shown to be computationally efficient. Finally, we studied M-ary hypothesis testing by considering settings where the channels are identical, non-identical, or adaptive. We showed that communication constraints may lead to an exponential increase in sample complexity even for adaptive channels. For identical channels, we developed an efficient algorithm and analyzed its sample complexity. At a technical level, our results rely on a reverse data processing inequality for communication-constrained channels, a reverse Markov inequality, and a sketching algorithm akin to the Johnson-Lindenstrauss theorem.

There are several research directions that are worth exploring: Is adaptivity of channels useful in simple binary hypothesis testing? Can one tighten the dependence of sample complexity on the minimum Hellinger distance between the distributions for M-ary hypothesis testing? (Our results use total variation as a proxy for Hellinger and

are likely to be loose.) It would also be interesting to study simple hypothesis testing under other constraints such as local differential privacy and memory. The technical tools developed in this paper, particularly the reverse data processing inequality, may also have applications in quantization via "single-shot compression" in information theory.

11 HYPOTHESIS TESTING UNDER LOCAL DIFFERENTIAL PRIVACY AND COMMUNICATION CONSTRAINTS

We study simple binary hypothesis testing under both local differential privacy (LDP) and communication constraints. We qualify our results as either minimax optimal or instance optimal: the former hold for the set of distribution pairs with prescribed Hellinger divergence and total variation distance, whereas the latter hold for specific distribution pairs. For the sample complexity of simple hypothesis testing under pure LDP constraints, we establish instance-optimal bounds for distributions with binary support; minimax-optimal bounds for general distributions; and (approximately) instance-optimal, computationally efficient algorithms for general distributions. When both privacy and communication constraints are present, we develop instance-optimal, computationally efficient algorithms that achieve the minimum possible sample complexity (up to universal constants). Our results on instance-optimal algorithms hinge on identifying the extreme points of the joint range set $\mathcal{A}$ of two distributions p and q, defined as $\mathcal{A} := \{(\mathbf{T}p, \mathbf{T}q) | \mathbf{T} \in \mathcal{C}\}$, where $\mathcal{C}$ is the set of channels characterizing the constraints.

11.1 Introduction

Statistical inference on distributed data is becoming increasingly common, due to the proliferation of massive datasets which cannot be stored on a single server, and greater awareness of the security and privacy risks of centralized data. An institution (or statistician) that wishes to infer an aggregate statistic of such distributed data needs to solicit information, such as the raw data or some relevant statistic, from data owners (individuals). Individuals may be wary of sharing their data due to its sensitive nature or their lack of trust in the institution. The local differential privacy (LDP) paradigm suggests a solution by requiring that individuals' responses divulge only a limited amount of in-

formation about their data to the institution. Privacy is typically ensured by deliberately randomizing individuals' responses, e.g., by adding noise. See Definition 11.1.1 below for a formal definition; we refer the reader to Dwork and Roth [DR13] for more details on differential privacy.

In this paper, we study distributed estimation under LDP constraints, focusing on simple binary hypothesis testing, a fundamental problem in statistical estimation. We will also consider LDP constraints in tandem with communication constraints. This is a more realistic setting, since bandwidth considerations often impose constraints on the size of individuals' communications. The case when only communication constraints are present was addressed previously by Pensia, Jog, and Loh [PJL22].

Recall that simple binary hypothesis testing is defined as follows: Let p and q be two distributions over a finite domain $\mathcal{X}$, and let $X_1, \dots, X_n \in \mathcal{X}^n$ be n i.i.d. samples drawn from either p or q. The goal of the statistician is to identify (with high probability) whether the samples were drawn from p or q. This problem has been extensively studied in both asymptotic and nonasymptotic settings [NP33; Wal45; Cam86]. For example, it is known that the optimal test for this problem is the likelihood ratio test, and its performance can be characterized in terms of divergences between p and q, such as the total variation distance, Hellinger divergence, or Kullback–Leibler divergence. In particular, the sample complexity of hypothesis testing, defined as the smallest sample size needed to achieve an error probability smaller than a small constant, say, 0.01, is $\Theta\left(\frac{1}{d_{\mathrm{h}}^2(p,q)}\right)$, where $d_{\mathrm{h}}^2(p,q)$ is the Hellinger divergence between p and q.

In the context of local differential privacy, the statistician no longer has access to the original samples $X_1, \dots, X_n$, but only their privatized counterparts: $Y_1, \dots, Y_n \in \mathcal{Y}^n$, for some set $\mathcal{Y}$.[25] Each X_i is transformed to Y_i via a private channel $\mathbf{T}_i$, which is simply a probability kernel specifying $\mathbf{T}_i(y, x) = \mathbb{P}(Y_i = y | X_i = x)$. With a slight abuse

[25] As shown in Kairouz, Oh, and Viswanath [KOV16], for simple binary hypothesis testing, we can take $\mathcal{Y}$ to be $\mathcal{X}$, with the same sample complexity (up to constant factors); see Fact 11.2.7.

of notation, we shall use $\mathbf{T}_i$ to denote the transition kernel in $\mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$, as well as the stochastic map $Y_i = \mathbf{T}_i(X_i)$. A formal definition of privacy is given below:

Definition 11.1.1 (ϵ-LDP)**.** *Let $\epsilon \in \mathbb{R}_+$, and let $\mathcal{X}$ and $\mathcal{Y}$ be two domains. A channel $\mathbf{T} : \mathcal{X} \to \mathcal{Y}$* satisfies ϵ-LDP *if*

$$\sup_{x,x'\in\mathcal{X}} \sup_{A\subseteq\mathcal{Y}} \mathbb{P}[\mathbf{T}(x) \in A] - e^{\epsilon} \cdot \mathbb{P}[\mathbf{T}(x') \in A] \leq 0,$$

where we interpret $\mathbf{T}$ as a stochastic map on $\mathcal{X}$. Equivalently, if $\mathcal{X}$ and $\mathcal{Y}$ are countable domains (as will be the case for us), a channel $\mathbf{T}$ is ϵ-LDP if $\sup_{x,x'\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \frac{\mathbf{T}(y,x)}{\mathbf{T}(y,x')} \leq e^{\epsilon}$, where we interpret $\mathbf{T}$ as the transition kernel.

When $\epsilon = \infty$, we may set Y_i equal to X_i with probability 1, and we recover the vanilla version of the problem with no privacy constraints.

Existing results on simple binary hypothesis testing under LDP constraints have focused on the high-privacy regime of $\epsilon \in (0, c)$, for a constant $c > 0$, and have shown that the sample complexity is $\Theta\left(\frac{1}{\epsilon^2 d_{\mathrm{TV}}^2(p,q)}\right)$, where $d_{\mathrm{TV}}(p, q)$ is the total variation distance between p and q (cf. Fact 11.2.7). Thus, when ϵ is a constant, the sample complexity is $\Theta\left(\frac{1}{d_{\mathrm{TV}}^2(p,q)}\right)$, and when $\epsilon = \infty$ (no privacy), the sample complexity is $\Theta\left(\frac{1}{d_{\mathrm{h}}^2(p,q)}\right)$. Although these two divergences satisfy $d_{\mathrm{TV}}^2(p, q) \lesssim d_{\mathrm{h}}^2(p, q) \lesssim d_{\mathrm{TV}}(p, q)$, the bounds are tight; i.e., the two sample complexities can be quadratically far apart. Existing results therefore do not inform sample complexity when $1 \ll \epsilon < \infty$. This is not an artifact of analysis: the optimal tests in the low and high privacy regimes are fundamentally different.

The large-ϵ regime has been increasingly used in practice, due to privacy amplification provided by shuffling [CSUZZ19; BEMMRLRKTS17; FMT21]. Our paper makes progress on the computational and statistical fronts in the large-ϵ regime, as will be highlighted in Section 11.1.3 below.

11.1.1 Problem Setup

For a natural number k, we use $[k]$ to denote the set $\{1, 2, \ldots, k\}$. In our paper, we focus on the private-coin, non-interactive protocol.[26] As we will be working with both privacy and communication constraints in this paper, we first define the generic protocol for distributed inference under an arbitrary set of channels $\mathcal{C}$ below:

Definition 11.1.2 (Simple binary hypothesis testing under channel constraints)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two countable sets. Let $\mathcal{C}$ be a set of channels from $\mathcal{X}$ to $\mathcal{Y}$, and let p and q be two distributions on $\mathcal{X}$. Let $\{U_i\}_{i=1}^n$ denote a set of n users who choose channels $\{\mathbf{T}_i\}_{i=1}^n \in \mathcal{C}^n$ according to a deterministic rule[27] $\mathcal{R} : [n] \to \mathcal{C}$. Each user U_i then observes X_i and generates $Y_i = \mathbf{T}_i(X_i)$ independently, where $X_1, \ldots, X_n$ is a sequence of i.i.d. random variables drawn from an (unknown) $r \in \{p, q\}$. The central server U_0 observes $(Y_1, \ldots, Y_n)$ and constructs an estimate $\widehat{r} = \phi(Y_1, \ldots, Y_n)$, for some test $\phi : \cup_{i=1}^{\infty} \mathcal{Y}^i \to \{p, q\}$. We refer to this problem as* simple binary hypothesis testing under channel constraints*.*

In the non-interactive setup, we can assume that all $\mathbf{T}_i$'s are identical equal to some $\mathbf{T}$, as it will increase the sample complexity by at most a constant factor [PJL22] (cf. Fact 11.2.7). We now specialize the setting of Definition 11.1.2 to the case of LDP constraints:

Definition 11.1.3 (Simple binary hypothesis testing under LDP constraints)**.** *Consider the problem in Definition 11.1.2 with $\mathcal{Y} = \mathbb{N}$, where $\mathcal{C}$ is the set of all ϵ-LDP channels from $\mathcal{X}$ to $\mathcal{Y}$. We denote this problem by $\mathcal{B}(p, q, \epsilon)$. For a given test-rule pair $(\phi, \mathcal{R})$ with $\phi : \cup_{j=1}^{\infty} \mathcal{Y}^j \to \{p, q\}$, we say that* $(\phi, \mathcal{R})$ solves $\mathcal{B}(p, q, \epsilon)$ with sample complexity n *if*

$$\mathop{\mathbb{P}}_{(X_1,\ldots,X_n)\sim p^{\otimes n}}(\phi(Y_1, \ldots, Y_n) \neq p) + \mathop{\mathbb{P}}_{(X_1,\ldots,X_n)\sim q^{\otimes n}}(\phi(Y_1, \ldots, Y_n) \neq q) \leq 0.1. \tag{11.1}$$

[26] We refer the reader to Acharya, Canonne, Liu, Sun, and Tyagi [ACLST22] for differences between various protocols.

[27] When $\mathcal{C}$ is a convex set of channels, as will be the case in this paper, the deterministic rules are equivalent to randomized rules (with independent randomness).

We use $n^*(p, q, \epsilon)$ to denote the sample complexity of this task, i.e., the smallest n so that there exists a $(\phi, \mathcal{R})$-pair that solves $\mathcal{B}(p, q, \epsilon)$. We use $\mathcal{B}(p, q)$ and $n^*(p, q)$ to refer to the setting of non-private testing, i.e., when $\epsilon = \infty$, which corresponds to the case when $\mathcal{C}$ is the set of all possible channels from $\mathcal{X}$ to $\mathcal{Y}$.

For any fixed rule $\mathcal{R}$, the optimal choice of ϕ corresponds to the likelihood ratio test on $\{Y_i\}_{i=1}^n$. Thus, in the rest of this paper, our focus will be optimizing the rule $\mathcal{R}$, with the choice of ϕ made implicitly. In fact, we can take $\mathcal{Y}$ to be $\mathcal{X}$, at the cost of a constant-factor increase in the sample complexity [KOV16] (cf. Fact 11.2.7).

We now define the threshold for free privacy, in terms of a large enough universal constant $C_{\text{thresh}} > 0$ which can be explicitly deduced from our proofs:

Definition 11.1.4 (Threshold for free privacy)**.** *We define $\epsilon^*(p, q)$ (also denoted by ϵ^* when the context is clear) to be the smallest ϵ such that $n^*(p, q, \epsilon) \leq C_{\text{thresh}} \cdot n^*(p, q)$; i.e., for all $\epsilon \geq \epsilon^*(p, q)$, we can obtain ϵ-LDP without any substantial increase in sample complexity compared to the non-private setting.*

Next, we study the problem of simple hypothesis testing under both privacy and communication constraints. By communication constraints, we mean that the channel $\mathbf{T}$ maps from $\mathcal{X}$ to $[\ell]$ for some $\ell \in \mathbb{N}$, which is potentially much smaller than $|\mathcal{X}|$.

Definition 11.1.5 (Simple binary hypothesis testing under LDP and communication constraints)**.** *Consider the problem in Definition 11.1.2 and Definition 11.1.3, with $\mathcal{C}$ equal to the set of all channels that satisfy ϵ-LDP and $\mathcal{Y} = [\ell]$. We denote this problem by $\mathcal{B}(p, q, \epsilon, \ell)$, and use $n^*(p, q, \epsilon, \ell)$ to denote its sample complexity.*

Communication constraints are worth studying not only for their practical relevance in distributed inference, but also for their potential to simplify algorithms without significantly impacting performance. Indeed, the sample complexities of simple hypothesis testing with and without communication constraints are almost identical [BNOP21;

PJL22] (cf. Fact 11.2.8), even for a single-bit ($\ell = 2$) communication constraint. As we explain later, a similar statement can be made for privacy constraints, as well.

11.1.2 Existing Results

As noted earlier, the problem of simple hypothesis testing with just communication constraints was addressed in Pensia, Jog, and Loh [PJL22]. Since communication and privacy constraints are the most popular information constraints studied in the literature, the LDP-only and LDP-with-communication-constraints settings considered in this paper are natural next steps. Many of our results, particularly those on minimax-optimal sample complexity bounds, are in a similar vein as those in Pensia, Jog, and Loh [PJL22]. Before describing our results, let us briefly mention the most relevant prior work. We discuss further related work in Section 11.1.4.

Existing results on sample complexity. Existing results (cf. Duchi, Jordan, and Wainwright [DJW18, Theorem 1] and Asoodeh and Zhang [AZ22, Theorem 2]) imply that

$$n^*(p, q, \epsilon) \gtrsim \begin{cases} \frac{1}{\epsilon^2 \cdot d_{\mathrm{TV}}^2(p,q)}, & \text{if } \epsilon \in (0, 1], \\ \frac{1}{e^\epsilon \cdot d_{\mathrm{TV}}^2(p,q)}, & \text{if } e^\epsilon \in \left(e, \frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{TV}}^2(p,q)}\right], \\ \frac{1}{d_{\mathrm{h}}^2(p,q)}, & \text{if } e^\epsilon > \frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{TV}}^2(p,q)}. \end{cases} \tag{11.2}$$

An upper bound on the sample complexity can be obtained by choosing a specific private channel $\mathbf{T}$ and analyzing the resulting test. A folklore result (see, for example, Joseph, Mao, Neel, and Roth [JMNR19, Theorem 5.1]) shows that setting $\mathbf{T} = \mathbf{T}_{\mathrm{RR}} \times \mathbf{T}_{\mathrm{Scheffe}}$, where $\mathbf{T}_{\mathrm{Scheffe}}$ maps $\mathcal{X}$ to $\{0, 1\}$ using a threshold rule based on $\frac{p(x)}{q(x)}$, and $\mathbf{T}_{\mathrm{RR}}$ is the binary-input binary-output randomized response channel, gives $n^*(p, q, \epsilon) \lesssim \frac{1}{\min(1,\epsilon^2) \cdot d_{\mathrm{TV}}^2(p,q)}$. This shows that when $\epsilon \in (0, 1]$ (or $(0, c]$, for some constant c), the lower bound is tight up to constants. Observe that for any $\ell \geq 2$, the sample complexity with privacy and

communication constraints $n^*(p, q, \epsilon, \ell)$ also satisfies the same lower and upper bounds, since the channel $\mathbf{T}$ has only two outputs.

However, the following questions remain unanswered:

What is the optimal sample complexity for $\epsilon \gg 1$? In particular, are the existing lower bounds Equation (11.2) tight? What is the threshold for free privacy?

In Section 11.1.3.1, we establish minimax-optimal bounds on the sample complexity for all values of ϵ, over sets of distribution pairs with fixed total variation distance and Hellinger divergence. In particular, we show that the lower bounds Equation (11.2) are tight for binary distributions, but may be arbitrarily loose for general distributions.

Existing results on computationally efficient algorithms. Recall that each user needs to select a channel $\mathbf{T}$ to optimize the sample complexity. Once $\mathbf{T}$ is chosen, the optimal test is simply a likelihood ratio test between $\mathbf{T}p$ and $\mathbf{T}q$. Thus, the computational complexity lies in determining $\mathbf{T}$. As noted earlier, for $\epsilon \leq 1$, the optimal channel is $\mathbf{T} = \mathbf{T}_{\text{RR}} \times \mathbf{T}_{\text{Scheffe}}$, and this can be computed efficiently. However, this channel $\mathbf{T}$ may no longer be optimal in the regime of $\epsilon \gg 1$.

As with statistical rates, prior literature on finding optimal channels for $\epsilon \gg 1$ is scarce. Existing algorithms either take time exponential in the domain size [KOV16], or their sample complexity is suboptimal by polynomial factors (depending on $\frac{1}{d^2_{\text{TV}}(p,q)}$, as opposed to $\frac{1}{d^2_{\text{h}}(p,q)}$). This raises the following natural question:

Is there a polynomial-time algorithm that finds a channel $\mathbf{T}$ whose sample complexity is (nearly) optimal?

We answer this question in the affirmative in Section 11.1.3.3.

11.1.3 Our Results

We are now ready to describe our results in this paper, which we outline in the next three subsections. In particular, Section 11.1.3.1 focuses on the sample complexity of simple hypothesis testing under local privacy, Section 11.1.3.2 focuses on structural properties of the extreme points of the joint range under channel constraints, and Section 11.1.3.3 states our algorithmic guarantees.

11.1.3.1 Statistical Rates

We begin by analyzing the sample complexity when both p and q are binary distributions. We prove the following result in Section 11.3.1, showing that the existing lower bounds Equation (11.2) are tight for binary distributions:

Theorem 11.1.6 (Sample complexity of binary distributions)**.** *Let p and q be two binary distributions. Then*

$$n^*(p, q, \epsilon) \asymp \begin{cases} \frac{1}{\epsilon^2 \cdot d_{\mathrm{TV}}^2(p,q)}, & \text{if } \epsilon \le 1, \\ \frac{1}{e^\epsilon \cdot d_{\mathrm{TV}}^2(p,q)}, & \text{if } e^\epsilon \in \left[e, \frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{TV}}^2(p,q)}\right], \\ \frac{1}{d_{\mathrm{h}}^2(p,q)}, & \text{if } e^\epsilon > \frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{TV}}^2(p,q)}. \end{cases} \tag{11.3}$$

In particular, the threshold ϵ^* for free privacy (Definition 11.1.4) satisfies $e^{\epsilon^*} \asymp \frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{TV}}^2(p,q)}$. Note that the sample complexity $n^*(p, q, \epsilon)$ for all ranges of ϵ is completely characterized by the total variation distance and Hellinger divergence between p and q. A natural set to consider is *all distribution pairs* (not just those with binary support) with a prescribed total variation distance and Hellinger divergence; we investigate minimax-optimal sample complexity over this set. Our next result shows that removing the binary support condition radically changes the sample complexity, even if the total variation distance and Hellinger divergence are the same. Specifically, we show that there are

ternary distribution pairs whose sample complexity (as a function of the total variation distance and Hellinger divergence) is significantly larger than the corresponding sample complexity for binary distributions.

Theorem 11.1.7 (Sample complexity lower bound for general distributions)**.** *For any $\rho \in (0, 0.5)$ and $\nu \in (0, 0.5)$ such that $2\nu^2 \le \rho \le \nu$, there exist ternary distributions p and q such that $d_{\mathrm{h}}^2(p, q) = \rho$, $d_{\mathrm{TV}}(p, q) = \nu$, and the sample complexity behaves as*

$$n^*(p, q, \epsilon) \asymp \begin{cases} \frac{1}{\epsilon^2 \cdot d_{\mathrm{TV}}^2(p,q)}, & \text{if } \epsilon \le 1, \\ \min\left(\frac{1}{d_{\mathrm{TV}}^2(p,q)}, \frac{1}{e^\epsilon \cdot d_{\mathrm{h}}^4(p,q)}\right), & \text{if } e^\epsilon \in \left[e, \frac{1}{d_{\mathrm{h}}^2(p,q)}\right], \\ \frac{1}{d_{\mathrm{h}}^2(p,q)}, & \text{if } e^\epsilon > \frac{1}{d_{\mathrm{h}}^2(p,q)}. \end{cases} \tag{11.4}$$

We prove this result in Section 11.3.2.

Remark 11.1.8. *We highlight the differences between the sample complexity in the binary setting (cf. equation Equation (11.3)) and the worst-case general distributions (cf. equation Equation (11.4)) below (also see Figure 11.1):*

1. *(Relaxing privacy may not lead to significant improvements in accuracy.) In equation Equation (11.4), there is an arbitrarily large range of ϵ where the sample complexity remains roughly constant. In particular, when $e \le e^\epsilon \lesssim \frac{d_{\mathrm{TV}}^2(p,q)}{d_{\mathrm{h}}^4(p,q)}$, the sample complexity of hypothesis testing remains roughly the same (up to constants). That is, we are sacrificing privacy without any significant gains in statistical efficiency. This is in stark contrast to the binary setting, where increasing e^ϵ by a large constant factor leads to a constant-factor improvement in sample complexity.*

2. *(The threshold for free privacy is larger.) Let $\epsilon^* := \epsilon(p, q)$ be the threshold for free privacy (cf. Definition 11.1.4). In the binary setting, one has $e^{\epsilon^*} \asymp \frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{TV}}^2(p,q)}$, whereas for general*

distributions, one may need $e^{\epsilon^*} \gtrsim \frac{1}{d_{\mathrm{h}}^2(p,q)}$. *The former* ϵ^* *can be arbitrarily smaller than the latter.*

To complement the result above, which provides a lower bound on the sample complexity for worst-case distributions, our next result provides an upper bound on the sample complexity that nearly matches the rates (up to logarithmic factors) for arbitrary distributions. Moreover, the proposed algorithm uses an ϵ-LDP channel with *binary* outputs. The following result is proved in Section 11.3.3:

Theorem 11.1.9 (Sample complexity upper bounds and an efficient algorithm for hypothesis testing for general distributions)**.** *Let* p *and* q *be two distributions on* $[k]$. *Let* $\epsilon > 0$. *Then the sample complexity behaves as*

$$n^*(p,q,\epsilon) \lesssim \begin{cases} \frac{1}{\epsilon^2 \cdot d_{\mathrm{TV}}^2(p,q)}, & \text{if } \epsilon \le 1, \\ \min\left(\frac{1}{d_{\mathrm{TV}}^2(p,q)}, \frac{\alpha^2}{e^\epsilon \cdot d_{\mathrm{h}}^4(p,q)}\right), & \text{if } e^\epsilon \in \left(e, \frac{\alpha}{d_{\mathrm{h}}^2(p,q)}\right], \\ \frac{\alpha}{d_{\mathrm{h}}^2(p,q)}, & \text{if } e^\epsilon > \frac{\alpha}{d_{\mathrm{h}}^2(p,q)}, \end{cases} \tag{11.5}$$

where $\alpha \lesssim \log(1/d_{\mathrm{h}}^2(p,q)) \asymp \log\left(n^*(p,q)\right)$.

Moreover, the rates above are achieved by an ϵ*-LDP channel* $\mathbf{T}$ *that maps* $[k]$ *to* $[2]$ *and can be found in time polynomial in* k, *for any choice of* p, q, *and* ϵ.

Theorems 11.1.7 and 11.1.9 imply that the above sample complexity is minimax optimal (up to logarithmic factors) over the class of distributions with total variation distance ν and Hellinger divergence ρ satisfying the conditions in Theorem 11.1.7. We summarize this in the following theorem:

Theorem 11.1.10 (Minimax-optimal bounds)**.** *Let* $\rho \in (0, 0.5)$ *and* $\nu \in (0, 0.5)$ *be such that* $2\nu^2 \le \rho \le \nu$. *Let* $S_{\rho,\nu}$ *be the set of all distribution pairs with discrete supports, with total*

variation distance and Hellinger divergence being ν and ρ, respectively:

$$S_{\rho,\nu} := \{(p,q) : k \in \mathbb{N}, p \in \Delta_k, q \in \Delta_k, d_{\mathrm{TV}}(p,q) = \nu, d^2_{\mathrm{h}}(p,q) = \rho\}.$$

Let $n^*(S_{\rho,\nu}, \epsilon)$ be the minimax-optimal sample complexity of hypothesis testing under ϵ-LDP constraints, defined as

$$n^*(S_{\rho,\nu}, \epsilon) = \min_{(\phi,\mathcal{R})} \max_{(p,q)\in S_{\rho,\nu}} n^*(p,q,\epsilon),$$

for test-rule pairs $(\phi, \mathcal{R})$, as defined in Definition 11.1.3. Then

$$n^*(S_{\rho,\nu}, \epsilon) = \begin{cases} \widetilde{\Theta}\left(\frac{1}{\epsilon^2 \cdot \nu^2}\right), & \text{if } \epsilon \le 1, \\ \widetilde{\Theta}\left(\min\left(\frac{1}{\nu^2}, \frac{1}{e^\epsilon \cdot \rho^2}\right)\right), & \text{if } e^\epsilon \in \left[e, \frac{1}{\rho}\right], \\ \widetilde{\Theta}\left(\frac{1}{\rho}\right), & \text{if } e^\epsilon > \frac{1}{\rho}. \end{cases} \tag{11.6}$$

Here, the $\widetilde{\Theta}$ notation hides poly-logarithmic factors in $1/\nu$ and $1/\rho$.

Remark 11.1.11. *A version of the above theorem may also be stated for privacy and communication constraints, by defining*

$$n^*(S_{\rho,\nu}, \epsilon, \ell) = \min_{(\phi,\mathcal{R})} \max_{(p,q)\in S_{\rho,\nu}} n^*(p,q,\epsilon,\ell).$$

In fact, it may seen that the same sample complexity bounds continue to hold for $n^*(S_{\rho,\nu}, \epsilon, \ell)$, with $\ell \ge 2$, since the lower bound in Theorem 11.1.7 continues to hold with communication constraints, as does the upper bound in Theorem 11.1.9, which uses a channel with only binary outputs.

Remark 11.1.12. *The above theorem mirrors a minimax optimality result for communication-constrained hypothesis testing from Pensia, Jog, and Loh [PJL22]. There, the set under con-*

sideration was S_ρ, where ρ is the Hellinger divergence between the distribution pair, and the minimax-optimal sample complexity was shown to be $\widetilde{\Theta}(1/\rho)$ even for a binary communication constraint.

Finally, we consider the threshold for free privacy ϵ^* for general distributions; see Definition 11.1.4. Observe that Theorem 11.1.9 does not provide any upper bounds on ϵ^*, since the sample complexity in Theorem 11.1.9 is bounded away from $n^*(p, q)$, due to the logarithmic multiplier α. Recall that Theorem 11.1.7 implies $e^{\epsilon^*} \gtrsim \frac{1}{d_{\mathrm{h}}^2(p,q)}$ in the worst case. Our next result, proved in Section 11.3.3, shows that this is roughly tight, and $e^{\epsilon^*} \lesssim \frac{1}{d_{\mathrm{h}}^2(p,q)} \cdot \log\left(\frac{1}{d_{\mathrm{h}}^2(p,q)}\right)$ for all distributions:

Theorem 11.1.13. *Let p and q be two distributions on $[k]$, and let $e^\epsilon \gtrsim \frac{1}{d_{\mathrm{h}}^2(p,q)} \log\left(\frac{1}{d_{\mathrm{h}}^2(p,q)}\right)$. Then $n^*(p, q, \epsilon) \asymp n^*(p, q)$. Moreover, there is a channel $\mathbf{T}$ achieving this sample complexity that maps $[k]$ to a domain of size $\lceil \log(n^*(p, q)) \rceil$, and which can be computed in $\mathrm{poly}(k, \log(\lceil n^*(p, q) \rceil))$ time.*

We thereby settle the question of minimax-optimal sample complexity (up to logarithmic factors) for simple binary hypothesis testing under LDP-only and LDP-with-communication constraints (over the class of distributions with a given total variation distance and Hellinger divergence). Moreover, the minimax-optimal upper bounds are achieved by computationally efficient, communication-efficient algorithms. However, there can be a wide gap between instance-optimal and minimax-optimal procedures; in the next two subsections, we present structural and computational results for instance-optimal algorithms.

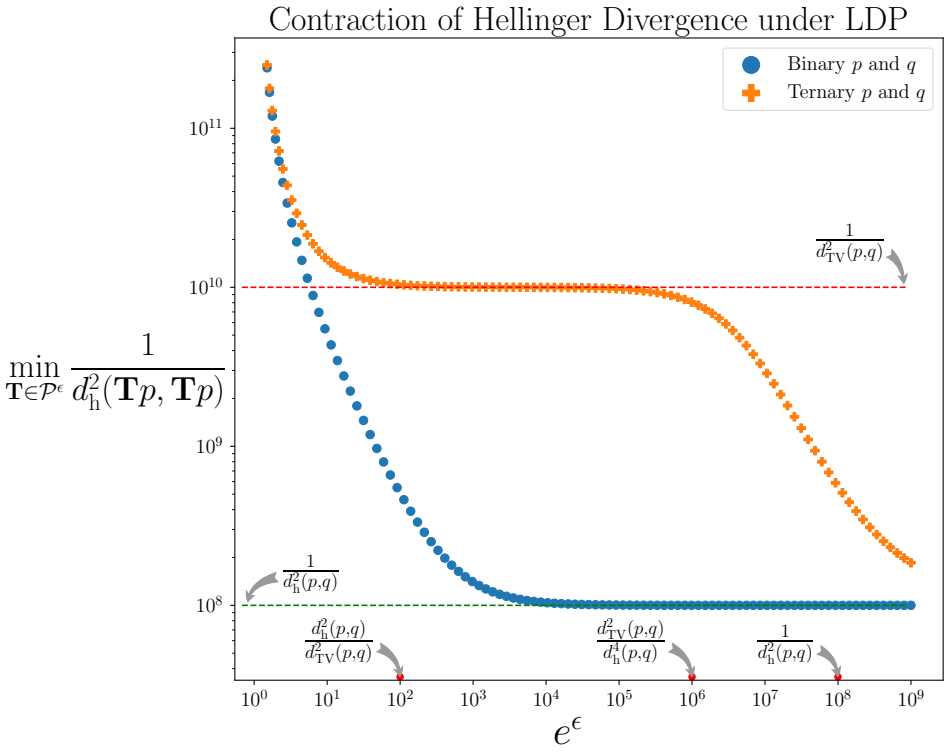

Figure 11.1: In this plot, we show the difference between the behavior of sample complexity under ϵ-LDP constraints for binary distributions and (worst-case) ternary distributions from Theorem 11.1.7. We take two pairs of distributions (p, q)—one pair of binary distributions (shown in blue, with marker $\circ$) and one pair of ternary distributions (shown in orange, with marker $+$)—such that the two pairs have Hellinger divergence $d_{\mathrm{h}}^2(p, q) = 10^{-8}$ and total variation distance $d_{\mathrm{TV}}(p, q) = 10^{-5}$. For each value of ϵ, shown on the horizontal axis after being mapped to e^ϵ, we compute $\min_{\mathbf{T}\in\mathcal{P}^\epsilon} 1/d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)$, where $\mathcal{P}^\epsilon$ is the set of all ϵ-LDP channels, and plot it on vertical axis. Thus, the vertical axis characterizes the sample complexity $n^*(p, q, \epsilon)$ of simple binary hypothesis testing between p and q with privacy constraints, up to constant factors (cf. Fact 11.2.7). Both axes are shown in log-scale here. Since the total variation distance between the two pairs is identical, we see that their curves overlap for small ϵ ($\epsilon \ll 1$, which is consistent with the fact that $n^*(p, q, \epsilon) \asymp \frac{1}{\epsilon^2 d_{\mathrm{TV}}^2(p,q)}$ for small ϵ). As predicted by Theorem 11.1.6, the curve for binary distributions decreases rapidly for $\epsilon \gg 1$ until it saturates at $1/d_{\mathrm{h}}^2(p, q)$. Moreover, for $e^\epsilon \asymp d_{\mathrm{h}}^2(p, q)/d_{\mathrm{TV}}^2(p, q)$, the predicted threshold for free privacy, the vertical axis is within constant factors of its asymptotic value, as predicted. On the other hand, the curve for ternary distributions seems to have three different phases, as predicted by Theorem 11.1.7: (i) for small ϵ, it behaves as $1/(\epsilon^2 d_{\mathrm{TV}}^2(p, q))$; (ii) for moderate values of ϵ, such that $e \ll e^\epsilon \ll \frac{d_{\mathrm{TV}}^2(p,q)}{d_{\mathrm{h}}^4(p,q)}$, it remains stagnant roughly at $\frac{1}{d_{\mathrm{TV}}^2(p,q)}$; and (iii) for $e^\epsilon \gg \frac{d_{\mathrm{TV}}^2(p,q)}{d_{\mathrm{h}}^4(p,q)}$, the curve decreases rapidly until it approaches $1/d_{\mathrm{h}}^2(p, q)$. The phase (ii) corresponds to the phenomenon that we are leaking privacy without any gains in statistical efficiency. Finally, e^ϵ needs to be as large as $1/d_{\mathrm{h}}^2(p, q)$ for the vertical axis to be within a factor of 10 of its asymptotic value. We refer the reader to Remark 11.1.8 for more details.

11.1.3.2 Structure of Extreme Points under the Joint Range

In this section, we present results for the extreme points of the joint range of an arbitrary pair of distributions when transformed by a set of channels. Formally, if $\mathcal{C}$ is a convex set of channels from $\mathcal{X}$ to $\mathcal{Y}$, and p and q are two distributions on $\mathcal{X}$, we are interested in the extreme points of the set $\mathcal{A} := \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{C}\}$, which is a convex subset of $\Delta_{|\mathcal{Y}|} \times \Delta_{|\mathcal{Y}|}$.[28] The extreme points of a convex set are naturally insightful for maximizing quasi-convex functions, and we will present the consequences of the results in this section in Section 11.1.3.3.

We consider two choices of $\mathcal{C}$: first, when $\mathcal{C}$ is the set of all channels from $\mathcal{X}$ to $\mathcal{Y} = [\ell]$, and second, when $\mathcal{C}$ is the set of all ϵ-LDP channels from $\mathcal{X}$ to $\mathcal{Y} = [\ell]$. We use $\mathcal{T}_{\ell,k}$ to denote the set of all channels that map from $[k]$ to $[\ell]$.

The following class of deterministic channels plays a critical role in our theory:

Definition 11.1.14 (Threshold channels)**.** *For some $k \in \mathbb{N}$, let p and q be two distributions on $[k]$. For any $\ell \in \mathbb{N}$, a deterministic channel $\mathbf{T} \in \mathcal{T}_{\ell,k}$ is a* threshold channel *if the following property holds for every $u, v \in [k]$: If $\frac{p(u)}{q(u)} < \frac{p(v)}{q(v)}$ and $\mathbf{T}(u) = \mathbf{T}(v)$, then any $w \in [k]$ such that $\frac{p(w)}{q(w)} \in \left(\frac{p(u)}{q(u)}, \frac{p(v)}{q(v)}\right)$ satisfies $\mathbf{T}(w) = \mathbf{T}(u)(= \mathbf{T}(v))$. (The likelihood ratios are assumed to take values on the extended real line; i.e., on $\mathbb{R} \cup \{-\infty, +\infty\}$.)*

Remark 11.1.15. *Threshold channels are intuitively easy to understand when all the likelihood ratios are distinct (this may be assumed without loss of generality in our paper, as explained later): Arrange the inputs in increasing order of their likelihood ratios and partition them into ℓ contiguous blocks. Thus, there are at most k^ℓ such threshold channels (up to reordering of output labels).*

Our first result proved in Section 11.4 is for the class of communication-constrained channels, and shows that all extreme points of the joint range are obtained using deter-

[28] For $k \in \mathbb{N}$, we use Δ_k to denote the probability simplex on a domain of alphabet size k.

ministic threshold channels:

Theorem 11.1.16 (Extreme points of the joint range under communication constraints)**.** *Let p and q be two distributions on $[k]$. Let $\mathcal{A}$ be the set of all pairs of distributions that are obtained by passing p and q through a channel of output size ℓ, i.e.,*

$$\mathcal{A} = \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{T}_{\ell,k}\}.$$

If $(\mathbf{T}p, \mathbf{T}q)$ is an extreme point of $\mathcal{A}$, then $\mathbf{T}$ is a threshold channel.

We note that the above result is quite surprising: $(\mathbf{T}p, \mathbf{T}q)$ is extreme point of $\mathcal{A}$ only if $\mathbf{T}$ is an extreme point of $\mathcal{T}_{\ell,k}$ (i.e., a deterministic channel), but Theorem 11.1.16 demands that $\mathbf{T}$ be a deterministic *threshold* channel, meaning it lies in a very small subset of deterministic channels. Indeed, even for $\ell = 2$, the number of deterministic channels from $[k]$ to $[2]$ is 2^k, whereas the number of threshold channels is just $2k$. We note that the result above is similar in spirit to Tsitsiklis [Tsi93, Proposition 2.4]. However, the focus there was on a particular objective, the probability of error in simple hypothesis testing, with non-identical channels for users. Our result is for identical channels and is generally applicable to quasi-convex objectives, as mentioned later.

We now consider the case where $\mathcal{C}$ is the set of ϵ-LDP channels from $[k]$ to $[\ell]$. Since $\mathcal{C}$ is a set of private channels, it does not contain any deterministic channels (thus, does not contain threshold channels). Somewhat surprisingly, we still show that the threshold channels play a fundamental role in the extreme points of the joint range under $\mathcal{C}$. The following result shows that any extreme point of the joint range $\mathcal{A}$ can be obtained by a threshold channel mapping into $[2\ell^2]$, followed by an ϵ-LDP channel from $[2\ell^2]$ to $[\ell]$:

Theorem 11.1.17 (Extreme points of the joint range under privacy and communication constraints)**.** *Let p and q be distributions on $[k]$. Let $\mathcal{C}$ be the set of ϵ-LDP channels from $[k]$ to $[\ell]$. Let $\mathcal{A}$ be the set of all pairs of distributions that are obtained by applying a channel from $\mathcal{C}$ to*

p and q, i.e.,

$$\mathcal{A} = \{(\mathbf{T}p, \mathbf{T}q) \mid \mathbf{T} \in \mathcal{C}\}. \tag{11.7}$$

If $(\mathbf{T}p, \mathbf{T}q)$ is an extreme point of $\mathcal{A}$ for $\mathbf{T} \in \mathcal{C}$, then $\mathbf{T}$ can be written as $\mathbf{T} = \mathbf{T}_2 \times \mathbf{T}_1$ for some threshold channel $\mathbf{T}_1 \in \mathcal{T}_{2\ell^2,k}$ and some $\mathbf{T}_2$ an extreme point of the set of ϵ-LDP channels from $[2\ell^2]$ to $[\ell]$.

We prove this structural result in Section 11.5, which leads to polynomial-time algorithms for constant ℓ for maximizing quasi-convex functions, as mentioned in Section 11.1.3.3.

11.1.3.3 Computationally Efficient Algorithms for Instance Optimality

The results from the previous sections characterized the minimax-optimal sample complexity, but did not address instance optimality. Instance-optimal performance may be substantially better than minimax-optimal performance, as seen by comparing the instance-optimal bounds for binary distributions to the minimax-optimal bounds for general distributions. In this section, we focus on identifying an instance-optimal channel $\mathbf{T}$ (satisfying the necessary constraints) for a given pair (p, q) of distributions.

Let p and q be fixed distributions over $[k]$. Let $\mathcal{P}^{\epsilon}_{\ell,k}$ be the set of all ϵ-LDP channels from $[k]$ to $[\ell]$, and let $\mathcal{T}_{\ell,k}$ be the set of all channels from $[k]$ to $[\ell]$. Let $\mathcal{C} \in \{\mathcal{P}^{\epsilon}_{\ell,k}, \mathcal{T}_{\ell,k}\}$. As before, define $\mathcal{A} = \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{C}\}$. Let $g : \mathcal{A} \to \mathbb{R}$ be a (jointly) quasi-convex function; i.e., for all $t \in \mathbb{R}$, the sublevel sets $\{(p', q') : g(p', q') \leq t\}$ are convex. In this paper, we are primarily interested in functions corresponding to divergences between the distribution pair. So, unless otherwise mentioned, we shall assume the quasi-convex functions g in this paper are permutation-invariant; i.e., $g(p', q') = g(\Pi p, \Pi q)$ for all permutation matrices Π. However, our algorithmic results will continue to hold even without this assumption, with an additional factor of $\ell!$ in the time complexity. We will

consider the problem of identifying $\mathbf{T}$ that solves

$$\max_{\mathbf{T} \in \mathcal{C}} g(\mathbf{T}p, \mathbf{T}q).$$

The quasi-convexity of g implies that the maximum is attained at some $\mathbf{T}$ such that $(\mathbf{T}p, \mathbf{T}q)$ is an extreme point of $\mathcal{A}$. We can thus leverage the results from Section 11.1.3.2 to search over the subset of channels satisfying certain structural properties.

Identifying $\mathbf{T}$ that maximizes the Hellinger divergence leads to an *instance-optimal* test for minimizing sample complexity for testing between p and q with channel constraints $\mathcal{C}$: This is because if each user chooses the channel $\mathbf{T}$, the resulting sample complexity will be $\Theta\left(\frac{1}{d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)}\right)$. Thus, the instance-optimal sample complexity will be obtained by a channel $\mathbf{T}$ that attains $\max_{\mathbf{T} \in \mathcal{C}} d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)$. Note that the Hellinger divergence is convex (and thus quasi-convex) in its arguments. Apart from the Hellinger divergence, other functions of interest such as the Kullback–Leibler divergence or Chernoff information (which are also convex) characterize the asymptotic error rates in hypothesis testing, so finding $\mathbf{T}$ for these functions identifies instance-optimal channels in the asymptotic (large-sample) regime. Other potential functions of interest include Rényi divergences of all orders, which are quasi-convex, but not necessarily convex [EH14].

As mentioned earlier, the results of Kairouz, Oh, and Viswanath [KOV16] give a linear program with 2^k variables to find an instance-optimal channel under privacy constraints, which is computationally prohibitive. It is also unclear if their result extends when the channels are further restricted to have communication constraints in addition to privacy constraints. We now show how to improve on the guarantees of Kairouz, Oh, and Viswanath [KOV16] in the presence of communication constraints, using the structural results from the previous subsection.

Corollary 11.1.18 (Computationally efficient algorithms for maximizing quasi-convex functions)**.** *Let p and q be fixed distributions over $[k]$, let $\mathcal{C} \in \{\mathcal{T}_{\ell,k}, \mathcal{P}^{\epsilon}_{\ell,k}\}$, and let $\mathcal{A} =$*

$\{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{C}\}$. *Let* $g : \mathcal{A} \to \mathbb{R}$ *be a jointly quasi-convex function. When* $\mathcal{C} = \mathcal{T}_{\ell,k}$*, there is an algorithm that solves* $\max_{\mathbf{T}\in\mathcal{C}} g(\mathbf{T}p, \mathbf{T}q)$ *in time polynomial in* k^ℓ*. When* $\mathcal{C} = \mathcal{P}^\epsilon_{\ell,k}$*, there is an algorithm that solves* $\max_{\mathbf{T}\in\mathcal{C}} g(\mathbf{T}p, \mathbf{T}q)$ *in time polynomial in* k^{ℓ^2} *and* $2^{\ell^3 \log \ell}$.

We prove Corollary 11.1.18 in Section 11.4 and Section 11.5.3 for $\mathcal{C} = \mathcal{T}_{\ell,k}$ and $\mathcal{C} = \mathcal{P}^\epsilon_{\ell,k}$, respectively.

Remark 11.1.19. *When* ℓ *is constant, we obtain a polynomial-time algorithm for maximizing any quasi-convex function under* $\mathcal{T}_{\ell,k}$ *or* $\mathcal{P}^\epsilon_{\ell,k}$ *channel constraints. When* $\mathcal{C} = \mathcal{T}_{\ell,k}$ *and* g *is the Kullback–Leibler divergence, this exactly solves (for small* ℓ*) a problem introduced in Carpi, Garg, and Erkip [CGE21], which proposed a polynomial-time heuristic.*

Applying the above result to the Hellinger divergence d^2_{h}, we obtain the following result for simple binary hypothesis testing, proved in Section 11.5.3:

Corollary 11.1.20 (Computationally efficient algorithms for instance-optimal results under communication constraints)**.** *Let* p *and* q *be two distributions on* $[k]$*. For any* ϵ *and any integer* $\ell > 1$*, there is an algorithm that runs in time polynomial in* k^{ℓ^2} *and* $2^{\ell^3 \log \ell}$ *and outputs an* ϵ*-LDP channel* $\mathbf{T}$ *mapping from* $[k]$ *to* $[\ell]$*, such that if* N *denotes the sample complexity of hypothesis testing between* p *and* q *when each individual uses the channel* $\mathbf{T}$*, then* $N \asymp n^*(p, q, \epsilon, \ell)$.

In particular, the sample complexity with $\mathbf{T}$ *satisfies*

$$N \lesssim n^*(p, q, \epsilon) \cdot \left(1 + \frac{\log\left(n^*\left(p, q, \epsilon\right)\right)}{\ell}\right). \tag{11.8}$$

The channel $\mathbf{T}$ *may be decomposed as a deterministic threshold channel to a domain of size* $[2\ell^2]$*, followed by an* ϵ*-LDP channel from* $[2\ell^2]$ *to* $[\ell]$.

Thus, by choosing $\ell = 2$, we obtain a polynomial-time algorithm with nearly instance-optimal sample complexity (up to logarithmic factors) under just ϵ-LDP constraints.

11.1.4 Related Work

Distributed estimation has been studied extensively under resource constraints such as memory, privacy, and communication. Typically, this line of research considers problems of interest such as distribution estimation [RT70; LR86; CKO21; BHÖ20], identity or independence testing [ACT20a; ACT20b; ACFST21], and parameter estimation [Hel74; DJWZ14; DJW18; BGMNW16; DR19; BCÖ20; DKPP22], and identifies minimax-optimal bounds on the error or sample complexity. In what follows, we limit our discussion to related work on hypothesis testing under resource constraints.

For *memory-constrained hypothesis testing*, the earliest works in Cover [Cov69] and Hellman and Cover [HC73a] derived tight bounds on the memory size needed to perform asymptotically error-free testing. Hellman and Cover [HC71] also highlighted the benefits of randomized algorithms. These benefits were also noted in recent work by Berg, Ordentlich, and Shayevitz [BOS20], which considered the error exponent in terms of the memory size. Recently, Braverman, Garg, and Zamir [BGZ22] showed tight bounds on the memory size needed to test between two Bernoulli distributions.

Communication-constrained hypothesis testing has two different interpretations. In the information theory literature, Berger [Ber79], Ahlswede and Csiszár [AC86], and Amari and Han [AH98] considered a family of problems where two nodes, one which only observes X_i's and the other which only observes Y_i's, try to distinguish between P_{XY} and Q_{XY}. Communication between the nodes occurs over rate-limited channels. The second interpretation, also called "decentralized detection" in Tsitsiklis [Tsi88], is more relevant to this work. Here, the observed X_i's are distributed amongst different nodes (one observation per node) that communicate a finite number of messages (bits) to a central node, which needs to determine the hypothesis. Tsitsiklis [Tsi88; Tsi93] identified the optimal decision rules for individual nodes and considered asymptotic error rates in terms of the number of bits. These results were recently extended to the nonasymptotic

regime in Pensia, Jog, and Loh [PJL22; PLJ22].

Privacy-constrained hypothesis testing has been studied in the asymptotic and nonasymptotic regimes under different notions of privacy. The local privacy setting, which is relevant to this paper, is similar to the decentralized detection model in Tsitsiklis [Tsi93], except that the each node's communication to the central server is private. This is achieved by passing observations through private channels. Liao, Sankar, Calmon, and Tan [LSCT17; LSTC17] considered maximizing the error exponent under local privacy notions defined via maximal leakage and mutual information. Sheffet [She18] analyzed the performance of the randomized response method for LDP for hypothesis testing. Gopi, Kamath, Kulkarni, Nikolov, Wu, and Zhang [GKKNWZ20] showed that M-ary hypothesis testing under pure LDP constraints requires exponentially more samples ($\Omega(M)$ instead of $O(\log M)$). Closely related to the instance-optimal algorithms in our paper, Kairouz, Oh, and Viswanath [KOV16] presented an algorithm to find LDP channels that maximize the output divergence for two fixed probability distributions at the channel input; the proposed algorithm runs in time exponential in the domain size of the input distributions.[29] Note that divergences are directly related to error exponents and sample complexities in binary hypothesis testing. The results of Kairouz, Oh, and Viswanath [KOV16] on extreme points of the polytope of LDP channels were strengthened in Holohan, Leith, and Mason [HLM17], which characterized the extreme points in special cases. We were able to find only two other papers that consider instance optimality, but in rather special settings [GGKMZ21; AFT22]. For simple binary hypothesis testing in the *global* differential privacy setting, Canonne, Kamath, McMillan, Smith, and Ullman [CKMSU19] identified the optimal test and corresponding sample complexity. Bun, Kamath, Steinke, and Wu [BKSW19] showed that $O(\log M)$ samples are enough for M-ary hypothesis testing in the global

[29] We remark, however, that the algorithm in Kairouz, Oh, and Viswanath [KOV16] is applicable to a wider class of objective functions, which they term "sublinear."

differential privacy setting.

11.1.5 Organization

This paper is organized as follows: Section 11.2 records standard results. Section 11.3 focuses on the sample complexity of hypothesis testing under privacy constraints. Section 11.4 considers extreme points of the joint range under communication constraints. Section 11.5 characterizes the extreme points under both privacy and communication constraints. Section 11.6 explores other notions of privacy beyond pure LDP. Finally, we conclude with a discussion in Section 11.7. We defer proofs of some intermediate results to the appendices.

11.2 Preliminaries and Facts

Notation: Throughout this paper, we will focus on discrete distributions. For a natural number $k \in \mathbb{N}$, we use $[k]$ to denote the set $\{1, \ldots, k\}$ and Δ_k to denote the set of distributions over $[k]$. We represent a probability distribution $p \in \Delta_k$ as a vector in $\mathbb{R}^k$. Thus, p_i denotes the probability of element i under p. Given two distributions p and q, let $d_{\mathrm{TV}}(p, q) := \frac{1}{2}\sum_i |p_i - q_i|$ and $d^2_{\mathrm{h}}(p, q) := \sum_i (\sqrt{p_i} - \sqrt{q_i})^2$ denote the total variation distance and Hellinger divergence between p and q, respectively.

We denote channels with bold letters such as $\mathbf{T}$. As the channels between discrete distributions can be represented by rectangular column-stochastic matrices (each column is nonnegative and sums to one), we also use bold capital letters, such as $\mathbf{T}$, to denote the corresponding matrices. In particular, if a channel $\mathbf{T}$ is from $[k]$ to $[\ell]$, we denote it by an $\ell \times k$ matrix, where each of the k columns is in Δ_ℓ. In the same vein, for a column index $c \in [k]$ and a row index $r \in [\ell]$, we use $\mathbf{T}(r, c)$ to refer to the entry at the corresponding location. For a channel $\mathbf{T} : \mathcal{X} \to \mathcal{Y}$ and a distribution p over $\mathcal{X}$, we use $\mathbf{T}p$ to denote the distribution over $\mathcal{Y}$ when $X \sim p$ passes through the channel $\mathbf{T}$. In the notation above,

when p is a distribution over $[k]$, represented as a vector in $\mathbb{R}^k$, and $\mathbf{T}$ is a channel from $[k] \to [\ell]$, represented as a matrix $\mathbf{T} \in \mathbb{R}^{\ell \times k}$, the output distribution $\mathbf{T}p$ corresponds to the usual matrix-vector product. We shall also use $\mathbf{T}$ to denote the stochastic map transforming the channel input X to the channel output $Y = \mathbf{T}(X)$. Similarly, for two channels $\mathbf{T}_1$ and $\mathbf{T}_2$ from $[k_1]$ to $[k_2]$ and $[k_2]$ to $[k_3]$, respectively, the channel $\mathbf{T}_3$ from $[k_1]$ to $[k_3]$ that corresponds to applying $\mathbf{T}_2$ to the output of $\mathbf{T}_1$ is equal to the matrix product $\mathbf{T}_2 \times \mathbf{T}_1$.

Let $\mathcal{T}_{\ell,k}$ be the set of all channels that map from $[k]$ to $[\ell]$. We use $\mathcal{T}_{\ell,k}^{\texttt{thresh}}$ to denote the subset of $\mathcal{T}_{\ell,k}$ that corresponds to threshold channels (cf. Definition 11.1.14). We use $\mathcal{P}_{\ell,k}^{\epsilon}$ to denote the set of all ϵ-LDP channels from $[k]$ to $[\ell]$. Recall that for two distributions p and q, we use $n^*(p, q, \epsilon)$ (respectively, $n^*(p, q, \epsilon, \ell)$) to denote the sample complexity of simple binary hypothesis testing under privacy constraints (respectively, both privacy and communication constraints).

For a set A, we use $\mathrm{conv}(A)$ to denote the convex hull of A. For a convex set A, we use $\mathrm{ext}(A)$ to denote the set of extreme points of A. Finally, we use the following notations for simplicity: (i) $\lesssim$, $\gtrsim$, and $\asymp$ to hide positive constants, and (ii) the standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$. Finally, we use $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$, and $\widetilde{\Theta}$ to hide poly-logarithmic factors in their arguments.

11.2.1 Convexity

We refer the reader to Bertsimas and Tsitsiklis [BT97] for further details. We will use the following facts repeatedly in the paper, often without mentioning them explicitly:

Fact 11.2.1 (Extreme points of linear transformations)**.** *Let $\mathcal{A}$ be a convex, compact set in a finite-dimensional space. Let $\mathbf{T}$ be a linear function on $\mathcal{A}$, and define the set $\mathcal{A}' := \{\mathbf{T}x : x \in \mathcal{A}\}$. Then $\mathcal{A}'$ is convex and compact, and $\mathrm{ext}(\mathcal{A}') \subseteq \{\mathbf{T}x : x \in \mathrm{ext}(\mathcal{A})\}$.*

Fact 11.2.2. *Let $\mathcal{A}$ be a convex, compact set. If $\mathcal{A} = \mathrm{conv}(\mathcal{B})$ for some set $\mathcal{B}$, then $\mathrm{ext}(\mathcal{A}) \subseteq \mathcal{B}$.*

Fact 11.2.3 (Number of vertices and vertex enumeration)**.** *Let $\mathcal{A} \subseteq \mathbb{R}^n$ be a bounded polytope defined by m linear inequalities. The number of vertices of $\mathcal{A}$ is at most $\binom{m}{n}$. Moreover, there is an algorithm that takes $e^{O(n \log m)}$ time and output all the vertices of $\mathcal{A}$.*[30]

Fact 11.2.4 (Extreme points of channels)**.** *The set of extreme points of $\mathcal{T}_{\ell,k}$ is the set of all deterministic channels from $[k]$ to $[\ell]$.*

11.2.2 Local Privacy

We state standard facts from the privacy literature here.

Definition 11.2.5 (Randomized response)**.** *For an integer $k \geq 2$, the k-ary* randomized response channel with privacy parameter *ϵ is a channel from $[k]$ to $[k]$ defined as follows: for any $i \in [k]$, $\mathbf{T}(i) = i$ with probability $\frac{e^\epsilon}{(k-1)+e^\epsilon}$ and $\mathbf{T}(i) = j$ with probability $\frac{1}{(k-1)+e^\epsilon}$, for any $j \in [k] \setminus \{i\}$. The standard randomized response [War65] corresponds to $k = 2$, which we denote by $\mathbf{T}^\epsilon_{\mathrm{RR}}$. We omit ϵ in the superscript when it is clear from context.*

We will also use the following result on the extreme points for Theorem 11.1.7.

Fact 11.2.6 (Extreme points of the LDP polytope in special cases [HLM17])**.** *We mention all the extreme points of $\mathcal{P}^\epsilon_{\ell,k}$ (up to permutation of rows and columns; if a channel is an extreme point, then any permutation of rows and/or columns is an extreme point) below for some special cases.*

1. (*Trivial extreme points*) *A channel with one row of all ones and the rest of the rows with zero values is always an extreme point of $\mathcal{P}^\epsilon_{\ell,k}$. We call such extreme points* trivial.

[30] Throughout this paper, we assume the bit-complexity of linear inequalities is bounded.

2. ($\ell = 2$ *and* $k \geq 2$) *All non-trivial extreme points of* $\mathcal{P}^{\epsilon}_{2,k}$ *are of the form* (*up to permutation of rows*)*:*

$$\begin{bmatrix} a & a & \cdots & a & 1-a & 1-a & \cdots & 1-a \\ 1-a & 1-a & \cdots & 1-a & a & a & \cdots & a \end{bmatrix},$$

where $a/(1-a) = e^{\epsilon}$*. In other words, the columns are of only two types, containing* a *and* $1-a$*.*

3. ($\ell = 3$ *and* $k = 3$) *There are two types of non-trivial extreme points of* $\mathcal{P}^{\epsilon}_{3,3}$*: one with two nonzero rows and another with three nonzero rows. For the former, the nonzero rows are exactly the extreme points of* $\mathcal{P}^{\epsilon}_{2,3}$*. For the latter, two extreme points exist, of the following form:*

$$\begin{bmatrix} 1-2a & a & a \\ a & 1-2a & a \\ a & a & 1-2a \end{bmatrix},$$

one with $\frac{1-2a}{a} = e^{\epsilon}$ *and one with* $\frac{a}{1-2a} = e^{\epsilon}$*. The case of* $\frac{1-2a}{a} = e^{\epsilon}$ *corresponds to the usual randomized response* (*Definition 11.2.5*).

11.2.3 Hypothesis Testing

In this section, we state some standard facts regarding hypothesis testing and divergences that will be used repeatedly.

Fact 11.2.7 (Hypothesis testing and divergences; see, for example, Tsybakov [Tsy09])**.** *Let* p *and* q *be two arbitrary distributions. Then:*

1. *We have* $d^2_{\mathrm{TV}}(p,q) \leq d^2_{\mathrm{h}}(p,q) \leq 2d_{\mathrm{TV}}(p,q)$*.*

2. (*Sample complexity of non-private hypothesis testing*) *We have* $n^*(p, q) \asymp \frac{1}{d_{\mathrm{h}}^2(p,q)}$.

3. (*Sample complexity in the high-privacy regime*) *For every* $\epsilon \leq 1$, *we have* $n^*(p, q, \epsilon) \asymp \frac{1}{\epsilon^2 d_{\mathrm{TV}}^2(p,q)}$. *See the references* [DJW18, *Theorem 1*], [AZ22, *Theorem 2*], *and* [JMNR19, *Theorem 5.1*].

4. (*Restricting the size of the output domain*) *Let* p *and* q *be distributions over* $[k]$. *Then* $n^*(p, q, \epsilon) \asymp n^*(p, q, \epsilon, k)$. *This follows by applying Theorem 2 in Kairouz, Oh, and Viswanath* [KOV16] *to* $d_{\mathrm{h}}^2(\cdot, \cdot)$.

5. (*Choice of identical channels in Definition 11.1.2*) *Let* $\mathbf{T}$ *be a channel that maximizes* $d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)$ *among all channels in* $\mathcal{C}$. *Then the sample complexity of hypothesis testing under the channel constraints of* $\mathcal{C}$ *is* $\Theta\left(\frac{1}{d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)}\right)$. *See Lemma 4.2 in Pensia, Jog, and Loh* [PJL22].

Fact 11.2.8 (Preservation of Hellinger distance under communication constraints (Theorem 1 in Bhatt, Nazer, Ordentlich, and Polyanskiy [BNOP21] and Corollary 3.4 in Pensia, Jog, and Loh [PJL22]))**.** *Let* p *and* q *be two distributions on* $[k]$. *Then for any* $\ell \in \mathbb{N}$, *there exists a channel* $\mathbf{T}$ *from* $[k]$ *to* $[\ell]$, *which can be computed in time polynomial in* k, *such that*

$$d_{\mathrm{h}}^2(p, q) \lesssim d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \cdot \left(1 + \frac{\ell}{\min(k, \log(1/d_{\mathrm{h}}^2(p, q)))}\right). \tag{11.9}$$

Moreover, this bound is tight in the following sense: for every choice of $\rho \in (0, 1)$, *there exist two distributions* p *and* q *such that* $d_{\mathrm{h}}^2(p, q) \asymp \rho$, *and for every channel* $\mathbf{T} \in \mathcal{T}_{\ell,k}$, *the right-hand side of inequality Equation (11.9) is further upper-bounded by* $O(\rho)$.

11.3 Locally Private Simple Hypothesis Testing

In this section, we provide upper and lower bounds for locally private simple hypothesis testing. This section is organized as follows: In Section 11.3.1, we derive instance-optimal

bounds when both distributions are binary. We then prove minimax-optimal bounds for general distributions (with support size at least three): Lower bounds on sample complexity are proved in Section 11.3.2 and upper bounds in Section 11.3.3. Proofs of some of the technical arguments are deferred to the appendices.

11.3.1 Binary Distributions and Instance-Optimality of Randomized Response

We first consider the special case when p and q are both binary distributions. Our main result characterizes the instance-optimal sample complexity in this setting:

Theorem 11.1.6 (Sample complexity of binary distributions)**.** *Let p and q be two binary distributions. Then*

$$n^*(p, q, \epsilon) \asymp \begin{cases} \frac{1}{\epsilon^2 \cdot d_{\mathrm{TV}}^2(p,q)}, & \text{if } \epsilon \le 1, \\ \frac{1}{e^\epsilon \cdot d_{\mathrm{TV}}^2(p,q)}, & \text{if } e^\epsilon \in \left[e, \frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{TV}}^2(p,q)}\right], \\ \frac{1}{d_{\mathrm{h}}^2(p,q)}, & \text{if } e^\epsilon > \frac{d_{\mathrm{h}}^2(p,q)}{d_{\mathrm{TV}}^2(p,q)}. \end{cases} \tag{11.3}$$

By Fact 11.2.7, the proof of Theorem 11.1.6 is a consequence of the following bound on the strong data processing inequality for randomized responses:

Proposition 11.3.1 (Strong data processing inequality for Hellinger divergence)**.** *Let p and q be two binary distributions. Then*

$$\max_{\mathbf{T} \in \mathcal{P}_{2,2}^\epsilon} d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \asymp \begin{cases} \epsilon^2 \cdot d_{\mathrm{TV}}^2(p, q), & \text{if } \epsilon \le 1 \\ \min\left(e^\epsilon \cdot d_{\mathrm{TV}}^2(p, q), d_{\mathrm{h}}^2(p, q)\right), & \text{otherwise.} \end{cases}$$

Moreover, the maximum is achieved by the randomized response channel.

Proof. Let $\mathcal{A} = \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{P}^{\epsilon}_{2,2}\}$ be the joint range of p and q under ϵ-LDP privacy constraints. Since $\mathcal{A}$ is a convex set and d^2_{h} is a convex function over $\mathcal{A}$, the maximizer of d^2_{h} in $\mathcal{A}$ is an extreme point of $\mathcal{A}$. Since $\mathcal{A}$ is a linear transformation of $\mathcal{P}^{\epsilon}_{2,2}$, Fact 11.2.1 implies that any extreme point of $\mathcal{A}$ is obtained by using a channel $\mathbf{T}$ corresponding to an extreme point of $\mathcal{P}^{\epsilon}_{2,2}$. By Fact 11.2.6, the only extreme point of $\mathcal{P}^{\epsilon}_{2,2}$ is the randomized response channel $\mathbf{T}^{\epsilon}_{\mathrm{RR}}$. Thus, in the rest of the proof, we consider $\mathbf{T} = \mathbf{T}^{\epsilon}_{\mathrm{RR}}$.

By abusing notation, we will also use p and q to denote the probabilities of observing 1 under the two respective distributions. Without loss of generality, we will assume that $0 \le p \le q$ and $p \le 1/2$. We will repeatedly use the following claim, which is proved in Appendix H.4:

Claim 11.3.2 (Approximation for Hellinger divergence of binary distributions)**.** *Let $p, q \in [0, 1]$. Let $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$ be the corresponding Bernoulli distributions with $\min(p, q) \le 1/2$. Then*

$$d^2_{\mathrm{h}}\left(\mathrm{Ber}(p), \mathrm{Ber}(q)\right) \asymp \frac{d^2_{\mathrm{TV}}(\mathrm{Ber}(p), \mathrm{Ber}(q))}{\max(p, q)}.$$

Applying Claim 11.3.2, we obtain

$$d^2_{\mathrm{h}}(p, q) \asymp \frac{d^2_{\mathrm{TV}}(p, q)}{q}. \tag{11.10}$$

We know that the transformed distributions $p' := \mathbf{T}^{\epsilon}_{\mathrm{RR}}p$ and $q' := \mathbf{T}^{\epsilon}_{\mathrm{RR}}q$ are binary distributions; by abusing notation, let p' and q' also be the corresponding real-valued parameters associated with these binary distributions. By the definition of the randomized response, we have

$$p' := \frac{p(e^{\epsilon} - 1) + 1}{1 + e^{\epsilon}}, \quad \text{and} \quad q' := \frac{q(e^{\epsilon} - 1) + 1}{1 + e^{\epsilon}}. \tag{11.11}$$

Consequently, we have $0 \le p' \le q'$ and $p' \le 1/2$. We directly see that

$$d_{\mathrm{TV}}(p', q') = q' - p' = \frac{(q-p)(e^\epsilon - 1)}{e^\epsilon + 1} = d_{\mathrm{TV}}(p, q) \cdot \frac{e^\epsilon - 1}{e^\epsilon + 1}.$$

We now apply Claim 11.3.2 below to the distributions p' and q':

$$\begin{aligned}
d_{\mathrm{h}}^2(p', q') &\asymp \frac{d_{\mathrm{TV}}^2(p', q')}{q'} \\
&= d_{\mathrm{TV}}^2(p, q) \cdot \left(\frac{e^\epsilon - 1}{e^\epsilon + 1}\right)^2 \cdot \frac{1 + e^\epsilon}{q(e^\epsilon - 1) + 1} \\
&\qquad \text{(using } \textit{Equation (11.11)}\text{)} \\
&\asymp d_{\mathrm{TV}}^2(p, q) \cdot \frac{(e^\epsilon - 1)^2}{e^\epsilon + 1} \cdot \min\left(1, \frac{1}{q(e^\epsilon - 1)}\right) \\
&\qquad \left(\text{using } \frac{1}{a+b} \asymp \min\left(\frac{1}{a}, \frac{1}{b}\right) \text{ for } a, b > 0\right) \\
&\asymp d_{\mathrm{TV}}^2(p, q) \cdot \frac{(e^\epsilon - 1)^2}{e^\epsilon + 1} \cdot \min\left(1, \frac{d_{\mathrm{h}}^2(p, q)}{d_{\mathrm{TV}}^2(p, q)(e^\epsilon - 1)}\right) \\
&\qquad \left(\text{using } \textit{Equation (11.10)} \text{ and } \min(1, a) \asymp \min(1, b) \text{ if } a \asymp b\right) \\
&\asymp \min\left(d_{\mathrm{TV}}^2(p, q) \cdot \frac{(e^\epsilon - 1)^2}{e^\epsilon + 1}, d_{\mathrm{h}}^2(p, q) \cdot \frac{e^\epsilon - 1}{e^\epsilon + 1}\right) \\
&\asymp \begin{cases} \min\left(d_{\mathrm{TV}}^2(p, q) \cdot \epsilon^2, d_{\mathrm{h}}^2(p, q) \cdot \epsilon\right), & \text{if } \epsilon \le 1, \\ \min\left(d_{\mathrm{TV}}^2(p, q) \cdot e^\epsilon, d_{\mathrm{h}}^2(p, q)\right), & \text{otherwise,} \end{cases} \\
&\qquad \text{(using } e^\epsilon - 1 \asymp \epsilon \text{ for } \epsilon \le 1 \text{ and } e^\epsilon \text{ otherwise)} \\
&\asymp \begin{cases} \epsilon^2 d_{\mathrm{TV}}^2(p, q), & \text{if } \epsilon \le 1, \\ \min\left(d_{\mathrm{TV}}^2(p, q) e^\epsilon, d_{\mathrm{h}}^2(p, q)\right), & \text{otherwise,} \end{cases}
\end{aligned}$$

where the last step uses the inequality $d_{\mathrm{TV}}^2(p, q) \le d_{\mathrm{h}}^2(p, q)$ from Fact 11.2.7. □

11.3.2 General Distributions: Lower Bounds and Higher Cost of Privacy

In this section, we establish lower bounds for the sample complexity of private hypothesis testing for general distributions. In the subsequent section, the lower bounds will be shown to be tight up to logarithmic factors.

We formally state the lower bound in the statement below:

Theorem 11.1.7 (Sample complexity lower bound for general distributions)**.** *For any $\rho \in (0, 0.5)$ and $\nu \in (0, 0.5)$ such that $2\nu^2 \le \rho \le \nu$, there exist ternary distributions p and q such that $d_{\mathrm{h}}^2(p, q) = \rho$, $d_{\mathrm{TV}}(p, q) = \nu$, and the sample complexity behaves as*

$$n^*(p, q, \epsilon) \asymp \begin{cases} \frac{1}{\epsilon^2 \cdot d_{\mathrm{TV}}^2(p,q)}, & \text{if } \epsilon \le 1, \\ \min\left(\frac{1}{d_{\mathrm{TV}}^2(p,q)}, \frac{1}{e^\epsilon \cdot d_{\mathrm{h}}^4(p,q)}\right), & \text{if } e^\epsilon \in \left[e, \frac{1}{d_{\mathrm{h}}^2(p,q)}\right], \\ \frac{1}{d_{\mathrm{h}}^2(p,q)}, & \text{if } e^\epsilon > \frac{1}{d_{\mathrm{h}}^2(p,q)}. \end{cases} \tag{11.4}$$

We provide the proof below. We refer the reader to Remark 11.1.8 for further discussion on differences between the worst-case sample complexity of general distributions and the sample complexity of binary distributions (cf. Theorem 11.1.6). We note that a similar construction is mentioned in Canonne, Kamath, McMillan, Smith, and Ullman [CKMSU19, Section 1.3]; however, their focus is on the *central model* of differential privacy.

11.3.2.1 Proof of Theorem 11.1.7

Proof. The case when $\epsilon \le 1$ follows from Fact 11.2.7. Thus, we set $\epsilon \ge 1$ in the remainder of this section. We start with a helpful approximation for computing the Hellinger divergence, proved in Appendix H.4:

Claim 11.3.3 (Additive approximation for $\sqrt{\cdot}$)**.** *There exist constants $0 < c_1 \le c_2$ such that for $0 < y \le x$, we have $c_1 \cdot \frac{y^2}{x} \le (\sqrt{x} - \sqrt{x-y})^2 \le c_2 \cdot \frac{y^2}{x}$.*

For some $\gamma \in (0, 0.25)$ and $\delta > 0$ to be decided later, let p and q be the following ternary distributions:

$$p = \begin{bmatrix} 0 \\ 1/2 \\ 1/2 \end{bmatrix}, \quad \text{and} \quad q = \begin{bmatrix} 2\gamma^{1+\delta} \\ 1/2 + \gamma - \gamma^{1+\delta} \\ 1/2 - \gamma - \gamma^{1+\delta} \end{bmatrix}.$$

Since $\gamma \le 0.25$ and $\delta \ge 0$, these two are valid distributions.

Observe that $d_{\mathrm{TV}}(p, q) = \gamma + \gamma^{1+\delta} \asymp \gamma$ and and $d_{\mathrm{h}}^2(p, q) \asymp \gamma^{1+\delta}$ by Claim 11.3.3. We choose γ and δ such that $\nu = d_{\mathrm{TV}}(p, q)$ and $\rho = d_{\mathrm{h}}^2(p, q)$. Such a choice of γ and δ can be made by the argument given in Appendix H.4.2 as long as $\nu \in (0, 0.5)$ and $\rho \in [2\nu^2, \nu]$. Thus, these two distributions satisfy the first two conditions of the theorem statement.

In the rest of the proof, we will use the facts that $\gamma^{1+\delta} \asymp \rho$ and $\gamma \asymp \nu$. In particular, we have $\gamma^2 \lesssim \gamma^{1+\delta} \lesssim \gamma$.

Since both p and q are supported on $[3]$, we can restrict our attention to ternary output channels (see Fact 11.2.7). Recall that $\mathcal{P}_{3,3}^{\epsilon}$ is the set of all ϵ-LDP channels from $[3]$ to $[3]$. We will establish the following result: for all ϵ such that $e \le e^{\epsilon} \lesssim \frac{1}{d_{\mathrm{h}}^2(p,q)}$, we have

$$\max_{\mathbf{T} \in \mathcal{P}_{3,3}^{\epsilon}} d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \asymp \max\left(d_{\mathrm{TV}}^2(p, q), d_{\mathrm{h}}^4(p, q)e^{\epsilon}\right) \asymp \max(\gamma^2, e^{\epsilon}\gamma^{2+2\delta}). \tag{11.12}$$

By Fact 11.2.7, equation Equation (11.12) implies that for $e \le e^{\epsilon} \lesssim \frac{1}{d_{\mathrm{h}}^2(p,q)}$, we have

$$n^*(p, q, \epsilon) \asymp \min\left(\frac{1}{d_{\mathrm{TV}}^2(p, q)}, \frac{1}{e^{\epsilon} \cdot d_{\mathrm{h}}^4(p, q)}\right). \tag{11.13}$$

Let ϵ_0 be the right endpoint of the range for ϵ above, i.e., $e^{\epsilon_0} \asymp \frac{1}{d_{\mathrm{h}}^2(p,q)}$. Then equation Equation (11.13) shows that $n^*(p, q, \epsilon_0) \asymp 1/d_{\mathrm{h}}^2(p, q) \asymp n^*(p, q)$. Since for any ϵ such

that $\epsilon > \epsilon_0$, we have $n^*(p, q, \epsilon) \in [n^*(p, q, \epsilon_0), n^*(p, q)]$, the desired conclusion in equation Equation (11.4) holds for $\epsilon > \epsilon_0$, as well. Thus, in the remainder of this proof, we will focus on establishing equation Equation (11.12).

Since $d_{\mathrm{h}}^2(\cdot, \cdot)$ is a convex, bounded function and the set of ϵ-LDP channels is a convex polytope, it suffices to restrict our attention only to the extreme points of the polytope. As mentioned in Fact 11.2.6, these extreme points are of three types:

Case I. (Exactly one nonzero row) Any such extreme point $\mathbf{T}$ maps the entire domain to a single point with probability 1. After transformation under this channel, all distributions become indistinguishable, giving $d_{\mathrm{h}}(\mathbf{T}p, \mathbf{T}q) = 0$.

Case II. (Exactly two nonzero rows) This corresponds to the case when $\mathbf{T} = \mathbf{T}_{\mathrm{RR}}^{\epsilon} \times \mathbf{T}'$, where $\mathbf{T}'$ is a deterministic threshold channel from $[3]$ to $[2]$.[31] There are two non-trivial options for choosing $\mathbf{T}'$, which we analyze below.

The first choice of $\mathbf{T}'$ maps $\{1\}$ and $\{2, 3\}$ to different elements. The transformed distributions p' and q' are $[0, 1]$ and $[2\gamma^{1+\delta}, 1 - 2\gamma^{1+\delta}]$, respectively. Using Claim 11.3.3, we obtain $d_{\mathrm{h}}^2(p', q') \asymp \gamma^{1+\delta}$ and $d_{\mathrm{TV}}(p', q') \asymp \gamma^{1+\delta}$. Let p'' and q'' be the corresponding distributions after applying the randomized response with parameter ϵ. Since p' and q' are binary distributions, we can apply Proposition 11.3.1 to obtain

$$d_{\mathrm{h}}^2(p'', q'') \asymp \min(d_{\mathrm{h}}^2(p', q'), e^{\epsilon} d_{\mathrm{TV}}^2(p', q')) \asymp \min(\gamma^{1+\delta}, e^{\epsilon}\gamma^{2+2\delta}) \asymp \gamma^{1+\delta} \cdot \min(1, e^{\epsilon}\gamma^{1+\delta}),$$

which is equal to $e^{\epsilon} \cdot \gamma^{2+2\delta}$ in the regime of interest and consistent with the desired expression in equation Equation (11.12).

The second choice of $\mathbf{T}'$ maps $\{1, 2\}$ and $\{3\}$ to different elements. The transformed distributions p' and q' are $[1/2, 1/2]$ and $[1/2 + \gamma + \gamma^{1+\delta}, 1/2 - \gamma - \gamma^{1+\delta}]$, respectively. Applying Claim 11.3.3, we observe that $d_{\mathrm{h}}^2(p', q') \asymp \gamma^2$ and $d_{\mathrm{TV}}(p', q') \asymp \gamma$. Let p'' and

[31] We use Theorem 11.1.17 to restrict our attention only to threshold channels.

q'' be the corresponding distributions after applying the randomized response with parameter ϵ. Applying Proposition 11.3.1, we obtain

$$d_{\mathrm{h}}^2(p'', q'') \asymp \min(d_{\mathrm{h}}^2(p', q'), e^{\epsilon} d_{\mathrm{TV}}^2(p', q')) \asymp \min(\gamma^2, e^{\epsilon}\gamma^2) \asymp \gamma^2$$

in the regime of interest. Again, this is consistent with equation Equation (11.12).

Case III. (All nonzero rows) There are two extreme points of this type (up to a permutation of the rows), both of the following form:

$$\mathbf{T}_1 = \begin{bmatrix} 1-2\alpha & \alpha & \alpha \\ \alpha & 1-2\alpha & \alpha \\ \alpha & \alpha & 1-2\alpha \end{bmatrix}.$$

For the first extreme point, α satisfies $\frac{1-2\alpha}{\alpha} = e^{\epsilon}$, while the second extreme point has $\frac{\alpha}{1-2\alpha} = e^{\epsilon}$. These channels are relatively easy to analyze, since they transform the distributions element-wise: each entry x of the original distribution is transformed to $\alpha + x(1-3\alpha)$. Consequently, the transformed distributions p' and q' are

$$p' = \begin{bmatrix} \alpha \\ \frac{1-\alpha}{2} \\ \frac{1-\alpha}{2} \end{bmatrix}, \quad \text{and} \quad q' = \begin{bmatrix} \alpha + 2\gamma^{1+\delta}(1-3\alpha) \\ \frac{1-\alpha}{2} + (\gamma - \gamma^{1+\delta})(1-3\alpha) \\ \frac{1-\alpha}{2} + (-\gamma - \gamma^{1+\delta})(1-3\alpha) \end{bmatrix}. \tag{11.14}$$

We now compute the Hellinger divergence between these two distributions for both the extreme points.

Let us first consider the case where $\frac{1-2\alpha}{\alpha} = e^{\epsilon}$. Then $\alpha = \frac{1}{2+e^{\epsilon}} \asymp e^{-\epsilon}$, since $\epsilon \geq 0$. Thus, in the desired range of $e^{\epsilon} \lesssim \gamma^{-(1+\delta)}$, the parameter α satisfies $\alpha \gtrsim \gamma^{1+\delta}$. We will now calculate the Hellinger divergence between p' and q' in equation Equation (11.14) by analyzing the contribution from each of the three terms in the sum $\sum_{i=1}^{3}(\sqrt{p_i'} - \sqrt{q_i'})^2$. For

the first term, we apply Claim 11.3.3 with $x = \alpha + 2\gamma^{1+\delta}(1-3\alpha)$ and $y = 2\gamma^{1+\delta}(1-3\alpha)$, to see that its contribution is $\Theta(\gamma^{2+2\delta}/\alpha) \asymp e^{\epsilon}\gamma^{2+2\delta}$ (since $1 - 3\alpha \geq 0.1$). Applying Claim 11.3.3 again, we see that the contributions from the second and third elements are $\Theta(\gamma^2)$, since $\alpha \ll 1$ and $\gamma \ll 1$, respectively. Overall, the Hellinger divergence is $O\left(\max(\gamma^2, e^{\epsilon}\gamma^{2+2\delta})\right)$, which satisfies equation Equation (11.12).

Finally, let us consider the case when $\frac{\alpha}{1-2\alpha} = e^{\epsilon}$. We set $\beta = 1 - 2\alpha$. Then $\beta = 1/(1+2e^{\epsilon})$, which is much less than 1 in the desired range and is of the order of $e^{-\epsilon}$. Thus, each entry x of the distribution is mapped to $\frac{1}{2}(1-\beta+x(3\beta-1))$. The transformed distributions are

$$p' = \frac{1}{2} \cdot \begin{bmatrix} 1-\beta \\ \frac{1+\beta}{2} \\ \frac{1+\beta}{2} \end{bmatrix}, \quad \text{and} \quad q' = \frac{1}{2} \cdot \begin{bmatrix} 1-\beta+2\gamma^{1+\delta}(3\beta-1) \\ \frac{1+\beta}{2} + (\gamma - \gamma^{1+\delta})(3\beta-1) \\ \frac{1+\beta}{2} + (-\gamma - \gamma^{1+\delta})(3\beta-1) \end{bmatrix}. \tag{11.15}$$

As β is much less than 1 in the desired range of ϵ, we can apply Claim 11.3.3 to see that contribution of the first element is $\Theta(\gamma^{2+2\delta})$, and the contributions of both the second and third elements are $\Theta(\gamma^2)$. Overall, the Hellinger divergence is $\Theta(\gamma^2)$, which is again consistent with equation Equation (11.12).

Combining all the cases above, the maximum Hellinger divergence after applying any ϵ-LDP channel is $\Theta(\gamma^2 \cdot \max(1, e^{\epsilon}\gamma^{2\delta}))$, as desired. □

11.3.3 General Distributions: Upper Bounds and Minimax Optimality

We now demonstrate an algorithm that finds a private channel matching the minimax rate in Theorem 11.1.7 up to logarithmic factors. Moreover, the proposed algorithm is both computationally efficient and communication efficient.

Theorem 11.1.9 (Sample complexity upper bounds and an efficient algorithm for hypothesis testing for general distributions)**.** *Let p and q be two distributions on $[k]$. Let $\epsilon > 0$. Then the sample complexity behaves as*

$$n^*(p,q,\epsilon) \lesssim \begin{cases} \frac{1}{\epsilon^2 \cdot d_{\mathrm{TV}}^2(p,q)}, & \text{if } \epsilon \le 1, \\ \min\left(\frac{1}{d_{\mathrm{TV}}^2(p,q)}, \frac{\alpha^2}{e^\epsilon \cdot d_{\mathrm{h}}^4(p,q)}\right), & \text{if } e^\epsilon \in \left(e, \frac{\alpha}{d_{\mathrm{h}}^2(p,q)}\right], \\ \frac{\alpha}{d_{\mathrm{h}}^2(p,q)}, & \text{if } e^\epsilon > \frac{\alpha}{d_{\mathrm{h}}^2(p,q)}, \end{cases} \tag{11.5}$$

where $\alpha \lesssim \log(1/d_{\mathrm{h}}^2(p,q)) \asymp \log(n^*(p,q))$.

Moreover, the rates above are achieved by an ϵ-LDP channel $\mathbf{T}$ that maps $[k]$ to $[2]$ and can be found in time polynomial in k, for any choice of p, q, and ϵ.

In comparison with Theorem 11.1.7, we see that the test above is minimax optimal up to logarithmic factors over the class of distributions with fixed Hellinger divergence and total variation distance. The channel $\mathbf{T}$ satisfying this rate is of the following simple form: a deterministic binary channel $\mathbf{T}'$, followed by the randomized response. In fact, we can take $\mathbf{T}'$ to be either Scheffe's test (which preserves the total variation distance) or the binary channel from Fact 11.2.8 (which preserves the Hellinger divergence), whichever of the two is better. We provide the complete proof in Section 11.3.3.1.

One obvious shortcoming of Theorem 11.1.9 is that even when $\epsilon \to \infty$, the test does not recover the optimal sample complexity of $1/d_{\mathrm{h}}^2(p,q)$, due to the logarithmic multiplier α. We now consider the case when $e^\epsilon \gtrsim \frac{1}{d_{\mathrm{h}}^2(p,q)}$ and exhibit a channel that achieves the optimal sample complexity as soon as $e^\epsilon \gtrsim \frac{1}{d_{\mathrm{h}}^2(p,q)} \log\left(\frac{1}{d_{\mathrm{h}}^2(p,q)}\right)$. Thus, privacy can be attained essentially for free in this regime.

Theorem 11.1.13. *Let p and q be two distributions on $[k]$, and let $e^\epsilon \gtrsim \frac{1}{d_{\mathrm{h}}^2(p,q)} \log\left(\frac{1}{d_{\mathrm{h}}^2(p,q)}\right)$. Then $n^*(p,q,\epsilon) \asymp n^*(p,q)$. Moreover, there is a channel $\mathbf{T}$ achieving this sample complexity that maps $[k]$ to a domain of size $\lceil \log(n^*(p,q)) \rceil$, and which can be computed in $\mathrm{poly}(k, \log(\lceil n^*(p,q) \rceil))$*

time.

We note that the size of the output domain of $\lceil \log(n^*(p,q)) \rceil$ is tight in the sense that any channel that achieves the sample complexity within constant factors of $n^*(p,q)$ must use an output domain of size at least $\Omega(\log(n^*(p,q)))$; this follows by the tightness of Fact 11.2.8 in the worst case. Consequently, the channel $\mathbf{T}$ achieving the rate above is roughly of the form (1) a communication-efficient channel from Fact 11.2.8 that preserves the Hellinger divergence up to constant factors, followed by (2) an ℓ-ary randomized response channel, for $\ell \geq 2$.

We give the proof of Theorem 11.1.13 in Section 11.3.3.2 and defer the proofs of some of the intermediate results to Appendix H.1.

11.3.3.1 Proof of Theorem 11.1.9

In this section, we provide the proof of Theorem 11.1.9. We first note that this result can be slightly strengthened, replacing α by $\frac{n^*_{\text{binary}}}{n^*}$, where n^*_{binary} is the sample complexity of hypothesis testing under binary communication constraints. This choice of α is smaller, by Pensia, Jog, and Loh [PJL22, Corollary 3.4 and Theorem 4.1].

Proof. The case of $\epsilon \leq 1$ follows from Fact 11.2.7; thus, we focus on the setting where $\epsilon > 1$.

We will establish these bounds via Proposition 11.3.1, by using a binary deterministic channel, followed by the binary randomized response channel. A sample complexity of $1/d^2_{\text{TV}}(p,q)$ is direct by using the channel for $\epsilon = 1$. Thus, our focus will be on the term $\frac{1}{e^\epsilon d^4_{\text{h}}(p,q)}$.

Let $\mathbf{T}' \in \mathcal{T}_{2,k}$ be a deterministic binary output channel to be decided later. Consider the channel $\mathbf{T} = \mathbf{T}^\epsilon_{\text{RR}} \times \mathbf{T}'$. By Proposition 11.3.1, we have

$$d^2_{\text{h}}\left(\mathbf{T}^\epsilon_{\text{RR}} \times \mathbf{T}'p, \mathbf{T}^\epsilon_{\text{RR}} \times \mathbf{T}'q\right) \asymp \min\left(e^\epsilon d^2_{\text{TV}}\left(\mathbf{T}'p, \mathbf{T}'q\right), d^2_{\text{h}}\left(\mathbf{T}'p, \mathbf{T}'q\right)\right)$$

$$\geq \min\left(e^{\epsilon} d_{\mathrm{h}}^4\left(\mathbf{T}'p, \mathbf{T}'q\right), d_{\mathrm{h}}^2\left(\mathbf{T}'p, \mathbf{T}'q\right)\right)$$
$$= d_{\mathrm{h}}^2\left(\mathbf{T}'p, \mathbf{T}'q\right) \cdot \min\left(e^{\epsilon} d_{\mathrm{h}}^2\left(\mathbf{T}'p, \mathbf{T}'q\right), 1\right), \tag{11.16}$$

where the first inequality uses Fact 11.2.7

If we choose the channel $\mathbf{T}'$ from Fact 11.2.8, we have $d_{\mathrm{h}}^2(\mathbf{T}'p, \mathbf{T}'q) \geq \frac{1}{\alpha} \cdot d_{\mathrm{h}}^2(p, q)$. Applying this to inequality Equation (11.16), we obtain

$$d_{\mathrm{h}}^2\left(\mathbf{T}_{\mathrm{RR}}^{\epsilon} \times \mathbf{T}'p, \mathbf{T}_{\mathrm{RR}}^{\epsilon} \times \mathbf{T}'q\right) \gtrsim \frac{1}{\alpha} \cdot d_{\mathrm{h}}^2(p, q) \cdot \min\left(\frac{1}{\alpha} \cdot e^{\epsilon} d_{\mathrm{h}}^2(p, q), 1\right).$$

By Fact 11.2.7, the sample complexity of $\mathbf{T}_{\mathrm{RR}}^{\epsilon} \times \mathbf{T}'$, which is ϵ-LDP, is at most $\frac{\alpha}{d_{\mathrm{h}}^2(p,q)} \cdot \max\left(1, \frac{\alpha}{e^{\epsilon} d_{\mathrm{h}}^2(p,q)}\right)$, which is equivalent to the desired statement.

Finally, the claim on the runtime is immediate, since the channel $\mathbf{T}'$ from Fact 11.2.8 can be found efficiently. □

11.3.3.2 Proof of Theorem 11.1.13

We will prove a slightly generalized version of Theorem 11.1.13 below that works for a wider range of ϵ:

Proposition 11.3.4. *Let p and q be two distributions on $[k]$ and $\epsilon > 1$. Then there exists an ϵ-LDP channel $\mathbf{T}$ from $[k]$ to $[\ell]$, for $\ell = \min(\lceil \log(1/d_{\mathrm{h}}^2(p, q)) \rceil, k)$, such that*

$$d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \gtrsim d_{\mathrm{h}}^2(p, q) \cdot \min\left(1, \frac{e^{\epsilon} \cdot d_{\mathrm{h}}^2(p, q)}{\log(1/d_{\mathrm{h}}^2(p, q))}\right) \cdot \min\left(1, \frac{e^{\epsilon}}{\log(1/d_{\mathrm{h}}^2(p, q))}\right).$$

Furthermore, the channel $\mathbf{T}$ can be be computed in $\mathrm{poly}(k, \ell)$ time.

By Fact 11.2.7, Proposition 11.3.4 implies the following:

$$n^*(p, q, \epsilon) \lesssim n^* \cdot \max\left(1, \frac{n^* \log(n^*)}{e^{\epsilon}}\right) \cdot \max\left(1, \frac{\log n^*}{e^{\epsilon}}\right),$$

where $n^* := n^*(p, q)$. Setting e^ϵ equal to $n^* \log(n^*)$ proves Theorem 11.1.13. Thus, we will focus on proving Proposition 11.3.4 in the rest of this section. We establish this result with the help of the following observations:

- (Lemma 11.3.5) First, we show that the randomized response preserves the contribution to the Hellinger divergence by "comparable elements" (elements whose likelihood ratio is in the interval $\left[\frac{1}{2}, 2\right]$) when ϵ is large compared to the support size. In particular, we first define the following sets:

$$A = \left\{ i \in [k] : \frac{p_i}{q_i} \in \left[\frac{1}{2}, 1\right) \right\} \text{ and } A' = \left\{ i \in [k] : \frac{p_i}{q_i} \in [1, 2] \right\}. \tag{11.17}$$

  Let $\mathbf{T}^{\epsilon,\ell}_{\mathrm{RR}}$ denote the randomized response channel from $[\ell]$ to $[\ell]$ with privacy parameter ϵ (cf. Definition 11.2.5). The following result is proved in Appendix H.1.1:

  Lemma 11.3.5 (Randomized response preserves contribution of comparable elements)**.** *Let p and q be two distributions on $[\ell]$. Suppose $\sum_{i \in A \bigcup A'} (\sqrt{q_i} - \sqrt{p_i})^2 \geq \tau$. Then $\mathbf{T}^{\epsilon,\ell}_{\mathrm{RR}}$, for $\ell \leq e^\epsilon$, satisfies*

$$d^2_{\mathrm{h}}(\mathbf{T}^{\epsilon,\ell}_{\mathrm{RR}} p, \mathbf{T}^{\epsilon,\ell}_{\mathrm{RR}} q) \gtrsim \min\left(1, e^\epsilon \frac{\tau}{\ell}\right) \cdot \tau \,.$$

  Thus, when $e^\epsilon \gtrsim \frac{\ell}{\tau}$, the randomized response preserves the original contribution of comparable elements.

- (Lemma 11.3.6) We then show in Lemma 11.3.6, proved in Appendix H.1.2, that either we can reduce the problem to the previous special case (small support size and main contribution to Hellinger divergence is from comparable elements) or to the case when the distributions are binary (where Proposition 11.3.1 is applicable and is, in a sense, the *easy case* for privacy).

Lemma 11.3.6 (Reduction to base case)**.** *Let p and q be two distributions on $[k]$. Then there is a channel $\mathbf{T}$, which can be computed in time polynomial in k, that maps $[k]$ to $[\ell]$ (for ℓ to be decided below) such that for $p' = \mathbf{T}p$ and $q' = \mathbf{T}q$, at least one of the following holds:*

1. *For any $\ell > 2$ and $\ell \leq \min\left(k, 1 + \log\left(1/d_{\mathrm{h}}^2(p,q)\right)\right)$, we have*

$$\sum_{i \in B \bigcup B'} \left(\sqrt{q'_i} - \sqrt{p'_i}\right)^2 \gtrsim d_{\mathrm{h}}^2(p,q) \cdot \frac{\ell}{\min\left(k, 1 + \log\left(1/d_{\mathrm{h}}^2(p,q)\right)\right)},$$

   where B and B' are defined analogously to A and A' in equation Equation (11.17), but with respect to distributions p' and q'.

2. *$\ell = 2$ and $d_{\mathrm{h}}^2(p',q') \gtrsim d_{\mathrm{h}}^2(p,q)$.*

We now provide the proof of Proposition 11.3.4, with the help of Lemmata 11.3.5 and 11.3.6.

Proof. (Proof of Proposition 11.3.4) The channel $\mathbf{T}$ will be of the form $\mathbf{T} = \mathbf{T}_{\mathrm{RR}}^{\epsilon,\ell} \times \mathbf{T}_1$, where $\mathbf{T}_1$ is a channel from $[k]$ to $[\ell]$ and ℓ is to be decided. The privacy of $\mathbf{T}$ is clear from the construction.

We begin by applying Lemma 11.3.6. Let $\mathbf{T}_1$ be the channel from Lemma 11.3.6 that maps from $[k]$ to $[\ell]$. Let $p' = \mathbf{T}_1 p$ and $q' = \mathbf{T}_1 q$, and define $\tilde{p} = \mathbf{T}_{\mathrm{RR}}^{\epsilon,\ell} p'$ and $\tilde{q} = \mathbf{T}_{\mathrm{RR}}^{\epsilon,\ell} q'$. The claim on runtime thus follows from Lemma 11.3.6.

Suppose for now that $\mathbf{T}_1$ from Lemma 11.3.6 is a binary channel. Then we know that $d_{\mathrm{h}}^2(p',q') \gtrsim d_{\mathrm{h}}^2(p,q)$ and $d_{\mathrm{TV}}(p',q') \gtrsim d_{\mathrm{h}}^2(p',q')$, where the latter holds by Fact 11.2.7. Applying Proposition 11.3.1, we have

$$\begin{aligned} d_{\mathrm{h}}^2(\tilde{p},\tilde{q}) &\gtrsim \min\left(d_{\mathrm{h}}^2(p',q'), e^{\epsilon} d_{\mathrm{TV}}^2(p',q')\right) \\ &\gtrsim \min\left(d_{\mathrm{h}}^2(p',q'), e^{\epsilon} d_{\mathrm{h}}^4(p',q')\right) \end{aligned}$$

$$\gtrsim d_{\mathrm{h}}^2(p,q)\min\left(1, e^{\epsilon} d_{\mathrm{h}}^2(p,q)\right),$$

which concludes the proof in this case.

We now consider the case when $\ell > 2$ in the guarantee of Lemma 11.3.6. Then the comparable elements of p' and q' preserve a significant fraction of the Hellinger divergence (depending on the chosen value of ℓ) between p and q. Let $k' = \min\left(k, 1 + \log(1/d_{\mathrm{h}}^2(p,q))\right)$, and choose ℓ to be $\min(k', e^{\epsilon})$. Then Lemma 11.3.6 implies that the contribution to the Hellinger divergence from comparable elements of p' and q' is at least τ, for $\tau \asymp d_{\mathrm{h}}^2(p,q)\frac{\ell}{k'}$. We will now apply Lemma 11.3.5 to p' and q' with the above choice of τ. Since $\ell \leq e^{\epsilon}$ by construction, applying Lemma 11.3.5 to p' and q', we obtain

$$\begin{aligned} d_{\mathrm{h}}^2(\tilde{p},\tilde{q}) &\gtrsim \tau \cdot \min\left(1, \frac{e^{\epsilon}\tau}{\ell}\right) \\ &\gtrsim d_{\mathrm{h}}^2(p,q)\frac{\ell}{k'} \cdot \min\left(1, \frac{e^{\epsilon} \cdot d_{\mathrm{h}}^2(p,q)}{\log(1/d_{\mathrm{h}}^2(p,q))}\right) \\ &\gtrsim d_{\mathrm{h}}^2(p,q) \cdot \min\left(1, \frac{e^{\epsilon}}{\log(1/d_{\mathrm{h}}^2(p,q))}\right) \cdot \min\left(1, \frac{e^{\epsilon} \cdot d_{\mathrm{h}}^2(p,q)}{\log(1/d_{\mathrm{h}}^2(p,q))}\right), \end{aligned}$$

where the last step uses the facts that $\ell = \min(e^{\epsilon}, k')$ and $k' \gtrsim \log(1/d_{\mathrm{h}}^2(p,q))$. This completes the proof. □

11.4 Extreme Points of Joint Range Under Communication Constraints

In this section, our goal is to understand the extreme points of the set $\mathcal{A} := \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{T}_{\ell,k}\}$. This will allow us to identify the structure of optimizers of quasi-convex functions over $\mathcal{A}$. The main result of this section is the following:

Theorem 11.1.16 (Extreme points of the joint range under communication constraints)**.** *Let p and q be two distributions on $[k]$. Let $\mathcal{A}$ be the set of all pairs of distributions that are obtained by passing p and q through a channel of output size ℓ, i.e.,*

$$\mathcal{A} = \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{T}_{\ell,k}\}.$$

If $(\mathbf{T}p, \mathbf{T}q)$ is an extreme point of $\mathcal{A}$, then $\mathbf{T}$ is a threshold channel.

We provide the proof of Theorem 11.1.16 in Section 11.4.1. Before proving Theorem 11.1.16, we discuss some consequences for optimizing quasi-convex functions over $\mathcal{A}$. The following result proves Corollary 11.1.18 for $\mathcal{C} = \mathcal{T}_{\ell,k}$:

Corollary 11.4.1 (Threshold channels maximize quasi-convex functions)**.** *Let p and q be two distributions on $[k]$. Let $\mathcal{A} := \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{T}_{\ell,k}\}$. Let g be a real-valued quasi-convex function over $\mathcal{A}$. Then*

$$\max_{\mathbf{T} \in \mathcal{T}_{\ell,k}} g(\mathbf{T}p, \mathbf{T}q) = \max_{\mathbf{T} \in \mathcal{T}_{\ell,k}^{\text{thresh}}} g(\mathbf{T}p, \mathbf{T}q).$$

Moreover, the above optimization problem can be solved in time $\mathrm{poly}(k^\ell)$.[32]

Proof. Observe that $\mathcal{A}$ is a closed polytope. Let $\mathcal{X}$ be the set of extreme points of $\mathcal{A}$. Observe that $\mathcal{X} \subseteq \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{T}_{\ell,k}$ and $\mathbf{T}$ is deterministic$\}$, and thus is finite. Since $\mathcal{A}$ is a closed polytope, $\mathcal{A}$ is convex hull of $\mathcal{X}$. Furthermore, the maximum of g on $\mathcal{X}$ is well-defined and finite, as $\mathcal{X}$ is a finite set. Any $y \in \mathcal{A}$ can be expressed as a convex combination $y = \sum_{x_i \in \mathcal{X}} \lambda_i x_i$. Recall that an equivalent definition of quasi-convexity is that g satisfies $g(\lambda x + (1-\lambda)y) \le \max(g(x), g(y))$ for all $\lambda \in [0,1]$. By repeatedly using this fact, we have

$$g(\lambda) = g\left(\sum_{x_i \in \mathcal{X}} \lambda_i x_i\right) \le \max_{x \in \mathcal{X}} g(x).$$

[32] Recall that g is assumed to be permutation invariant. If not, an extra factor of $\ell!$ will appear in the time complexity.

By Theorem 11.1.16, any extreme point $x \in \mathcal{X}$ is obtained by passing p and q through a threshold channel. Thus, the maximum of g over $\mathcal{X}$ is attained by passing p and q through a threshold channel. The claimed runtime is obtained by trying all possible threshold channels. □

Remark 11.4.2. *(Quasi-)convex functions of interest include all f-divergences, Rényi divergences, Chernoff information, and L_p norms. We note that the above result also holds for post-processing: For any fixed channel $\mathbf{H} \in \mathcal{T}_{\ell',\ell}$, we have*

$$\max_{\mathbf{T} \in \mathcal{T}_{\ell,k}} g(\mathbf{HT}p, \mathbf{HT}q) = \max_{\mathbf{T} \in \mathcal{T}_{\ell,k}^{\texttt{thresh}}} g(\mathbf{HT}p, \mathbf{HT}q).$$

This is because $g(\mathbf{H}p', \mathbf{H}q')$ is a quasi-convex function of $(p', q') \in \mathcal{A}$.

Remark 11.4.3. *If g is the Hellinger divergence and $\mathcal{C} = \mathcal{T}_{\ell,k}$, we can conclude the following result for the communication-constrained setting: There exists a $\mathbf{T} \in \mathcal{T}_{\ell,k}^{\texttt{thresh}}$ that attains the instance-optimal sample complexity (up to universal constants) for hypothesis testing under a communication constraint of size ℓ. This result is implied in Pensia, Jog, and Loh [PJL22] by Theorem 2.9 (which is a result from Tsitsiklis [Tsi93]) and Lemma 4.2. The above argument provides a more straightforward proof.*

11.4.1 Proof of Theorem 11.1.16

We now provide the proof of Theorem 11.1.16.

Proof. (Proof of Theorem 11.1.16) We first make the following simplifying assumption about the likelihood ratios: there is at most a single element i^* with $q_{i^*} = 0$, and for all other elements $i \in [k] \setminus \{i^*\}$, p_i/q_i is a unique value. If there are two or more elements with the same likelihood ratio, we can merge those elements into a single alphabet without loss of generality, as we explain next. Let p' and q' be the distributions after merging these elements, and let $k' \le k$ be the new cardinality. Then for any channel

$\mathbf{T} \in \mathcal{T}_{\ell,k}$, there exists another channel $\mathbf{T}' \in \mathcal{T}_{\ell,k'}$ such that $(\mathbf{T}p, \mathbf{T}q) = (\mathbf{T}'p', \mathbf{T}'q')$. We can then apply the following arguments to p' and q'. See Appendix H.4.1 for more details.

In the following, we will consider p_i/q_i to be ∞ if $q_i = 0$, and we introduce the notation $\theta_i := p_i/q_i$. We will further assume, without loss of generality, that p_i/q_i is strictly increasing in i. Since the elements are ordered with respect to the likelihood ratio, a threshold channel corresponds to a map that partitions the set $[k]$ into contiguous blocks. Formally, we have the following definition:

Definition 11.4.4 (Partitions and threshold partitions)**.** *We say that $\mathcal{S} = (S_1, S_2, \ldots, S_\ell)$ forms an* ℓ-partition *of $[k]$ if $\cup_{i=1}^{\ell} S_i = [k]$ and $S_i \cap S_j = \emptyset$ for $1 \le i \ne j \le \ell$. We say that $\mathcal{S}$ forms an* ℓ-threshold partition *of $[k]$ if in addition, for all $i < j \in \ell$, every entry of S_i is less than every entry of S_j.*

As mentioned before, channels corresponding to ℓ-threshold partitions are precisely the threshold channels up to a permutation of output labels. The channels corresponding to ℓ-partitions are the set of all deterministic channels that map $[k]$ to ℓ, which are the extreme points of $\mathcal{T}_{\ell,k}$ (cf. Fact 11.2.4).

Observe that $\mathcal{A}$ is a convex, compact set, which is a linear transformation of the convex, compact set $\mathcal{T}_{\ell,k}$, and any extreme point of $\mathcal{A}$ is of the form $(\mathbf{T}p, \mathbf{T}q)$, where $\mathbf{T}$ is an extreme point of $\mathcal{T}_{\ell,k}$ (cf. Fact 11.2.1). Now suppose $(\mathbf{T}p, \mathbf{T}q)$ is an extreme point of $\mathcal{A}$, but $\mathbf{T}$ is not a threshold channel. Thus, $\mathbf{T}$ corresponds to some ℓ-partition $\mathcal{S}$ of $[k]$ that is not an ℓ-threshold partition. We will now show that $(\mathbf{T}p, \mathbf{T}q)$ is not an extreme point of $\mathcal{A}$, by showing that there exist two distinct channels $\mathbf{T}_1 \in \mathcal{T}_{\ell,k}$ and $\mathbf{T}_2 \in \mathcal{T}_{\ell,k}$ such that the following holds:

$$\frac{1}{2} \cdot \mathbf{T}_1 p + \frac{1}{2} \cdot \mathbf{T}_2 p = \mathbf{T}p, \quad \text{and} \quad \frac{1}{2} \cdot \mathbf{T}_1 q + \frac{1}{2} \cdot \mathbf{T}_2 q = \mathbf{T}q, \tag{11.18}$$

and $\mathbf{T}_1 p \ne \mathbf{T}p$.

Since $\mathcal{S}$ is not a ℓ-threshold partition, there exist $1 \leq a < b < c \leq k$ and $m \neq n$ in $[\ell]$ such that $a, c \in S_m$ and $b \in S_n$, and $p_a/q_a < p_b/q_b < p_c/q_c$. Among q_a, q_b, and q_c, only q_c is potentially zero. Suppose for now that $q_c \neq 0$; we will consider the alternative case shortly.

For some $\epsilon_1 \in (0,1)$ and $\epsilon_2 \in (0,1)$ to be determined later, let $\mathbf{T}_1$ be the following channel:

1. For $x \notin \{a, b\}$, $\mathbf{T}_1$ maps x to $\mathbf{T}(x)$.
2. For $x = a$ (respectively b), $\mathbf{T}_1$ maps x to m (respectively n) with probability $1 - \epsilon_1$ (respectively $1 - \epsilon_2$) and to n (respectively m) with probability ϵ_1 (respectively ϵ_2).

Thus, the channels $\mathbf{T}$ and $\mathbf{T}_1$ have all columns identical, except for those corresponding to inputs a and b. Let v_i be the i^{th} column of $\mathbf{T}$. Observe that v_a is a degenerate distribution at $m \in [\ell]$ and v_b is a degenerate distribution at $n \in [\ell]$ (equivalently, $\mathbf{T}(m, a) = 1$ and $\mathbf{T}(n, b) = 1$). Thus, we can write

$$\begin{aligned}
\mathbf{T}_1 q &= \mathbf{T} q + (\epsilon_2 q_b - \epsilon_1 q_a)(v_a - v_b), \\
\mathbf{T}_1 p &= \mathbf{T} p + (\epsilon_2 p_b - \epsilon_1 p_a)(v_a - v_b).
\end{aligned}$$

If we choose $\epsilon_1 q_a = \epsilon_2 q_b$, we have $\mathbf{T}_1 q = \mathbf{T} q$ and

$$\begin{aligned}
\mathbf{T}_1 p &= \mathbf{T} p + (\epsilon_2 p_b - \epsilon_1 p_a)(v_a - v_b) \\
&= \mathbf{T} p + (\epsilon_2 q_b \theta_b - \epsilon_1 q_a \theta_a)(v_a - v_b) \\
&= \mathbf{T} p + \epsilon_1 q_a (\theta_b - \theta_a)(v_a - v_b). \qquad (11.19)
\end{aligned}$$

Recall that $\theta_b > \theta_a$, as mentioned above.

We now define $\mathbf{T}_2$. For some $\epsilon_3 \in (0,1)$ and $\epsilon_4 \in (0,1)$ to be decided later, we have:

1. For $x \notin \{b, c\}$, $\mathbf{T}_2$ maps x to $\mathbf{T}(x)$.

2. For $x = c$ (respectively b), $\mathbf{T}_2$ maps x to m (respectively n) with probability $1 - \epsilon_3$ (respectively $1 - \epsilon_4$) and to n (respectively m) with probability ϵ_3 (respectively ϵ_4).

With the same arguments as before, we have

$$\mathbf{T}_2 q = \mathbf{T} q + (\epsilon_4 q_b - \epsilon_3 q_c)(v_c - v_b),$$
$$\mathbf{T}_1 p = \mathbf{T} p + (\epsilon_4 p_b - \epsilon_3 p_c)(v_c - v_b).$$

If we choose $\epsilon_3 q_c = \epsilon_4 q_b$, we have $\mathbf{T}_2 q = \mathbf{T} q$ and

$$\begin{aligned}\mathbf{T}_2 p &= \mathbf{T} p + (\epsilon_4 q_b \theta_b - \epsilon_3 q_c \theta_c)(v_c - v_b) \\ &= \mathbf{T} p + \epsilon_3 q_c (\theta_b - \theta_c)(v_c - v_b) \\ &= \mathbf{T} p + \epsilon_3 q_c (\theta_b - \theta_c)(v_a - v_b), \end{aligned} \tag{11.20}$$

where the last line follows by the fact that $v_a = v_c$, since $\mathbf{T}$ maps both a and c to m almost surely.

Let $\epsilon_1 \in (0, 1)$ and $\epsilon_3 \in (0, 1)$ be such that $\epsilon_1 q_a(\theta_b - \theta_a) = -\epsilon_3 q_c(\theta_b - \theta_c)$. Such a choice always exists because $\theta_b - \theta_a$ and $-(\theta_b - \theta_c)$ are both strictly positive and finite. Then equations Equation (11.19) and Equation (11.20) imply that $\mathbf{T} p = \frac{1}{2}\mathbf{T}_1 p + \frac{1}{2}\mathbf{T}_2 p$ and $\mathbf{T} q = \frac{1}{2}\mathbf{T}_1 q + \frac{1}{2}\mathbf{T}_2 q$, and $\mathbf{T}_1 p \neq \mathbf{T} p$. Moreover, $\mathbf{T}_1 p \neq \mathbf{T} p$. Thus, $(\mathbf{T} p, \mathbf{T} q)$ is not an extreme point of $\mathcal{A}$.

We now outline how to modify the construction above when q_c is zero. By setting ϵ_4 to be zero, we obtain $\mathbf{T}_2 q = \mathbf{T} q$ and $\mathbf{T}_2 p = \mathbf{T} p + (-\epsilon_3 p_c)\,(v_a - v_b)$. The desired conclusion follows by choosing ϵ_1 and ϵ_3 small enough such that $\epsilon_1 q_a(\theta_b - \theta_a) = -\epsilon_3 p_c$.

□

11.5 Extreme Points of Joint Range under Privacy Constraints

In the previous section, we considered the extreme points of the joint range under communication constraints. Such communication constraints are routinely applied in practice in the presence of additional constraints such as local differential privacy. However, the results of the previous section do not apply directly, as the joint range is now a strict subset of the set in Theorem 11.1.16, and the extreme points differ significantly. For example, the threshold channels are not even private. However, we show in this section that threshold channels still play a fundamental role. Our main result in this section is the following theorem:

Theorem 11.1.17 (Extreme points of the joint range under privacy and communication constraints)**.** *Let p and q be distributions on $[k]$. Let $\mathcal{C}$ be the set of ϵ-LDP channels from $[k]$ to $[\ell]$. Let $\mathcal{A}$ be the set of all pairs of distributions that are obtained by applying a channel from $\mathcal{C}$ to p and q, i.e.,*

$$\mathcal{A} = \{(\mathbf{T}p, \mathbf{T}q) \mid \mathbf{T} \in \mathcal{C}\}. \tag{11.7}$$

If $(\mathbf{T}p, \mathbf{T}q)$ is an extreme point of $\mathcal{A}$ for $\mathbf{T} \in \mathcal{C}$, then $\mathbf{T}$ can be written as $\mathbf{T} = \mathbf{T}_2 \times \mathbf{T}_1$ for some threshold channel $\mathbf{T}_1 \in \mathcal{T}_{2\ell^2,k}$ and some $\mathbf{T}_2$ an extreme point of the set of ϵ-LDP channels from $[2\ell^2]$ to $[\ell]$.

Actually, our result applies to a broader family of linear programming (LP) channels that we describe below:

Definition 11.5.1 (LP family of channels)**.** *For any $\ell \in \mathbb{N}$, let $\nu = (\nu_1, \nu_2, \ldots, \nu_\ell)$ and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_\ell)$ be two nonnegative vectors in $\mathbb{R}^\ell_+$. For $k \in \mathbb{N}$, define the set of linear programming (LP) channels $\mathcal{J}^{\gamma,\nu}_{\ell,k}$, a subset of $\mathcal{T}_{\ell,k}$, to be the (convex) set of all channels from $[k]$*

to $[\ell]$ that satisfy the following constraints:

$$\textit{For each row } j \in [\ell], \textit{ and for each } i, i' \in [k], \textit{ we have } \mathbf{T}(j,i) \leq \gamma_j \mathbf{T}(j,i') + \nu_j. \tag{11.21}$$

When $\gamma_j = e^\epsilon$ and $\nu_j = 0$ for all $j \in [\ell]$, we recover the set of ϵ-LDP channels from $[k]$ to $[\ell]$. Another example will be mentioned in Section 11.6 for a relaxed version of approximate LDP.

The rest of this section is organized as follows: In Section 11.5.1, we show that any $\mathbf{T}$ that leads to an extreme point of $\mathcal{A}$ cannot have more than $2\ell^2$ unique columns (Lemma 11.5.2). We use this result to prove Theorem 11.1.17 in Section 11.5.2. In Section 11.5.3, we apply Theorem 11.1.17 to prove Corollaries 11.1.18 and 11.1.20.

11.5.1 Bound on the Number of Unique Columns

The following result will be critical in the proof of Theorem 11.1.17, the main result of this section.

Lemma 11.5.2. *Let p and q be distributions on $[k]$. Let $\mathcal{C}$ be the set of channels from $[k]$ to $[\ell]$, from Definition 11.5.1. Let $\mathcal{A}$ be the set of all pairs of distributions that are obtained by applying a channel from $\mathcal{C}$ to p and q, i.e.,*

$$\mathcal{A} = \{(\mathbf{T}p, \mathbf{T}q) \mid \mathbf{T} \in \mathcal{C}\}. \tag{11.22}$$

If $\mathbf{T}$ has more than $2\ell^2$ unique columns, then $(\mathbf{T}p, \mathbf{T}q)$ is not an extreme point of $\mathcal{A}$.

We prove this result in Section 11.5.1.2 after proving a quantitatively weaker, but simpler, result in Section 11.5.1.1.

11.5.1.1 Warm-Up: An Exponential Bound on the Number of Unique Columns

In this section, we first prove a weaker version of Lemma 11.5.2, where we upper-bound the number of unique columns in the extreme points of $\mathcal{C}$ from Definition 11.5.1 (not just those that lead to extreme points of $\mathcal{A}$) by an exponential in ℓ. In fact, this bound will be applicable for a broader class of channels that satisfy the following property:

Condition 11.5.3 (Only one free entry per column)**.** *Let $\mathcal{C}$ be a convex set of channels from $[k]$ to $[\ell]$. Let $\mathbf{T}$ be an extreme point of $\mathcal{C}$. Then there exist numbers $\{m_1, \dots, m_\ell\}$ and $\{M_1, \dots, M_\ell\}$ such that for every column $c \in [k]$, there exists at most a single row $r \in [\ell]$ such that $\mathbf{T}(r, c) \notin \{m_r, M_r\}$. We call such entries* free*.*

We show in Appendix H.2 that extreme points of the LP channels from Definition 11.5.1 satisfy Condition 11.5.3. The following claim bounds the number of unique columns in any extreme point of $\mathcal{C}$, and thus also implies a version of Theorem 11.1.17 with $\ell \cdot 2^{\ell-1}$ instead of $2\ell^2$ (cf. Fact 11.2.1).

Claim 11.5.4 (Number of unique columns in an extreme point is at most exponential in ℓ)**.** *Let $\mathcal{C}$ be a set of channels satisfying the property of Condition 11.5.3. Let $\mathbf{T}$ be an extreme point of $\mathcal{C}$. Then the number of unique columns in $\mathbf{T}$ is at most $\ell \cdot 2^{\ell-1}$. In particular, $\mathbf{T}$ can be written as $\mathbf{T}_2 \times \mathbf{T}_1$, where $\mathbf{T}_1$ is a deterministic map from $[k]$ to $[\ell']$ and $\mathbf{T}_2$ is a map from $[\ell']$ to $[\ell]$, for $\ell' = \ell \cdot 2^{\ell-1}$.*

Proof. We use the notation from Condition 11.5.3. For each column, there are ℓ possible locations of a potential free entry. Let this location be j^*; the value at this location is still flexible. Now let us consider the number of ways to assign values at the remaining locations. For each $j \in [\ell] \setminus \{j^*\}$, the entry is either m_j or M_j (since j is not a free entry). Thus, there are $2^{\ell-1}$ such possible assignments. Since the column entries sum to one, each of those $2^{\ell-1}$ assignments fixes the value at the j^* location, as well. Thus, there are at most $\ell \cdot 2^{\ell-1}$ unique columns in $\mathbf{T}$. □

11.5.1.2 Forbidden Structure in Extreme Points Using the Joint Range

In Claim 11.5.4, we considered the extreme points of LP channels. However, we are actually interested in a (potentially much) smaller set: the extreme points that correspond to the extreme points of the joint range $\mathcal{A}$. In this section, we identify a necessary structural property for extreme points of LP channels that lead to extreme points of the joint range. We begin by defining the notion of a "loose" entry in a channel in $\mathcal{C}$:

Definition 11.5.5 (Loose and tight entries)**.** *Let $\mathbf{T}$ be a channel in $\mathcal{J}_{\ell,k}^{\gamma,\nu}$ from Definition 11.5.1 that maps from $[k]$ to $[\ell]$. Let $\{m_1, \ldots, m_\ell\}$ and $\{M_1, \ldots, M_\ell\}$ be the row-wise minimum and maximum entries, respectively. For $c \in [k]$ and $r \in [\ell]$, we say an entry $\mathbf{T}(r, c)$ is* max-tight *if $\mathbf{T}(r, c) = M_r$ and $M_r = \gamma_r m_r + \nu_r$. An entry $\mathbf{T}(r, c)$ is* min-tight *if $\mathbf{T}(r, c) = m_r$ and $M_r = \gamma_r m_r + \nu_r$. An entry that is neither max-tight nor min-tight is called* loose*.*

Remark 11.5.6. *Our results in this section continue to hold for a slightly more general definition, where we replace the linear functions $\gamma_j x + \nu_j$ by arbitrary monotonically increasing functions $f_j(x)$. We focus on linear functions for simplicity and clarity. (Also see Remark 11.5.11.)*

If the rest of the row is kept fixed, a max-tight entry cannot be increased without violating privacy constraints, but it can be decreased. Similarly, a min-tight entry cannot be decreased without violating privacy constraints, but it can be increased. Loose entries can be either increased or decreased without violating privacy constraints. These perturbations need to be balanced by adjusting other entries in the same column to satisfy column stochasticity; for example, a max-tight entry can be decreased while simultaneously increasing a min-tight or loose entry in the same column. This is formalized below:

Condition 11.5.7 (Mass can be transferred from entries that are not tight)**.** *Let $\mathcal{C}$ be a set of channels from $[k]$ to $[\ell]$. Let $\mathbf{T}$ be any extreme point of $\mathcal{C}$. Suppose there are two rows (r, r') and two columns (c, c') (in the display below, we take $r < r'$ and $c < c'$ without loss of*

generality) with values (m, m', M, M')*, as shown below:*

$$\begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & M & \cdots & m & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & m' & \cdots & M' & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix},$$

such that:

- $\mathbf{T}(r, c)$ *and* $\mathbf{T}(r', c')$ *are not min-tight* (M *and* M' *above, respectively*).
- $\mathbf{T}(r, c')$ *and* $\mathbf{T}(r', c)$ *are not max-tight* (m *and* m' *above, respectively*).

Then there exist $\epsilon' > 0$ *and* $\delta' > 0$ *such that for all* $\epsilon \in [0, \epsilon')$ *and* $\delta \in [0, \delta')$*, the following matrix* $\mathbf{T}'$ *also belongs to* $\mathcal{C}$*:*

$$\mathbf{T}' = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & M - \epsilon & \cdots & m + \delta & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & m' + \epsilon & \cdots & M' - \delta & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix},$$

where the omitted entries of $\mathbf{T}$ *and* $\mathbf{T}'$ *are the same.*

We show that the channels from Definition 11.5.1 satisfy Condition 11.5.7 in Appendix H.2. Using Condition 11.5.7, we show that the following structure is forbidden in the channels that lead to extreme points of the joint range:

Lemma 11.5.8. *Let* p *and* q *be two distributions on* $[k]$*. Let* $\mathcal{C}$ *be the set of LP channels from Definition 11.5.1* (*or, more generally, a convex set of channels satisfying Condition 11.5.7*) *from*

$[k]$ to $[\ell]$. Suppose p_i/q_i is strictly increasing in i. Let $\mathbf{T} \in \mathcal{C}$ have the following structure: there are two rows (r, r') (in the display below, $r < r'$ is taken without loss of generality) and three columns $i_1 < i_2 < i_3$ with values (m, m', m'', M, M', M''), as shown below:

$$\begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & M & \cdots & m & \cdots & M' & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & m' & \cdots & M'' & \cdots & m'' & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix},$$

such that:

- $\mathbf{T}(r, i_1), \mathbf{T}(r, i_3)$, *and* $\mathbf{T}(r', i_2)$ *are not min-tight* (M, M', *and* M'' *above, respectively*).
- $\mathbf{T}(r, i_2), \mathbf{T}(r', i_1)$, *and* $\mathbf{T}(r', i_3)$ *are not max-tight* (m, m', *and* m'' *above, respectively*).

Let $\mathcal{A} := \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{C}\}$. Then $(\mathbf{T}p, \mathbf{T}q)$ cannot be an extreme point of $\mathcal{A}$.

Proof. Firstly, the set $\mathcal{A}$ is convex since $\mathcal{C}$ is convex. For some $\epsilon > 0$ and $\delta > 0$ to be decided later, consider the following perturbed matrices:

$$\mathbf{T}' = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & M - \epsilon & \cdots & m + \delta & \cdots & M' & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & m' + \epsilon & \cdots & M'' - \delta & \cdots & m'' & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix},$$

$$\mathbf{T}'' = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & M & \cdots & m+\epsilon' & \cdots & M'-\delta' & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & m' & \cdots & M''-\epsilon' & \cdots & m''+\delta' & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

To be specific, the entries of $\mathbf{T}'$, $\mathbf{T}''$, and $\mathbf{T}$ match except in the six locations highlighted here. Since $\mathcal{C}$ satisfies Condition 11.5.7 (see Claim H.2.2), both $\mathbf{T}'$ and $\mathbf{T}''$ belong to the set $\mathcal{C}$ if ϵ, ϵ', δ, and δ are small enough and positive. We will now show that there exist choices of these parameters such that $(\mathbf{T}p, \mathbf{T}q)$ is a convex combination of $(\mathbf{T}'p, \mathbf{T}'q)$ and $(\mathbf{T}''p, \mathbf{T}''q)$, and these three points are distinct elements of $\mathcal{A}$. Consequently, $(\mathbf{T}p, \mathbf{T}q)$ will not be an extreme point of $\mathcal{A}$.

For any $j \in \ell$, let v_j denote the vector in $\mathbb{R}^\ell$ that is 1 at the j^{th} location and 0 otherwise. Define $\theta_i := p_i/q_i$ to be the likelihood ratio. If $\theta_i < \infty$, then $p_i = \theta_i q_i$. Since θ_i is strictly increasing in i, only θ_{i_3} may be infinity. Let us first suppose that $\theta_{i_3} < \infty$. We will consider the case when θ_{i_3} might be infinity in the end.

Let us begin by analyzing how $\mathbf{T}'$ transforms p and q. Since $\mathbf{T}'$ differs from $\mathbf{T}$ only in the four locations mentioned above, $\mathbf{T}'p$ and $\mathbf{T}p$, both of which are distributions on $[\ell]$, differ only in the elements r and r' of $[\ell]$. On the element r, $(\mathbf{T}'q)_r - (\mathbf{T}q)_r$ is equal to $-\epsilon q_{i_1} + \delta q_{i_2}$, and equal to its negation on the element r'. In particular, they satisfy the relation

$$\mathbf{T}'q = \mathbf{T}q + (-\epsilon q_{i_1} + \delta q_{i_2})(v_r - v_{r'}).$$

If $\epsilon q_{i_1} = \delta q_{i_2}$, we have $\mathbf{T}'q = \mathbf{T}q$. Under the same setting, p is transformed as follows:

$$\mathbf{T}'p = \mathbf{T}p + (-\epsilon p_{i_1} + \delta p_{i_2})(v_r - v_{r'})$$

$$= \mathbf{T}p + (-\epsilon\theta_{i_1} q_{i_1} + \delta\theta_{i_2} q_{i_2})\,(v_r - v_{r'})$$
$$= \mathbf{T}p + \epsilon q_{i_1}(-\theta_{i_1} + \theta_{i_2})\,(v_r - v_{r'})\,.$$

We now analyze the effect of $\mathbf{T}''$, which satisfies

$$\mathbf{T}''q = \mathbf{T}q + (\epsilon' q_{i_2} - \delta' q_{i_3})\,(v_r - v_{r'})\,.$$

If $\epsilon' q_{i_2} = \delta' q_{i_3}$, we have $\mathbf{T}''q = \mathbf{T}q$. Under the same setting, p is transformed as follows:

$$\mathbf{T}''p = \mathbf{T}p + (\epsilon' p_{i_2} - \delta' p_{i_3})\,(v_r - v_{r'})$$
$$= \mathbf{T}p + (-\epsilon'\theta_{i_2} q_{i_2} + \delta'\theta_{i_3} q_{i_3})\,(v_r - v_{r'})$$
$$= \mathbf{T}p + \epsilon' q_{i_2}(-\theta_{i_2} + \theta_{i_3})\,(v_r - v_{r'})\,.$$

Now observe that $\theta_{i_1} < \theta_{i_2} < \theta_{i_3}$. By choosing $\epsilon > 0$ and $\epsilon' > 0$ small enough such that $\epsilon q_{i_1}(-\theta_{i_1} + \theta_{i_2}) = \epsilon' q_{i_2}(-\theta_{i_2} + \theta_{i_3})$, we obtain

$$(\mathbf{T}p, \mathbf{T}q) = \frac{1}{2} \cdot (\mathbf{T}'p, \mathbf{T}'q) + \frac{1}{2} \cdot (\mathbf{T}'p, \mathbf{T}'q)\,,$$

and all three points are distinct elements of $\mathcal{A}$. Such a choice of ϵ and ϵ' always exists, since both $q_{i_1}(-\theta_{i_1} + \theta_{i_2})$ and $q_{i_2}(-\theta_{i_2} + \theta_{i_3})$ are positive and finite. Thus, $(\mathbf{T}p, \mathbf{T}q)$ is not an extreme point of $\mathcal{A}$.

Let us now consider the case when $\theta_{i_3} = \infty$, or equivalently, $q_{i_3} = 0$. Define ϵ' to be 0, so that $\mathbf{T}''q = \mathbf{T}q$ and $\mathbf{T}''p = \mathbf{T}p - \delta' p_{i_3}\,(v_r - v_{r'})$. Then choose $\delta' > 0$ and $\epsilon > 0$ small enough such that $\epsilon q_{i_1}\,(\theta_{i_2} - \theta_{i_1}) = \delta' p_{i_3}$, which is possible since both sides are positive and finite. Thus, $(\mathbf{T}p, \mathbf{T}q)$ is a non-trivial convex combination of $(\mathbf{T}'p, \mathbf{T}'q)$ and $(\mathbf{T}''p, \mathbf{T}''q)$, so is not an extreme point of $\mathcal{A}$. □

11.5.1.3 Proof of Lemma 11.5.2

Proof. Without loss of generality, we assume that the likelihood ratios p_i/q_i are unique and strictly increasing in i. We refer the reader to the proof of Theorem 11.1.16 and Claim H.4.1 for more details.

Let $\mathbf{T} \in \mathcal{C}$ be a channel from $[k]$ to $[\ell]$ such that $(\mathbf{T}p, \mathbf{T}q)$ is an extreme point of $\mathcal{A}$. Suppose that there are ℓ' unique columns in $\mathbf{T}$ with $\ell' > 2\ell^2$. From now on, we assume that $\ell' = 2\ell^2$; otherwise, we apply the following argument to the first $2\ell^2$ distinct columns.

Let $c, c' \in [k]$ be such that the c^{th} and c'^{th} columns of $\mathbf{T}$ are distinct. Observe that for every pair of distinct columns c and c', there are two rows such that c^{th} column has a strictly bigger value than the c'^{th} column on one row, and vice versa on the another row. This is because both of the columns sum up to 1, so if a particular column has a larger entry in a row, its entry must be smaller in a different row. In particular, there exist two rows $g(c, c')$ and $h(c, c')$ such that $\mathbf{T}(g(c, c'), c) > \mathbf{T}(g(c, c'), c')$ and $\mathbf{T}(h(c, c'), c) < \mathbf{T}(h(c, c'), c')$. As a result, $\mathbf{T}(g(c, c'), c)$ and $\mathbf{T}(h(c, c'), c')$ are not min-tight, and $\mathbf{T}(g(c, c'), c')$ and $\mathbf{T}(h(c, c'), c)$ are not max-tight.

We now order the distinct columns of $\mathbf{T}$ in the order of their appearance from left to right. Let $i_1, i_2, \ldots, i_{\ell'}$ be the indices of the unique columns. For example, the first distinct column i_1 is the first column of $\mathbf{T}$ (corresponding to the element 1). The second distinct column i_2 is the first column of $\mathbf{T}$ that is different from the first column. The third distinct column is the first column of $\mathbf{T}$ that is different from the first two columns. Let $\mathcal{I}$ be the set of unique column indices of $\mathbf{T}$.

Now, we divide the distinct columns in $\mathbf{T}$ into pairs: $\mathcal{H} = \{(i_1, i_2), (i_3, i_4), \ldots, (i_{\ell'-1}, i_{\ell'})\}$. The total number of possible choices in $\mathcal{H}$ is $\ell'/2$, and for every $(m, m+1)$ in $\mathcal{H}$, the possible number of choices of $(g(i_m, i_{m+1}), h(i_m, i_{m+1}))$ is at most $\ell(\ell - 1)$, since both of these lie in $[\ell]$ and are distinct. Thus, there must exist two pairs in $\mathcal{H}$ whose

corresponding indices are the same, since $\frac{\ell'}{2} = \ell^2 > \ell(\ell-1)$.

Without loss of generality, we let these pairs of columns be (i_1, i_2) and (i_3, i_4). Let $r := g(i_1, i_2) = g(i_3, i_4)$ and $r' := h(i_1, i_2) = h(i_3, i_4)$. Then the previous discussion implies that:

- $\mathbf{T}(r, i_1)$ and $\mathbf{T}(r, i_3)$ are not min-tight, and $\mathbf{T}(r', i_1)$ and $\mathbf{T}(r', i_3)$ are not max-tight.
- $\mathbf{T}(r, i_2)$ and $\mathbf{T}(r, i_4)$ are not max-tight, and $\mathbf{T}(r', i_2)$ and $\mathbf{T}(r', i_4)$ are not min-tight.

Thus, the columns i_1, i_2, and i_3 satisfy the conditions of Lemma 11.5.8, i.e., they exhibit the forbidden structure. This implies that $(\mathbf{T}p, \mathbf{T}q)$ cannot be an extreme point of $\mathcal{A}$. Therefore, $\ell' \leq 2\ell^2$. □

11.5.2 Proof of Theorem 11.1.17: Unique Columns to Threshold Channels

In this section, we provide the proof of Theorem 11.1.17 using Lemma 11.5.2. Noting that our main structural result is more widely applicable (Condition 11.5.7), we prove a more general version of Theorem 11.1.17 below for Definition 11.5.1. Before doing so, we require an additional property on the set of our channels, proved in Appendix H.2:

Claim 11.5.9 (Closure under pre-processing)**.** *The set $\mathcal{J}_{\ell,k}^{\gamma,\nu}$ satisfies the following closure property under pre-processing:*

$$\mathcal{J}_{\ell,k}^{\gamma,\nu} = \bigcup_{\ell'=1}^{k} \left\{ \mathbf{T}_2 \times \mathbf{T}_1 : \mathbf{T}_2 \in \mathcal{J}_{\ell,\ell'}^{\gamma,\nu} \text{ and } \mathbf{T}_1 \in \mathcal{T}_{\ell',k} \right\}. \tag{11.23}$$

Informally, if we take an arbitrary channel $\mathbf{T}_1$ and compose it with an LP private channel $\mathbf{T}_2$, the composition $\mathbf{T}_2 \times \mathbf{T}_1$ is also LP private.

The following result is thus a more general version of Theorem 11.1.17:

Theorem 11.5.10 (Structure of optimal channels)**.** *Let p and q be distributions on $[k]$. For any $\ell \in \mathbb{N}$, let $\mathcal{C}$ be the set of channels $\mathcal{J}_{\ell,k}^{\gamma,\nu}$ from Definition 11.5.1. Let $\mathcal{A}$ be the set of all pairs of distributions that are obtained by applying a channel from $\mathcal{C}$ to p and q, i.e.,*

$$\mathcal{A} = \{(\mathbf{T}p, \mathbf{T}q) \mid \mathbf{T} \in \mathcal{C}\}. \tag{11.24}$$

If $(\mathbf{T}p, \mathbf{T}q)$ is an extreme point of $\mathcal{A}$, then $\mathbf{T}$ can be written as $\mathbf{T} = \mathbf{T}_2 \times \mathbf{T}_1$, for some $\mathbf{T}_1 \in \mathcal{T}_{\ell',k}^{\texttt{thresh}}$ and $\mathbf{T}_2$ an extreme point of the set $\mathcal{J}_{\ell,\ell'}^{\gamma,\nu}$.

Proof. Since $\mathcal{C}$ is convex, the joint range $\mathcal{A}$ is convex. By Lemma 11.5.2, we know that if $(\mathbf{T}p, \mathbf{T}q)$ is an extreme point of $\mathcal{A}$, then $\mathbf{T}$ can be written as $\mathbf{T}_2 \times \mathbf{T}_1$, where $\mathbf{T}_1 \in \mathcal{T}_{\ell',k}$ for $\ell' := 2\ell^2$. Using Claim 11.5.9, any such channel in $\mathcal{C}$ is of the form $\mathbf{T}_2 \times \mathbf{T}_1$, where $\mathbf{T}_1 \in \mathcal{T}_{\ell',k}$ and $\mathbf{T}_2 \in \mathcal{J}_{\ell,\ell'}^{\gamma,\nu}$. Combining the last two observations, we obtain the following:

$$\mathcal{A} = \operatorname{conv}\left(\left\{(\mathbf{T}_2 \times \mathbf{T}_1 p, \mathbf{T}_2 \times \mathbf{T}_1 q) : \mathbf{T}_2 \in \mathcal{J}_{\ell,\ell'}^{\gamma,\nu}, \mathbf{T}_1 \in \mathcal{T}_{\ell',k}\right\}\right). \tag{11.25}$$

We now claim that we can further take $\mathbf{T}_1$ to be a threshold channel $\mathbf{T}_1 \in \mathcal{T}_{\ell',k}^{\texttt{thresh}}$ and $\mathbf{T}_2$ to be an extreme point of $\mathcal{J}_{\ell,\ell'}^{\gamma,\nu}$. This claim follows if we can write an arbitrary point in $\mathcal{A}$ as a convex combination of elements of the set $\left\{(\mathbf{T}_2 \times \mathbf{T}_1 p, \mathbf{T}_2 \times \mathbf{T}_1 q) : \mathbf{T}_2 \in \operatorname{ext}\left(\mathcal{J}_{\ell,\ell'}^{\gamma,\nu}\right), \mathbf{T}_1 \in \mathcal{T}_{\ell',k}^{\texttt{thresh}}\right\}$. By equation Equation (11.25), it suffices to demonstrate this convex combination for all points of the form $(\mathbf{T}_2 \times \mathbf{T}_1 p, \mathbf{T}_2 \times \mathbf{T}_1 q)$, for some $\mathbf{T}_2 \in \mathcal{J}_{\ell,\ell'}^{\gamma,\nu}$ and $\mathbf{T}_1 \in \mathcal{T}_{\ell',k}$.

Let $\mathbf{H}_1, \mathbf{H}_2, \ldots$ be extreme points of $\mathcal{J}_{\ell,\ell'}^{\gamma,\nu}$, and let $\mathbf{L}_1, \mathbf{L}_2, \ldots$ be an enumeration of the threshold channels $\mathcal{T}_{\ell',k}^{\texttt{thresh}}$. By definition, any $\mathbf{T}_1 \in \mathcal{J}_{\ell,\ell'}^{\gamma,\nu}$ can be written as $\sum_i \alpha_i \mathbf{H}_i$ for some convex combination $\alpha_1, \alpha_2, \ldots$. Furthermore, Theorem 11.1.16 implies that any $(\mathbf{T}_2 p, \mathbf{T}_2 q)$, for $\mathbf{T}_2 \in \mathcal{T}_{\ell',k}$, can be written as $\sum_j \beta_j (\mathbf{L}_j p, \mathbf{L}_j q) = (\sum_j \beta_j \mathbf{L}_j p, \sum_j \beta_j \mathbf{L}_j q)$, for some convex combination $\beta_1, \beta_2, \ldots$.

Thus, any arbitrary point $(\mathbf{T}_2 \times \mathbf{T}_1 p, \mathbf{T}_2 \times \mathbf{T}_1 q)$, for $\mathbf{T}_2 \in \mathcal{J}_{\ell,\ell'}^{\gamma,\nu}$ and $\mathbf{T}_1 \in \mathcal{T}_{\ell',k}$, can be

written as

$$\begin{aligned}
(\mathbf{T}_2 \times \mathbf{T}_1 p, \mathbf{T}_2 \times \mathbf{T}_1 q) &= \left(\left(\sum_i \alpha_i \mathbf{H}_i \right) \times \mathbf{T}_1 p, \left(\sum_i \alpha_i \mathbf{H}_i \right) \times \mathbf{T}_1 q \right) \\
&= \sum_i \alpha_i \left(\mathbf{H}_i \times \mathbf{T}_1 p, \mathbf{H}_i \times \mathbf{T}_1 q \right) \\
&= \sum_i \alpha_i \left(\mathbf{H}_i \left(\mathbf{T}_1 p \right), \mathbf{H}_i \left(\mathbf{T}_1 q \right) \right) \\
&= \sum_i \alpha_i \left(\mathbf{H}_i \left(\sum_j \beta_j \mathbf{L}_j p \right), \mathbf{H}_i \left(\sum_j \beta_j \mathbf{L}_j q \right) \right) \\
&= \sum_i \alpha_i \left(\sum_j \beta_j \mathbf{H}_i \times \mathbf{L}_j p, \sum_j \beta_j \mathbf{H}_i \times \mathbf{L}_j q \right) \\
&= \sum_i \sum_j \alpha_i \beta_j \left(\mathbf{H}_i \times \mathbf{L}_j p, \mathbf{H}_i \times \mathbf{L}_j q \right).
\end{aligned}$$

Finally, note that $\{\alpha_i \beta_j\}$ are also valid convex combinations of $(\mathbf{H}_i \times \mathbf{L}_j p, \mathbf{H}_i \times \mathbf{L}_j q)$, since they are nonnegative and sum to 1. □

Remark 11.5.11 (Extending Theorems 11.1.17 and 11.5.10 to a more general set of constraints)**.** *We note that Theorem 11.5.10 can be extended to an arbitrary convex set of channels $\mathcal{C}$ that satisfy* (*appropriately modified versions of*) *Condition 11.5.7 and equation Equation* (11.23)*.* (*Also see Remark 11.5.6.*)

11.5.3 Application to Hypothesis Testing

In Section 11.3, we showed that the minimax-optimal sample complexity can be obtained by a communication-efficient and efficiently computable channel, up to logarithmic factors. However, for a particular (p, q), these guarantees can be significantly improved. For example, consider the extreme case when p and q are the following two distributions on $[k]$: for γ small enough,

$$p = [\alpha, 1 - \alpha - (k-2)\gamma, \gamma, \gamma, \ldots, \gamma],$$

$$q = [\beta, 1 - \beta - (k-2)\gamma, \gamma, \gamma, \ldots, \gamma].$$

Let $\mathbf{T}'$ be a deterministic binary channel that maps the first and second elements to different elements, while assigning the remaining elements arbitrarily. Now consider the following private channel $\mathbf{T}$: the channel $\mathbf{T}'$, followed by the randomized response over binary distributions. Then as $\gamma \to 0$, the performance of $\mathbf{T}$ mirrors equation Equation (11.3), which is much better than the minimax bound of equation Equation (11.4). Thus, there is a wide gap between instance-optimal and minimax-optimal performance. We thus consider the computational question of optimizing a quasi-convex function $g(\mathbf{T}p, \mathbf{T}q)$ over all possible ϵ-private channels that map to a domain of size ℓ.

The following result proves Corollary 11.1.18 for $\mathcal{C}$ equal to $\mathcal{P}^{\epsilon}_{\ell,k}$:

Corollary 11.5.12 (Computationally efficient algorithms for maximizing quasi-convex functions under privacy constraints)**.** *Let p and q be fixed distributions over $[k]$, let $\mathcal{C}$ be the set of channels $\mathcal{J}^{\gamma,\nu}_{\ell,k}$ from Definition 11.5.1, and let $\mathcal{A} = \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{C}\}$. Let $g : \mathcal{A} \to \mathbb{R}$ be a jointly quasi-convex function. Then there is an algorithm that solves $\max_{\mathbf{T}\in\mathcal{C}} g(\mathbf{T}p, \mathbf{T}q)$ in time polynomial in k^{ℓ^2} and $2^{O(\ell^3 \log \ell)}$.*[33]

Proof. The algorithm is as follows: we try all threshold channels in $\mathbf{T}_1 \in \mathcal{T}^{\texttt{thresh}}_{\ell',k}$ and all extreme points of $\mathcal{J}^{\gamma,\nu}_{\ell,\ell'}$, and output the channel $\mathbf{T} = \mathbf{T}_2 \times \mathbf{T}_1$ that attains the maximum value of $g(\mathbf{T}p, \mathbf{T}q)$. By Theorem 11.5.10 and quasi-convexity of g, we know the algorithm will output a correct value, since all extreme points are of this form. Thus, we focus on bounding the runtime. We know that the cardinality of $\mathcal{T}^{\texttt{thresh}}_{\ell',k}$ is bounded by $k^{\ell'}$ (up to a rotation of output rows). By Fact 11.2.3, the time taken to iterate through all the extreme points of ϵ-LDP channels from $[\ell']$ to $[\ell]$ is at most polynomial in $2^{\ell^3 \log \ell}$, since $\mathcal{J}_{\ell,\ell'}$ is a polytope in $\mathbb{R}^{2\ell^3}$ with $\mathrm{poly}(\ell)$ inequalities. This completes the proof. □

[33] Recall that g is assumed to be permutation invariant. If not, an extra factor of $\ell!$ will appear in the time complexity.

The proof of Corollary 11.1.20 is immediate from Fact 11.2.7, Corollary 11.1.18, and Proposition 11.6.2, stated later.

11.6 Extensions to Other Notions of Privacy

In this section, we explore computational and statistical aspects of hypothesis testing under other notions of privacy. Section 11.6.1 is on approximate privacy, in which we first focus on (ϵ, δ)-LDP and then our proposed definition of approximate privacy. Next, we focus on binary communication constraints for Rényi differential privacy in Section 11.6.2. This will be possible since our algorithmic and structural results were not restricted to the case of pure LDP.

We begin by noting that communication constraints have a benign effect on the sample complexity of hypothesis testing for many notions of privacy:

Condition 11.6.1 (Closure under post-processing)**.** *Let $k \in \mathbb{N}$. For each $r \in \mathbb{N}$, consider sets $\mathcal{C}_r \subseteq \mathcal{T}_{r,k}$ and define $\mathcal{C} = \cup_{r \in \mathbb{N}} \mathcal{C}_r$. We say $\mathcal{C}$* satisfies *ℓ-post-processing if for every $r \in \mathbb{N}$, if $\mathbf{T} \in \mathcal{C}_r$ and $\mathbf{H}$ is a deterministic channel from $[r]$ to $[\ell]$, the channel $\mathbf{H} \times \mathbf{T}$ also belongs to $\mathcal{C}_\ell$, and thus to $\mathcal{C}$.*

Post-processing is satisfied by various notions of privacy: ϵ-pure privacy, (ϵ, δ)-approximate privacy (see Dwork and Roth [DR13, Proposition 2.1]), and Rényi privacy [Mir17]. For a set of channels $\mathcal{C}$, we use the notation $n^*(p, q, \mathcal{C})$ to denote the sample complexity of hypothesis testing under channel constraints of $\mathcal{C}$ in Definition 11.1.2. The following result shows that even with binary communication constraints, the sample complexity increases by at most a logarithmic factor:

Proposition 11.6.2 (Benign effect of communication constraints on sample complexity under closure)**.** *Let p and q be any two distributions on $[k]$. Let $\mathcal{C}$ be a set of channels that satisfy ℓ-post-processing (Condition 11.6.1) for some $\ell > 1$. Let $\mathcal{C}_\ell$ denote the subset of channels*

in $\mathcal{C}$ that map to a domain of size ℓ. Then

$$n^*(p, q, \mathcal{C}_\ell) \lesssim n^*(p, q, \mathcal{C}) \cdot \left(1 + \frac{\log\left(n^*(p, q, \mathcal{C})\right)}{\ell}\right). \tag{11.26}$$

Proof. Let $\mathbf{T}$ be the optimal channel in $\mathcal{C}$ that maximizes $d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)$. Let k' be the size of the range of $\mathbf{T}$. By Fact 11.2.7, we have $n^*(p, q, \mathcal{C}) \asymp 1/d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)$. By Fact 11.2.8, we know that there exists $\mathbf{T}' \in \mathcal{T}_{\ell,k'}$ such that[34]

$$d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \lesssim d_{\mathrm{h}}^2(\mathbf{T}'(\mathbf{T}p), \mathbf{T}'(\mathbf{T}q)) \cdot \left(1 + \frac{\log(1/d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q))}{\ell}\right). \tag{11.27}$$

By the assumed closure of $\mathcal{C}$ under post-processing, the channel $\mathbf{T}' \times \mathbf{T}$ belongs to $\mathcal{C}$. Thus, the channel $\mathbf{T}' \times \mathbf{T}$ also belongs to $\mathcal{C}_\ell$, since its output is of size ℓ. This implies that the sample complexity $n^*(p, q, \mathcal{C}_\ell)$ is at most $1/d_{\mathrm{h}}^2(\mathbf{T}' \times \mathbf{T}p, \mathbf{T}' \times \mathbf{T}q)$. Using the fact that $n^*(p, q, \mathcal{C}) \asymp 1/d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)$, we obtain the desired result. □

Thus, in the rest of this section, our main focus will be on the setting of binary channels.

11.6.1 Approximate Local Privacy

In this section, we first focus on (ϵ, δ)-approximate LDP (Definition 11.6.3). We begin by showing upper bounds on the associated sample complexity. On the computational front, we present efficient algorithms for the case of binary constraints and then propose a relaxation for the case of larger output domains.

We first recall the definition of (ϵ, δ)-LDP:

[34] If the supremum is not attained, the proof can be modified by considering a suitable sequence of channels and applying a similar argument.

Definition 11.6.3 ((ϵ, δ)-LDP)**.** *We say a channel from $\mathcal{X}$ to $\mathcal{Y}$ is* (ϵ, δ)-LDP *if for all $S \subseteq \mathcal{Y}$, we have*

$$\sup_{x,x' \in \mathcal{X}} \mathbb{P}[\mathbf{T}(x) \in S)] - e^{\epsilon} \cdot \mathbb{P}[\mathbf{T}(x) \in S)] - \delta \leq 0. \tag{11.28}$$

What makes the analysis of (ϵ, δ)-LDP different from ϵ-LDP is that when $|\mathcal{Y}| > 2$, the condition in inequality (11.28) should be verified for *all* sets $S \subseteq \mathcal{Y}$, not just singleton sets ($|S| = 1$). Only when $|\mathcal{Y}| = 2$ is it enough to consider singleton sets S.

Let $n^*(p, q, (\epsilon, \delta))$ denote the sample complexity for the setting in Definition 11.1.3, with $\mathcal{C}$ equal to the set of all (ϵ, δ)-LDP channels. We directly obtain the following upper bound on the sample complexity, proved in Appendix H.3, which happens to be tight for the case of binary distributions:

Claim 11.6.4 (Sample complexity of approximate LDP)**.** *For all $\delta \in (0, 1)$, we have*

$$n^*(p, q, (\epsilon, \delta)) \lesssim \min\left(n^*(p, q, \epsilon) \cdot \frac{1}{1-\delta}\ , n^*(p, q) \cdot \frac{1}{\delta}\right).$$

Moreover, this is tight (up to constant factors) when both p and q are binary distributions.

In the rest of this section, we focus on efficient algorithms in the presence of both privacy and communication constraints.

Turning to computationally efficient algorithms for the case of privacy and communication constraints, we present two kinds of results: exact results for the case of binary outputs, and sharp relaxations for the case of multiple outputs.

Binary channels: Let $\mathcal{C}$ be the set of all (ϵ, δ)-approximate LDP channels from $[k]$ to $[2]$, i.e., binary channels. Let $\gamma = (e^{\epsilon}, e^{\epsilon})$ and $\nu = (\delta, \delta)$. Observe that $\mathcal{C}$ is then equal to $\mathcal{J}_{2,k}^{\gamma,\nu}$, defined in Definition 11.5.1. Thus, Theorem 11.5.10 and Corollary 11.5.12 hold in this case, as well.

Channels with larger output spaces: Here, we define a new notion of privacy that relaxes (ϵ, δ)-LDP. It is enough to verify whether the privacy condition holds for singleton events S:

Definition 11.6.5 ((ϵ, δ)-SLDP)**.** *We say a channel $\mathcal{X}$ to $\mathcal{Y}$ is* (ϵ, δ)-singleton-based-LDP ((ϵ, δ)-SLDP) *if for all $S \subseteq \mathcal{Y}$, we have*

$$\sup_{x,x' \in \mathcal{X}} \mathbb{P}[\mathbf{T}(x) \in S)] - e^{\epsilon} \cdot \mathbb{P}[\mathbf{T}(x) \in S)] - \delta \cdot |S| \leq 0.$$

The following result shows that (ϵ, δ)-SLDP is a good approximation to (ϵ, δ)-LDP when the output space is small:

Claim 11.6.6 (Relations between LDP and SLDP)**.** *Consider a channel $\mathbf{T}$ from $\mathcal{X}$ to $[\ell]$.*

1. *If $\mathbf{T}$ is (ϵ, δ)-SLDP, it is $(\epsilon, \ell\delta)$-LDP.*

2. *If $\mathbf{T}$ is (ϵ, δ)-LDP, it is (ϵ, δ)-SLDP.*

The proof is immediate from the definitions of (ϵ, δ)-LDP and (ϵ, δ)-SLDP, and we omit it. We now show that it is easy to optimize over SLDP channels in the presence of communication constraints. For any $\ell \in \mathbb{N}$, let $\mathcal{C}$ be the set of all channels from $[k]$ to $[\ell]$ that satisfy (ϵ, δ)-SLDP. Let $\gamma = (e^{\epsilon}, e^{\epsilon}, \ldots, e^{\epsilon})$ and $\nu = (\delta, \delta, \ldots, \delta)$. Observe that $\mathcal{C}$ is then equal to $\mathcal{J}_{\ell,k}^{\gamma,\nu}$, defined in Definition 11.5.1. Thus, Theorem 11.5.10 and Corollary 11.5.12 imply that we can efficiently optimize over SLDP channels.

11.6.2 Other Notions of Privacy

We briefly note that our computationally efficient algorithms hold for a wider family of channels defined in Definition 11.5.1; see also Remark 11.5.11.

Finally, we consider the case of Rényi differential privacy introduced in Mironov [Mir17]:

Definition 11.6.7 ((ϵ, α)-Rényi differential privacy)**.** *Let $\epsilon \in \mathbb{R}_+$ and $\alpha > 1$, and let $\mathcal{X}$ and $\mathcal{Y}$ be two domains. A channel $\mathbf{T} : \mathcal{X} \to \mathcal{Y}$ satisfies* (ϵ, α)-RDP *if for all $x, x' \in \mathcal{X}$, we have*

$$D_\alpha(\mathbf{T}(x) \| \mathbf{T}(x')) \leq \epsilon,$$

where $D_\alpha(p\|q)$ is the Rényi divergence of order α between two distributions p and q on the same probability space, defined as

$$D_\alpha(p\|q) := \frac{1}{\alpha - 1} \log \mathop{\mathbb{E}}_{X \sim q} \left[\left(\frac{p(X)}{q(X)} \right)^\alpha \right].$$

Rényi divergence is also defined for $\alpha = 1$ and $\alpha = \infty$ by taking limits. When $\alpha = 1$, the limit yields the Kullback–Leibler divergence, and when $\alpha = \infty$, it leads to the supremum of the log-likelihood ratio between p and q. In fact, (∞, ϵ)-RDP is identical to ϵ-LDP. Similarly, $(1, \epsilon)$-RDP is closely related to mutual information-based privacy [CY16], since the corresponding channel $\mathbf{T}$ has Shannon capacity at most ϵ.

Proposition 11.6.8 (Rényi differential privacy and binary constraints)**.** *Let $\epsilon > 0$ and $\alpha > 1$. Let $\mathcal{C}$ be the set of (ϵ, α)-RDP channels from $[k]$ to $[2]$. Let p and q be two distributions on $[k]$, and define $\mathcal{A} := \{(\mathbf{T}p, \mathbf{T}q) : \mathbf{T} \in \mathcal{C}\}$. If $(\mathbf{T}p, \mathbf{T}q)$ is an extreme point of $\mathcal{A}$ for $\mathbf{T} \in \mathcal{C}$, then $\mathbf{T}$ can be written as $\mathbf{T}_1 \times \mathbf{T}_2$, where $\mathbf{T}_1$ is an extreme point of the set of (ϵ, α)-RDP channels from $[2]$ to $[2]$, and $\mathbf{T}_2$ is a binary threshold channel from $[k]$.*

Proof. Consider two binary distributions $[x, 1-x]$ and $[y, 1-y]$, where $0 \leq x, y \leq 1$. The α-Rényi divergence between the distributions is given by

$$D_\alpha(x\|y) := \frac{1}{\alpha - 1} \log \left(x^\alpha y^{1-\alpha} + (1-x)^\alpha (1-y)^{1-\alpha} \right).$$

Observe that the term inside the logarithm is convex in y for fixed x, and is minimized when $y = x$. Hence, the Rényi divergence above, as a function of y, is decreasing for

$y \in [0, x]$ and increasing for $y \in [x, 1]$. A similar conclusion holds for fixed y and varying x.

Consider a channel $\mathbf{T} \in \mathcal{C}$ given by

$$\mathbf{T} = \begin{bmatrix} x_1 & x_2 & \dots & x_k \\ 1 - x_1 & 1 - x_2 & \dots & 1 - x_k \end{bmatrix}.$$

Without loss of generality, assume $x_1 \leq x_2 \leq \cdots \leq x_k$ Suppose there is an index j such that $x_1 < x_j < x_k$. Observe that $x_j \notin \{0, 1\}$. By the monotonicity property of the Rényi divergence noted above, for any index i, we have

$$\max\{D_\alpha(x_j \| x_i), D_\alpha(x_i \| x_j)\} < \max\{D_\alpha(x_1 \| x_k), D_\alpha(x_k \| x_1)\} \leq \epsilon.$$

This means that x_j can be perturbed up and down by a small enough δ such that the Rényi divergence constraints continue to be satisfied. Such perturbations will allow $\mathbf{T}$ to be written as a convex combination of two distinct matrices, so $\mathbf{T}$ cannot be an extreme point of the (convex) set $\mathcal{C}$. Thus, an extreme point must have only two distinct columns; i.e., it must have the form

$$\mathbf{T} = \begin{bmatrix} x_1 & x_1 & \dots & x_1 & x_k & x_k & \dots & x_k \\ 1 - x_1 & 1 - x_1 & \dots & 1 - x_1 & 1 - x_k & 1 - x_k & \dots & 1 - x_k \end{bmatrix}.$$

Equivalently, any extreme point is a deterministic channel from $[k] \to [2]$ followed by an RDP-channel from $[2] \to [2]$. Since we are only concerned with extreme points that correspond to extreme points of the joint range $\mathcal{A}$, an argument identical to the one in the proof of Theorem 11.5.10 yields that an extreme point must admit a decomposition $\mathbf{T}_1 \times \mathbf{T}_2$, where $\mathbf{T}_2$ is a threshold channel from $[k] \to [2]$ and $\mathbf{T}_1$ is an extreme point of the set of RDP channels from $[2] \to [2]$. □

The above result implies that given a quasi-convex function $g : \mathcal{A} \to \mathbb{R}$, if we are interested in maximizing $g(\mathbf{T}p, \mathbf{T}q)$ over $\mathbf{T} \in \mathcal{C}$, the optimal $\mathbf{T}$ can be written as $\mathbf{T}_1 \times \mathbf{T}_2$, where $\mathbf{T}_1$ is a binary-input, binary-output Rényi private channel and $\mathbf{T}_2$ is a threshold channel. Since there are only $2k$ threshold channels, we can try all those choices of $\mathbf{T}_2$, and then try to optimize over $\mathbf{T}_1$ for each of those choices. However, each such problem is over binary inputs and binary outputs, and thus is amenable to grid search.

Remark 11.6.9. *In addition to the convexity of RDP channels, we also used the closure-under-pre-processing property (see Claim 11.5.9) and the unimodality of $D_\alpha(x\|y)$ when one of the variables is fixed and the other is varied. The above proof technique will therefore work for any set of convex channels from $[k] \to [2]$ that are closed under pre-processing, and are defined in terms of such a unimodal function. In particular, our results will continue to hold for all f-divergence-based private channels, defined as all $\mathbf{T}$ satisfying*

$$D_f(\mathbf{T}(x)\|\mathbf{T}(x')) \leq \epsilon.$$

Our results also hold for zero-concentrated differential privacy (z-CDP) [BS16], which is a notion of privacy defined using Rényi divergences.

11.7 Conclusion

In this paper, we considered the sample complexity of simple binary hypothesis testing under privacy and communication constraints. We considered two families of problems: finding minimax-optimal bounds and algorithms, and finding instance-optimal bounds and algorithms.

For minimax optimality, we considered the set of distributions with fixed Hellinger divergences and total variation distances. This is a natural family to consider, because these two metrics characterize the sample complexity in the low- and high-privacy

regimes. Prior work did not resolve the question of sample complexity in the moderate-privacy regime; our work has addressed this gap in the literature, by establishing a sample-complexity lower bound via a carefully constructed family of distribution pairs on the ternary alphabet. Our results highlight a curious separation between the binary and ternary (and larger alphabet) settings, roughly implying that the binary case is substantially easier (i.e., has a lower sample complexity) than the general case.

Our focus on instance optimality sets our paper apart from most prior work on information-constrained estimation, which exclusively considered minimax optimality. When only privacy constraints are imposed, we established approximately instance-optimal algorithms; i.e., for any distribution pair, we proposed a protocol whose sample complexity is within logarithmic factors of the true sample complexity. Importantly, the algorithm we proposed to identify this protocol is computationally efficient, taking time polynomial in k, the support size of the distributions. When both privacy and communication constraints are in force, we developed instance-optimal algorithms, i.e., protocols whose sample complexity is within constant factors of the true sample complexity. As before, these algorithms take time polynomial in k, for any constant communication constraint of size ℓ.

Our results highlight the critical role played by threshold channels in both communication- and privacy-constrained settings. We showed that for any distribution pair, the channel with output size ℓ that maximizes the output divergence (Hellinger, Kullback–Leibler, or any quasi-convex function in general) among all channels with fixed output size ℓ must be a threshold channel. Furthermore, optimal private channels with output size ℓ admit a decomposition into a threshold channel cascaded with a private channel. These two results underpin our algorithmic contributions.

There are many interesting open problems stemming from our work that would be worth exploring. We did not characterize instance-optimal sample complexity in

the moderate-privacy regime; our work shows that it is not characterized in terms of the Hellinger divergence and total variation distance, but leaves open the possibility of some other divergence, such as the E_γ divergence, capturing the sample complexity. We identified a forbidden structure for optimal private channels; however, the best algorithm from Kairouz, Oh, and Viswanath [KOV16] does not use this information at all. It would be interesting to see if that algorithm could be made more efficient by incorporating the extra structural information. Many open questions remain for the approximate LDP setting, as well. There is no known upper bound on the number of outputs that suffice for optimal approximate LDP channels. It is plausible, but unknown, if instance-optimal private channels with $\ell > 2$ outputs admit decompositions into threshold channels cascaded with private channels, similar to the pure LDP setting. It would be interesting to see if optimal SLDP channels, which are efficient to find, are nearly instance optimal for approximate LDP.

[AAR99] G. E. Andrews, R. Askey, and R. Roy. *Special Functions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1999.

[AB99] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. First. Cambridge University Press, Nov. 1999.

[ABC04] C. Abraham, G. Biau, and B. Cadre. "On the asymptotic properties of a simple estimate of the mode". In: *ESAIM: Probability and statistics* 8 (2004), pp. 1–11.

[ABHHRT72] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. *Robust Estimates of Location: Survey and Advances*. Princeton, NJ, USA: Princeton University Press, 1972.

[AC86] R. Ahlswede and I. Csiszár. "Hypothesis testing with communication constraints". In: *IEEE Transactions on Information Theory* 32.4 (1986), pp. 533–542.

[ACFST21] J. Acharya, C. L. Canonne, C. Freitag, Z. Sun, and H. Tyagi. "Inference under information constraints III: Local privacy constraints". In: *IEEE Journal on Selected Areas in Information Theory* 2.1 (2021), pp. 253–267.

[Ach03] D. Achlioptas. "Database-friendly random projections: Johnson-Lindenstrauss with binary coins". In: *Journal of computer and System Sciences* 66.4 (2003), pp. 671–687.

[ACLST22] J. Acharya, C. L. Canonne, Y. Liu, Z. Sun, and H. Tyagi. "Interactive Inference Under Information Constraints". In: *IEEE Transactions on Information Theory* 68.1 (2022), pp. 502–516.

[ACT20a] J. Acharya, C. L. Canonne, and H. Tyagi. "Inference Under Information Constraints I: Lower Bounds From Chi-Square Contraction". In: *IEEE Transactions on Information Theory* 66.12 (2020).

[ACT20b] J. Acharya, C. L. Canonne, and H. Tyagi. "Inference Under Information Constraints II: Communication Constraints and Shared Randomness". In: *IEEE Transactions on Information Theory* 66.12 (2020).

[AFT22] H. Asi, V. Feldman, and K. Talwar. "Optimal Algorithms for Mean Estimation under Local Differential Privacy". In: *Proc. 39th International Conference on Machine Learning (ICML)*. 2022.

[AH98] S. Amari and T. S. Han. "Statistical inference under multiterminal data compression". In: *IEEE Transactions on Information Theory* 44.6 (1998), pp. 2300–2324.

[AK01] S. Arora and R. Kannan. "Learning mixtures of arbitrary Gaussians". In: *Proc. 33rd Annual ACM Symposium on Theory of Computing (STOC)*. 2001.

[AKMVV16] A. Abdullah, R. Kumar, A. McGregor, S. Vassilvitskii, and S. Venkatasubramanian. "Sketching, Embedding and Dimensionality Reduction in Information Theoretic Spaces". In: *Proc. 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2016.

[AM05] D. Achlioptas and F. McSherry. "On spectral learning of mixtures of distributions". In: *Proc. 18th Annual Conference on Learning Theory (COLT)*. 2005.

[AMS99] N. Alon, Y. Matias, and M. Szegedy. "The Space Complexity of Approximating the Frequency Moments". In: *Journal of Computer and System Sciences* 58.1 (1999), pp. 137–147.

[AS98] P. K. Agarwal and M. Sharir. "Efficient algorithms for geometric optimization". In: *ACM Computing Surveys* 30.4 (1998), pp. 412–458.

[AV19] M. E. Ahsen and M. Vidyasagar. "An Approach to One-Bit Compressed Sensing Based on Probably Approximately Correct Learning Theory". In: *Journal of Machine Learning Research* 20.11 (2019), pp. 1–23.

[AZ22] S. Asoodeh and H. Zhang. "Contraction of Locally Differentially Private Mechanisms". In: *CoRR* abs/2210.13386 (2022).

[BB20] M. Brennan and G. Bresler. "Reducibility and Statistical-Computational Gaps from Secret Leakage". In: *Proc. 33rd Annual Conference on Learning Theory (COLT)*. 2020.

[BBHLS21] M. Brennan, G. Bresler, S. B. Hopkins, J. Li, and T. Schramm. "Statistical query algorithms and low-degree tests are almost equivalent". In: *Proc. 34th Annual Conference on Learning Theory* (*COLT*). 2021.

[BBV08] M. F. Balcan, A. Blum, and S. Vempala. "A discriminative framework for clustering via similarity functions". In: *Proc. 40th Annual ACM Symposium on Theory of Computing* (*STOC*). 2008.

[BCL13] Sebastien Bubeck, N. Cesa-Bianchi, and G. Lugosi. "Bandits With Heavy Tail". In: *IEEE Transactions on Information Theory* 59.11 (Nov. 2013), pp. 7711–7717.

[BCN18] L. Bottou, F. E. Curtis, and J. Nocedal. "Optimization Methods for Large-Scale Machine Learning". In: *SIAM Review* 60.2 (2018), pp. 223–311.

[BCÖ20] L. P. Barnes, W-N. Chen, and A. Özgür. "Fisher information under local differential privacy". In: *IEEE Journal on Selected Areas in Information Theory* 1.3 (2020), pp. 645–659.

[BDHKKK20] A. Bakshi, I. Diakonikolas, S. B. Hopkins, D. M. Kane, S. Karmalkar, and P. K. Kothari. "Outlier-Robust Clustering of Gaussians and Other Non-Spherical Mixtures". In: *Proc. 61st IEEE Symposium on Foundations of Computer Science* (*FOCS*). 2020.

[BDJKKV22] A. Bakshi, I. Diakonikolas, H. Jia, D. M. Kane, P. K. Kothari, and S. S. Vempala. "Robustly Learning Mixtures of k Arbitrary Gaussians". In: *Proc. 54th Annual ACM Symposium on Theory of Computing* (*STOC*). 2022.

[BDLS17] Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. "Computationally Efficient Robust Sparse Estimation in High Dimensions". In: *Proc. 30th Annual Conference on Learning Theory* (*COLT*). Vol. 65. 2017, pp. 1–44.

[BEMMRLRKTS17] A. Bittau, Ú. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld. "Prochlo: Strong Privacy for Analytics in the Crowd". In: *Proc. of the 26th Symposium on Operating Systems Principles*. 2017.

[Ber79] T. Berger. “Decentralized estimation and decision theory”. In: *IEEE Seven Springs Workshop on Information Theory*. 1979.

[BFJKMR94] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. “Weakly learning DNF and characterizing statistical query learning using Fourier analysis”. In: *Proc. 26th Annual ACM Symposium on Theory of Computing (STOC)*. 1994.

[BGKS20] J. Banks, J. Garza-Vargas, A. Kulkarni, and N. Srivastava. “Pseudospectral shattering, the sign function, and diagonalization in nearly matrix multiplication time”. In: *Proc. 61st IEEE Symposium on Foundations of Computer Science (FOCS)*. 2020.

[BGMNW16] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff. “Communication Lower Bounds for Statistical Estimation Problems via a Distributed Data Processing Inequality”. In: *Proc. 49th Annual ACM Symposium on Theory of Computing (STOC)*. 2016.

[BGZ22] M. Braverman, S. Garg, and O. Zamir. “Tight space complexity of the coin problem”. In: *Proc. 62nd IEEE Symposium on Foundations of Computer Science (FOCS)*. 2022.

[BHK20] A. Blum, J. Hopcroft, and R. Kannan. *Foundations of Data Science*. First. Cambridge University Press, Jan. 2020.

[BHÖ20] L. P. Barnes, Y. Han, and A. Özgür. “Lower bounds for learning distributions under communication constraints via Fisher information”. In: *Journal of Machine Learning Research* 21.1 (2020), pp. 9583–9612.

[Bir01a] L. Birgé. “An Alternative Point of View on Lepski’s Method”. In: *Lecture Notes-Monograph Series* (2001), pp. 113–133.

[Bir01b] L. Birgé. “An alternative point of view on Lepski’s method”. In: *Lecture Notes-Monograph Series* (2001), pp. 113–133.

[BJK15] K. Bhatia, P. Jain, and P. Kar. “Robust Regression via Hard Thresholding”. In: *Advances in Neural Information Processing Systems 28 (NeurIPS)*. 2015.

[BJKK17] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar. “Consistent Robust Regression”. In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*. 2017.

[BK21] A. Bakshi and P. K. Kothari. "List-Decodable Subspace Recovery: Dimension Independent Error in Polynomial Time". In: *Proc. 32nd Annual Symposium on Discrete Algorithms* (*SODA*). 2021.

[BKSW19] M. Bun, G. Kamath, T. Steinke, and Z. S. Wu. "Private Hypothesis Selection". In: *Advances in Neural Information Processing Systems 32* (*NeurIPS*). 2019.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BM22] D. Bartl and S. Mendelson. "On Monte-Carlo methods in convex stochastic optimization". In: *The Annals of Applied Probability* 32.4 (2022), pp. 3146–3198.

[BNJT10] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar. "The security of machine learning". In: *Machine Learning* 81.2 (2010), pp. 121–148.

[BNL12] B. Biggio, B. Nelson, and P. Laskov. "Poisoning Attacks against Support Vector Machines". In: *Proc. 29th International Conference on Machine Learning* (*ICML*). 2012.

[BNOP21] A. Bhatt, B. Nazer, O. Ordentlich, and Y. Polyanskiy. "Information-Distilling Quantizers". In: *IEEE Transactions on Information Theory* 67.4 (2021), pp. 2472–2487.

[Bog98] V. Bogachev. *Gaussian measures*. Mathematical surveys and monographs, vol. 62, 1998.

[BOS20] T. Berg, O. Ordentlich, and O. Shayevitz. "Binary Hypothesis Testing with Deterministic Finite-Memory Decision Rules". In: *Proc. 2020 IEEE International Symposium on Information Theory*. 2020.

[Bot10] L. Bottou. "Large-scale machine learning with stochastic gradient descent". In: *Proc. 19th International Conference on Computational Statistics* (*COMPSTAT*). 2010.

[BP21] A. Bakshi and A. Prasad. "Robust Linear Regression: Optimal Rates in Polynomial Time". In: *Proc. 53rd Annual ACM Symposium on Theory of Computing* (*STOC*). ACM Press, 2021, pp. 102–115.

[BP22] A. Bhatt and A. Pensia. “Sharp Concentration Inequalities for the Centered Relative Entropy”. In: *Information and Inference: A Journal of the IMA* (2022).

[BS16] M. Bun and T. Steinke. “Concentrated differential privacy: Simplifications, extensions, and lower bounds”. In: *Theory of Cryptography Conference*. 2016.

[BT97] D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, 1997.

[Bub15] S. Bubeck. “Convex Optimization: Algorithms and Complexity”. In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[BV04] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge, UK ; New York: Cambridge University Press, 2004.

[Cam86] L. L. Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. New York, NY: Springer New York, 1986.

[Can22] C. L. Canonne. “Topics and Techniques in Distribution Testing: A Biased but Representative Sample”. 2022.

[Cat12] O. Catoni. “Challenging the Empirical Mean and Empirical Variance: A Deviation Study”. In: *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* 48.4 (2012), pp. 1148–1185.

[CATJFB20] Y. Cherapanamjeri, E. Aras, N. Tripuraneni, M. I. Jordan, N. Flammarion, and P. L. Bartlett. “Optimal Robust Linear Regression in Nearly Linear Time”. In: abs/2007.08137 (2020).

[CDG19] Y. Cheng, I. Diakonikolas, and R. Ge. “High-Dimensional Robust Mean Estimation in Nearly-Linear Time”. In: *Proc. 30th Annual Symposium on Discrete Algorithms (SODA)*. SIAM, 2019, pp. 2755–2771.

[CDGS20] Y. Cheng, I. Diakonikolas, R. Ge, and M. Soltanolkotabi. “High-Dimensional Robust Mean Estimation via Gradient Descent”. In: *Proc. 37th International Conference on Machine Learning (ICML)*. 2020.

[CDGW19] Y. Cheng, I. Diakonikolas, R. Ge, and D. P. Woodruff. "Faster Algorithms for High-Dimensional Robust Covariance Estimation". In: *Proc. 32nd Annual Conference on Learning Theory* (*COLT*). 2019.

[CDKGGS22] Y. Cheng, I. Diakonikolas, D. M. Kane, R. Ge, S. Gupta, and M. Soltanolkotabi. "Outlier-Robust Sparse Estimation via Non-Convex Optimization". In: *Advances in Neural Information Processing Systems 35* (*NeurIPS*). 2022.

[CDKL14] F. Chierichetti, A. Dasgupta, R. Kumar, and S. Lattanzi. "Learning Entangled Single-Sample Gaussians". In: *Proc. 25th Annual Symposium on Discrete Algorithms* (*SODA*). SIAM, 2014, pp. 511–522.

[CDKS18] Y. Cheng, I. Diakonikolas, D. Kane, and A. Stewart. "Robust Learning of Fixed-Structure Bayesian Networks". In: *Advances in Neural Information Processing Systems 31* (*NeurIPS*). 2018.

[CFB19] Y. Cherapanamjeri, N. Flammarion, and P. L. Bartlett. "Fast Mean Estimation with Sub-Gaussian Rates". In: *Proc. 32nd Annual Conference on Learning Theory* (*COLT*). 2019.

[CFJ13] T. Cai, J. Fan, and T. Jiang. "Distributions of angles in random packing on spheres". In: *Journal of Machine Learning Research* 14.1 (2013), pp. 1837–1864.

[CGE21] F. Carpi, S. Garg, and E. Erkip. "Single-shot compression for hypothesis testing". In: *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications* (*SPAWC*). IEEE. 2021, pp. 176–180.

[CGR16] M. Chen, C. Gao, and Z. Ren. "A General Decision Theory for Huber's ϵ-Contamination Model". In: *Electronic Journal of Statistics* 10.2 (2016), pp. 3752–3774.

[Che64] H. Chernoff. "Estimation of the mode". In: *Annals of the Institute of Statistical Mathematics* 16.1 (1964), pp. 31–41.

[CHKRT20] Y. Cherapanamjeri, S. B. Hopkins, T. Kathuria, P. Raghavendra, and N. Tripuraneni. "Algorithms for Heavy-Tailed Statistics: Regression, Covariance Estimation, and Beyond". In: *Proc. 52nd An-*

nual ACM Symposium on Theory of Computing (*STOC*). ACM Press, 2020, pp. 601–609.

[CKMSU19] C. L. Canonne, G. Kamath, A. McMillan, A. Smith, and J. Ullman. "The Structure of Optimal Private Tests for Simple Hypotheses". In: *Proc. 51st Annual ACM Symposium on Theory of Computing* (*STOC*). 2019.

[CKO21] W.-N. Chen, P. Kairouz, and A. Ozgur. "Pointwise Bounds for Distribution Estimation under Communication Constraints". In: *Advances in Neural Information Processing Systems 34* (*NeurIPS*). 2021.

[CLS20] S. Chen, J. Li, and Z. Song. "Learning mixtures of linear regressions in subexponential time via Fourier moments". In: *Proc. 52nd Annual ACM Symposium on Theory of Computing* (*STOC*). 2020.

[CMY20] Y. Cherapanamjeri, S. Mohanty, and M. Yau. "List decodable mean estimation in nearly linear time". In: *Proc. 61st IEEE Symposium on Foundations of Computer Science* (*FOCS*). 2020.

[Cov69] T. M. Cover. "Hypothesis Testing with Finite Statistics". In: *The Annals of Mathematical Statistics* 40.3 (1969), pp. 828–835.

[CS02] M. Charikar and A. Sahai. "Dimension Reduction in the ℓ_1 Norm". In: *Proc. 43rd IEEE Symposium on Foundations of Computer Science* (*FOCS*). 2002.

[CSUZZ19] A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev. "Distributed Differential Privacy via Shuffling". In: *Advances in Cryptology – EUROCRYPT 2019*. 2019.

[CSV17] M. Charikar, J. Steinhardt, and G. Valiant. "Learning from Untrusted Data". In: *Proc. 49th Annual ACM Symposium on Theory of Computing* (*STOC*). ACM Press, 2017, pp. 47–60.

[CTBJ22] Y. Cherapanamjeri, N. Tripuraneni, P. L. Bartlett, and M. I. Jordan. "Optimal Mean Estimation without a Variance". In: *Proc. 35th Annual Conference on Learning Theory* (*COLT*). 2022.

[CW82] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. New York: Chapman and Hall, 1982.

[CY16] P. Cuff and L. Yu. "Differential privacy as a mutual information constraint". In: *Proc. 2016 ACM SIGSAC Conference on Computer and Communications Security*. 2016, pp. 43–54.

[Das99] S. Dasgupta. "Learning mixtures of Gaussians". In: *Proc. 40th IEEE Symposium on Foundations of Computer Science (FOCS)*. 1999.

[Dav93] P. L. Davies. "Aspects of robust linear regression". In: *The Annals of Statistics* (1993), pp. 1843–1899.

[Dep20a] J. Depersin. "A Spectral Algorithm for Robust Regression with Subgaussian Rates". In: *CoRR* abs/2007.06072 (2020).

[Dep20b] J. Depersin. "Robust Subgaussian Estimation with VC-dimension". In: abs/2004.11734 (2020).

[DeV89] R. D. DeVeaux. "Mixtures of linear regressions". In: *Computational Statistics & Data Analysis* 8.3 (Nov. 1989), pp. 227–245.

[DG92] D. L. Donoho and M. Gasko. "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness". In: *The Annals of Statistics* 20.4 (Dec. 1992), pp. 1803–1827.

[dG99] V. H. de la Peña and E. Giné. *Decoupling*. Springer New York, 1999.

[DGKR19] I. Diakonikolas, T. Gouleakis, D. M. Kane, and S. Rao. "Communication and Memory Efficient Testing of Discrete Distributions". In: *Proc. 32nd Annual Conference on Learning Theory (COLT)*. 2019.

[DHL19] Y. Dong, S. B. Hopkins, and J. Li. "Quantum Entropy Scoring for Fast Robust Mean Estimation and Improved Outlier Detection". In: *Advances in Neural Information Processing Systems 32 (NeurIPS)*. 2019.

[Die01] T. E. Dielman. *Applied Regression Analysis for Business and Economics*. Duxbury/Thomson Learning, 2001.

[DJW18] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. "Minimax Optimal Procedures for Locally Private Estimation". In: *Journal of the American Statistical Association* 113.521 (2018), pp. 182–201.

[DJWZ14] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang. "Optimality Guarantees for Distributed Statistical Estimation". In: *CoRR* abs/1405.0782 (2014).

[DK14] S. Dasgupta and S. Kpotufe. "Optimal rates for k-NN density and mode estimation". In: *Advances in Neural Information Processing Systems 27* (*NeurIPS*). 2014.

[DK19] I. Diakonikolas and D. M. Kane. "Recent Advances in Algorithmic High-Dimensional Robust Statistics". In: *CoRR* abs/1911.05911 (2019).

[DK20] I. Diakonikolas and D. M. Kane. "Small Covers for Near-Zero Sets of Polynomials and Learning Latent Variable Models". In: *Proc. 61st IEEE Symposium on Foundations of Computer Science* (*FOCS*). 2020.

[DKBR07] M. Dundar, B. Krishnapuram, J. Bi, and R. B. Rao. "Learning Classifiers When the Training Data Is Not IID". In: *Proc. 20th International Joint Conference on Artificial Intelligence* (*IJCAI*). 2007.

[DKK20] I. Diakonikolas, D. M. Kane, and D. Kongsgaard. "List-decodable mean estimation via iterative multi-filtering". In: *Advances in Neural Information Processing Systems 33* (*NeurIPS*). 2020.

[DKKLMS16] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. "Robust Estimators in High Dimensions without the Computational Intractability". In: *Proc. 57th IEEE Symposium on Foundations of Computer Science* (*FOCS*). 2016, pp. 655–664.

[DKKLMS17] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. "Being Robust (in High Dimensions) Can Be Practical". In: *Proc. 34th International Conference on Machine Learning* (*ICML*). 2017.

[DKKLMS21] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. "Robustness meets algorithms". In: *Communications of the ACM* 64.5 (2021), pp. 107–115.

[DKKLSS19] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, J. Steinhardt, and A. Stewart. "Sever: A Robust Meta-Algorithm for Stochastic Optimization". In: *Proc. 36th International Conference on Machine Learning* (*ICML*). 2019.

[DKKLT21] I. Diakonikolas, D. M. Kane, D. Kongsgaard, J. Li, and K. Tian. "List-decodable mean estimation in nearly-pca time". In: *Advances in Neural Information Processing Systems 34* (*NeurIPS*). Vol. 34. 2021.

[DKKLT22] I. Diakonikolas, D. M. Kane, D. Kongsgaard, J. Li, and K. Tian. "Clustering Mixture Models in Almost-Linear Time via List-Decodable Mean Estimation". In: *Proc. 54th Annual ACM Symposium on Theory of Computing* (*STOC*). 2022.

[DKKPP22a] I. Diakonikolas, D. M. Kane, S. Karmalkar, A. Pensia, and T. Pittas. "List-Decodable Sparse Mean Estimation via Difference-of-Pairs Filtering". In: *Advances in Neural Information Processing Systems 35* (*NeurIPS*). 2022.

[DKKPP22b] I. Diakonikolas, D. M. Kane, S. Karmalkar, A. Pensia, and T. Pittas. "Robust Sparse Mean Estimation via Sum of Squares". In: *Proc. 35th Annual Conference on Learning Theory* (*COLT*). 2022.

[DKKPS19] I. Diakonikolas, D. M. Kane, S. Karmalkar, E. Price, and A. Stewart. "Outlier-Robust High-Dimensional Sparse Estimation via Iterative Filtering". In: *Advances in Neural Information Processing Systems 32* (*NeurIPS*). 2019.

[DKLP22] I. Diakonikolas, D. M. Kane, J. C. H. Lee, and A. Pensia. "Outlier-Robust Sparse Mean Estimation for Heavy-Tailed Distributions". In: *Advances in Neural Information Processing Systems 35* (*NeurIPS*). 2022.

[DKP20] I. Diakonikolas, D. M. Kane, and A. Pensia. "Outlier Robust Mean Estimation with Subgaussian Rates via Stability". In: *Advances in Neural Information Processing Systems 33* (*NeurIPS*). 2020.

[DKP23] I. Diakonikolas, D. M. Kane, and A. Pensia. "Gaussian Mean Testing Made Simple". In: *Proc. 6th Symposium on Simplicity in Algorithms* (*SOSA*). 2023.

[DKPP22] I. Diakonikolas, D. M. Kane, A. Pensia, and T. Pittas. "Streaming Algorithms for High-Dimensional Robust Statistics". In: *Proc. 39th International Conference on Machine Learning* (*ICML*). 2022.

[DKPPS21] I. Diakonikolas, D. M. Kane, A. Pensia, T. Pittas, and A. Stewart. "Statistical query lower bounds for list-decodable linear regression". In: *Advances in Neural Information Processing Systems 34* (*NeurIPS*). 2021.

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. "Statistical Query Lower Bounds for Robust Estimation of High-Dimensional Gaussians and Gaussian Mixtures". In: *Proc. 58th IEEE Symposium on Foundations of Computer Science* (*FOCS*). 2017, pp. 73–84.

[DKS18] I. Diakonikolas, D. M. Kane, and A. Stewart. "List-Decodable Robust Mean Estimation and Learning Mixtures of Spherical Gaussians". In: *Proc. 50th Annual ACM Symposium on Theory of Computing* (*STOC*). 2018.

[DKS19] I. Diakonikolas, W. Kong, and A. Stewart. "Efficient Algorithms and Lower Bounds for Robust Linear Regression". In: *Proc. 30th Annual Symposium on Discrete Algorithms* (*SODA*). 2019.

[DKSS21] I. Diakonikolas, D. M. Kane, A. Stewart, and Y. Sun. "Outlier-Robust Learning of Ising Models Under Dobrushin's Condition". In: *Proc. 34th Annual Conference on Learning Theory* (*COLT*). 2021.

[DKTZ20] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. "Non-Convex SGD Learns Halfspaces with Adversarial Label Noise". In: *Advances in Neural Information Processing Systems 33* (*NeurIPS*). 2020.

[DL01] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. New York, NY: Springer New York, 2001.

[DL21] J. Depersin and G. Lecué. "Optimal Robust Mean and Location Estimation via Convex Programs with Respect to Any Pseudo-Norms". In: *abs/2102.00995* (2021).

[DL22a] J. Depersin and G. Lecué. "Robust Sub-Gaussian Estimation of a Mean Vector in Nearly Linear Time". In: *The Annals of Statistics* 50.1 (Feb. 2022), pp. 511–536.

[DL22b] J. Depersin and G. Lecué. “Robust Subgaussian Estimation of a Mean Vector in Nearly Linear Time”. In: *The Annals of Statistics* 50.1 (2022), pp. 511–536.

[DL88] D. L. Donoho and R. C. Liu. “The "Automatic" Robustness of Minimum Distance Functionals”. In: *The Annals of Statistics* 16.2 (1988), pp. 552–586.

[DLLO16] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. “Sub-Gaussian Mean Estimators”. In: *The Annals of Statistics* 44.6 (Dec. 2016), pp. 2695–2725.

[DLLZ23] L. Devroye, S. Lattanzi, G. Lugosi, and N. Zhivotovskiy. “On mean estimation for heteroscedastic random variables”. In: *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* 59.1 (2023), pp. 1–20.

[DMT07] C. Dwork, F. McSherry, and K. Talwar. “The Price of Privacy and the Limits of LP Decoding”. In: *Proc. 39th Annual ACM Symposium on Theory of Computing (STOC)*. 2007, pp. 85–94.

[DR13] C. Dwork and A. Roth. “The Algorithmic Foundations of Differential Privacy”. In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4 (2013), pp. 211–407.

[DR19] J. C. Duchi and R. Rogers. “Lower Bounds for Locally Private Estimation via Communication Complexity”. In: *Proc. 32nd Annual Conference on Learning Theory (COLT)*. 2019.

[DS18] Y. Dagan and O. Shamir. “Detecting Correlations with Little Memory and Communication”. In: *Proc. 31st Annual Conference on Learning Theory (COLT)*. 2018.

[EE94] D. Eppstein and J. Erickson. “Iterated nearest neighbors and finding minimal polytopes”. In: *Discrete & Computational Geometry* 11.3 (1994), pp. 321–350.

[EH14] T. van Erven and P. Harremos. “Rényi Divergence and Kullback-Leibler Divergence”. In: *IEEE Transactions on Information Theory* 60.7 (2014), pp. 3797–3820.

[EH99] F. El Bantli and M. Hallin. "L_1-Estimation in Linear Models with Heterogeneous White Noise". en. In: *Statistics & Probability Letters* 45.4 (Dec. 1999), pp. 305–315.

[EK12] Y. C. Eldar and G. Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[EY07] A. Eremenko and P. Yuditskii. "Uniform approximation of sgn x by polynomials and entire functions". In: *Journal d'Analyse Mathématique* 101.1 (2007), pp. 313–324.

[Fel16] V. Feldman. "Statistical Query Learning". In: *Encyclopedia of Algorithms*. Springer New York, 2016, pp. 2090–2095.

[Fel17] V. Feldman. "A General Characterization of the Statistical Query Complexity". In: *Proc. 30th Annual Conference on Learning Theory (COLT)*. 2017.

[FGRVX17] V. Feldman, E. Grigorescu, L. Reyzin, S. S. Vempala, and Y. Xiao. "Statistical Algorithms and a Lower Bound for Detecting Planted Cliques". In: *Journal of the ACM* 64.2 (2017), 8:1–8:37.

[FGV17] V. Feldman, C. Guzman, and S. S. Vempala. "Statistical Query Algorithms for Mean Vector Estimation and Stochastic Convex Optimization". In: *Proc. 28th Annual Symposium on Discrete Algorithms (SODA)*. 2017.

[FLW17] J. Fan, Q. Li, and Y. Wang. "Estimation of High Dimensional Mean Regression in the Absence of Symmetry and Light Tail Assumptions". In: *Journal of the Royal Statistical Society Series B* 79.1 (2017), pp. 247–265.

[FMT21] V. Feldman, A. McMillan, and K. Talwar. "Hiding Among the Clones: A Simple and Nearly Optimal Analysis of Privacy Amplification by Shuffling". In: *Proc. 62nd IEEE Symposium on Foundations of Computer Science (FOCS)*. 2021.

[FNS16] S. R. Flaxman, D. B. Neill, and A. J. Smola. "Gaussian processes for independence tests with non-iid data in causal inference". In: *ACM Transactions on Intelligent Systems and Technology* 7.2 (2016), p. 22.

[FPV18] V. Feldman, W. Perkins, and S. Vempala. "On the Complexity of Random Satisfiability Problems with Planted Solutions". In: *SIAM Journal on Computing* 47.4 (2018), pp. 1294–1338.

[Gan02] M. I. Ganzburg. "Limit theorems for polynomial approximation with Hermite and Freud weights". In: *Approximation Theory X: Abstract and Classical Analysis (CK Chui, et al, eds.)* (2002), pp. 211–221.

[GGKMZ21] B. Ghazi, N. Golowich, R. Kumar, P. Manurangsi, and C. Zhang. "Deep Learning with Label Differential Privacy". In: *Advances in Neural Information Processing Systems 34 (NeurIPS)*. 2021.

[Gil06] G. L. Gilardoni. "On the Minimum F-Divergence for given Total Variation". In: *Comptes Rendus Mathematique* 343.11-12 (2006), pp. 763–766.

[GJ95] P. W. Goldberg and M. R. Jerrum. "Bounding the Vapnik-Chervonenkis Dimension of Concept Classes Parameterized by Real Numbers". In: *Machine Learning* 18.2 (1995), pp. 131–148.

[GKKNWZ20] S. Gopi, G. Kamath, J. Kulkarni, A. Nikolov, Z. S. Wu, and H. Zhang. "Locally Private Hypothesis Selection". In: *Proc. 33rd Annual Conference on Learning Theory (COLT)*. 2020.

[GKT51] D. Gale, H. W. Kuhn, and A. W. Tucker. "Linear programming and the theory of games". In: *Activity analysis of production and allocation* 13 (1951), pp. 317–335.

[GR08] M. I. Ganzburg and J. Rognes. *Limit theorems of polynomial approximation with exponential weights*. American Mathematical Soc., 2008.

[GRT18] S. Garg, R. Raz, and A. Tal. "Extractor-Based Time-Space Lower Bounds for Learning". In: *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*. 2018.

[HC71] M. E. Hellman and T. M. Cover. "On memory saved by randomization". In: *The Annals of Mathematical Statistics* (1971), pp. 1075–1078.

[HC73a] M. Hellman and T. Cover. “A Review of Recent Results on Learning with Finite Memory”. In: *International Symposium on Information Theory (ISIT)*. 1973, pp. 289–294.

[HC73b] M. E. Hellman and T. M. Cover. “Learning with finite memory”. In: *Matematika* 17.3 (1973), pp. 137–156.

[Hel74] M. Hellman. “Finite-memory algorithms for estimating the mean of a Gaussian distribution”. In: *IEEE Transactions on Information Theory* 20.3 (1974), pp. 382–384.

[HKPRSS17] S. B. Hopkins, P. K. Kothari, A. Potechin, P. Raghavendra, T. Schramm, and D. Steurer. “The Power of Sum-of-Squares for Detecting Hidden Structures”. In: *Proc. 58th IEEE Symposium on Foundations of Computer Science (FOCS)*. 2017.

[HL18] S. B. Hopkins and J. Li. “Mixture Models, Robustness, and Sum of Squares Proofs”. In: *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*. 2018.

[HL19] S. B. Hopkins and J. Li. “How Hard Is Robust Mean Estimation?” In: *Proc. 32nd Annual Conference on Learning Theory (COLT)*. 2019.

[HLM17] N. Holohan, D. J. Leith, and O. Mason. “Extreme Points of the Local Differential Privacy Polytope”. In: *Linear Algebra and its Applications* 534 (2017), pp. 78–96.

[HLZ20] S. B. Hopkins, J. Li, and F. Zhang. “Robust and Heavy-Tailed Mean Estimation Made Simple, via Regret Minimization”. In: *Advances in Neural Information Processing Systems 33 (NeurIPS)*. 2020.

[HM01] M. Hallin and I. Mizera. “Sample Heterogeneity and M-Estimation”. In: *Journal of Statistical Planning and Inference* 93 (Feb. 2001), pp. 139–160.

[HM97] M. Hallin and I. Mizera. *Unimodality and the Asymptotics of M-Estimators*. Vol. 31. 1997, pp. 47–56.

[Hoe56] W. Hoeffding. “On the distribution of the number of successes in independent trials”. In: *The Annals of Mathematical Statistics* 27 (1956), pp. 713–721.

[Hop18] S. B. Hopkins. “Statistical inference and the sum of squares method”. PhD thesis. Cornell University, 2018.

[Hop20] S. B. Hopkins. "Mean Estimation with Sub-Gaussian Rates in Polynomial Time". In: *The Annals of Statistics* 48.2 (2020), pp. 1193–1213.

[Hor07] K. J. Horadam. *Hadamard Matrices and Their Applications*. Princeton, N.J: Princeton University Press, 2007.

[HÖW21] Y. Han, A. Özgür, and T. Weissman. "Geometric Lower Bounds for Distributed Parameter Estimation Under Communication Constraints". In: *IEEE Transactions on Information Theory* 67.12 (2021), pp. 8248–8263.

[HR09] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley & Sons, 2009.

[HRRS11] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Vol. 196. John Wiley & Sons, 2011.

[HS16] D. Hsu and S. Sabato. "Loss Minimization and Parameter Estimation with Heavy Tails". In: *Journal of Machine Learning Research* 17.18 (2016), pp. 1–40.

[HS17] S. B. Hopkins and D. Steurer. "Efficient Bayesian Estimation from Few Samples: Community Detection and Related Problems". In: *Proc. 58th IEEE Symposium on Foundations of Computer Science (FOCS)*. 2017.

[HS73] P. J. Huber and V. Strassen. "Minimax Tests and the Neyman-Pearson Lemma for Capacities". In: *The Annals of Statistics* 1.2 (1973).

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

[Hub64] P. J. Huber. "Robust Estimation of a Location Parameter". In: *The Annals of Mathematical Statistics* 35.1 (Mar. 1964), pp. 73–101.

[Hub65] P. J. Huber. "A Robust Version of the Probability Ratio Test". In: *The Annals of Mathematical Statistics* 36.6 (1965), pp. 1753–1758.

[Hub73] P. J. Huber. "Robust regression: Asymptotics, conjectures and Monte Carlo". In: *The Annals of Statistics* 1.5 (1973), pp. 799–821.

[Jia17] H. Jiang. "Uniform convergence rates for kernel density estimation". In: *Proc. 34th International Conference on Machine Learning* (*ICML*). 2017.

[JJ94] M. I. Jordan and R. A. Jacobs. "Hierarchical Mixtures of Experts and the EM Algorithm". In: *Neural Computation* 6.2 (1994), pp. 181–214.

[JK17] P. Jain and P. Kar. "Non-Convex Optimization for Machine Learning". In: *Foundations and Trends® in Machine Learning* 10.3-4 (2017), pp. 142–336.

[JLT20] A. Jambulapati, J. Li, and K. Tian. "Robust sub-gaussian principal component analysis and width-independent schatten packing". In: *Advances in Neural Information Processing Systems 33* (*NeurIPS*) (2020). arxiv preprint at https://arxiv.org/abs/2006.06980.

[JMNR19] M. Joseph, J. Mao, S. Neel, and A. Roth. "The Role of Interactivity in Local Differential Privacy". In: *Proc. 60th IEEE Symposium on Foundations of Computer Science* (*FOCS*). 2019.

[JTK14] P. Jain, A. Tewari, and P. Kar. "On Iterative Hard Thresholding Methods for High-Dimensional M-Estimation". In: *Advances in Neural Information Processing Systems 27* (*NeurIPS*). 2014.

[JVV86] M. Jerrum, L. G. Valiant, and V. V. Vazirani. "Random Generation of Combinatorial Structures from a Uniform Distribution". In: *Theor. Comput. Sci.* 43 (1986), pp. 169–188.

[KC20] J. Kwon and C. Caramanis. "EM Converges for a Mixture of Many Linear Regressions". In: *Proc. 23rd International Conference on Artificial Intelligence and Statistics* (*AISTATS*). PMLR. 2020.

[Kea98] M. J. Kearns. "Efficient noise-tolerant Learning from Statistical Queries". In: *Journal of the ACM* 45.6 (1998), pp. 983–1006.

[KKK19] S. Karmalkar, A. Klivans, and P. K. Kothari. "List-Decodable Linear Regression". In: *Advances in Neural Information Processing Systems 32* (*NeurIPS*). 2019.

[KKM18] A. Klivans, P. K. Kothari, and R. Meka. "Efficient Algorithms for Outlier-Robust Regression". In: *Proc. 31st Annual Conference on Learning Theory* (*COLT*). 2018.

[KM15] V. Koltchinskii and S. Mendelson. "Bounding the Smallest Singular Value of a Random Matrix without Concentration". In: *International Mathematics Research Notices* 2015.23 (Mar. 2015), pp. 12991–13008.

[Kni99] K. Knight. "Asymptotics for L_1-Estimators of Regression Parameters under heteroscedasticity". In: *Canadian Journal of Statistics* 27.3 (1999), pp. 497–507.

[KOV16] P. Kairouz, S. Oh, and P. Viswanath. "Extremal Mechanisms for Local Differential Privacy". In: *Journal of Machine Learning Research* 17 (2016), 17:1–17:51.

[KP19] S. Karmalkar and E. Price. "Compressed Sensing with Adversarial Sparse Noise via L1 Regression". In: *Proc. 2nd Symposium on Simplicity in Algorithms (SOSA)*. 2019.

[KP90] J. Kim and D. Pollard. "Cube Root Asymptotics". In: *The Annals of Statistics* (1990), pp. 191–219.

[Kra04] I. Krasikov. "New Bounds on the Hermite Polynomials". In: *arXiv preprint math/0401310* (2004).

[KS53] S. Karlin and L. S. Shapley. *Geometry of moment spaces*. Vol. 12. American Mathematical Society, 1953.

[KSS18] P. K. Kothari, J. Steinhardt, and D. Steurer. "Robust Moment Estimation and Improved Clustering via Sum of Squares". In: *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*. ACM Press, 2018, pp. 1035–1046.

[KSV05] R. Kannan, H. Salmasian, and S. Vempala. "The spectral method for general mixture models". In: *Proc. 18th Annual Conference on Learning Theory (COLT)*. 2005.

[LATSCR+08] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers. "Worldwide human relationships inferred from genome-wide patterns of variation". In: *Science* 319 (5866 2008), pp. 1100–1104.

[LDB09] J. N. Laska, M. A. Davenport, and R. G. Baraniuk. "Exact Signal Recovery from Sparsely Corrupted Measurements through the Pursuit of Justice". In: *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*. 2009, pp. 1556–1560.

[Lep91] O. V. Lepskii. "On a Problem of Adaptive Estimation in Gaussian White Noise". In: *Theory of Probability & Its Applications* 35.3 (1991), pp. 454–466.

[Li18] J. Li. "Principled Approaches to Robust Machine Learning and Beyond". PhD thesis. Massachusetts Institute of Technology, Cambridge, USA, 2018.

[Lin95] B. G. Lindsay. "Mixture Models: Theory, Geometry and Applications". In: *NSF-CBMS Regional Conference Series in Probability and Statistics*. 1995, pp. i–163.

[Liu88] R. Y. Liu. "Bootstrap procedures under some non-iid models". In: *The Annals of Statistics* 16.4 (1988), pp. 1696–1708.

[LL18] Y. Li and Y. Liang. "Learning Mixtures of Linear Regressions with Nearly Optimal Complexity". In: *Proc. 31st Annual Conference on Learning Theory (COLT)*. 2018.

[LL20] G. Lecué and M. Lerasle. "Robust Machine Learning by Median-of-Means: Theory and Practice". In: *The Annals of Statistics* 48.2 (2020), pp. 906–931.

[LLVZ20] Z. Lei, K. Luh, P. Venkat, and F. Zhang. "A Fast Spectral Algorithm for Mean Estimation with Sub-Gaussian Rates". In: *Proc. 33rd Annual Conference on Learning Theory (COLT)*. 2020.

[LM19a] G. Lugosi and S. Mendelson. "Mean Estimation and Regression Under Heavy-Tailed Distributions: A Survey". In: *Foundations of Computational Mathematics* 19.5 (2019), pp. 1145–1190.

[LM19b] G. Lugosi and S. Mendelson. "Near-Optimal Mean Estimators with Respect to General Norms". In: *Probability Theory and Related Fields* 175.3-4 (2019), pp. 957–973.

[LM19c] G. Lugosi and S. Mendelson. "Risk Minimization by Median-of-Means Tournaments". In: *Journal of the European Mathematical Society* 22.3 (2019), pp. 925–965.

[LM19d] G. Lugosi and S. Mendelson. "Sub-Gaussian Estimators of the Mean of a Random Vector". In: *The Annals of Statistics* 47.2 (2019), pp. 783–794.

[LM21a] A. Liu and A. Moitra. "Settling the Robust Learnability of Mixtures of Gaussians". In: *Proc. 53rd Annual ACM Symposium on Theory of Computing (STOC)*. 2021.

[LM21b] G. Lugosi and S. Mendelson. "Robust Multivariate Mean Estimation: The Optimality of Trimmed Mean". In: *The Annals of Statistics* 49.1 (2021), pp. 393–410.

[LMM20] J. Li, A. Marsiglietti, and J. Melbourne. "Further Investigations of Rényi Entropy Power Inequalities and an Entropic Characterization of s-Concave Densities". In: *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2017-2019 Volume II*. Ed. by Bo'az Klartag and Emanuel Milman. 2020, pp. 95–123.

[LMN05] J. R. Lee, M. Mendel, and A. Naor. "Metric Structures in L_1 : Dimension, Snowflakes, and Average Distortion". In: *European Journal of Combinatorics* 26.8 (2005), pp. 1180–1190.

[LP84] D.-T. Lee and F. P. Preparata. "Computational geometry—a survey". In: *IEEE Transactions on Computers* 12 (1984), pp. 1072–1101.

[LR86] F. Leighton and R. Rivest. "Estimating a probability using finite memory". In: *IEEE Transactions on Information Theory* 32.6 (1986), pp. 733–742.

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala. "Agnostic Estimation of Mean and Covariance". In: *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*. 2016, pp. 665–674.

[LSCT17] J. Liao, L. Sankar, F. P. Calmon, and V. YF. Tan. "Hypothesis testing under maximal leakage privacy constraints". In: *Proc. 2017 IEEE International Symposium on Information Theory*. 2017.

[LSTC17] J. Liao, L. Sankar, V. YF. Tan, and F. P. Calmon. "Hypothesis testing under mutual information privacy constraints in the high privacy regime". In: *IEEE Transactions on Information Forensics and Security* 13.4 (2017).

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991.

[LV20] J. C. H. Lee and P. Valiant. "Optimal Sub-Gaussian Mean Estimation in $\mathbb{R}$". In: *Proc. 62nd IEEE Symposium on Foundations of Computer Science (FOCS)* (Nov. 2020).

[LY20] Y. Liang and H. Yuan. "Learning Entangled Single-Sample Gaussians in the Subset-of-Signals Model". In: *Proc. 33rd Annual Conference on Learning Theory (COLT)*. 2020.

[Mal75] C. L. Mallows. "On some topics in robustness". In: *Unpublished Memorandum, Bell Telephone Laboratories, Murray Hill, NJ* 37 (1975).

[Mas90] P. Massart. "The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality". In: *The Annals of Probability* 18.3 (July 1990), pp. 1269–1283.

[McD09] J. H. McDonald. *Handbook of Biological Statistics*. Sparky House Publishing, 2009.

[Men15] S. Mendelson. "Learning without Concentration". In: *Journal of the ACM* 62.3 (2015), pp. 1–25.

[MGJK19] B. Mukhoty, G. Gopakumar, P. Jain, and P. Kar. "Globally-Convergent Iteratively Reweighted Least Squares for Robust Regression Problems". In: *Proc. 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 89. 2019, pp. 313–322.

[Min15] S. Minsker. "Geometric Median and Robust Estimation in Banach Spaces". In: *Bernoulli* 21.4 (2015), pp. 2308–2335.

[Min17] S. Minsker. "On Some Extensions of Bernstein's Inequality for Self-Adjoint Operators". In: *Statistics & Probability Letters* 127 (Aug. 2017), pp. 111–119.

[Min19] S. Minsker. "Uniform Bounds for Robust Mean Estimators". In: *CoRR* abs/1812.03523 (2019).

[Min22] S. Minsker. "U-statistics of growing order and sub-Gaussian mean estimators with sharp constants". In: *CoRR* abs/2202.11842 (2022).

[Mir17] I. Mironov. "Rényi differential privacy". In: *Proc. 2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE. 2017.

[MMYS19] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust Statistics: Theory and Methods (With R)*. John Wiley & Sons, 2019.

[MV18] M. Meister and G. Valiant. "A Data Prism: Semi-verified learning in the small-alpha regime". In: *Proc. 31st Annual Conference on Learning Theory (COLT)*. 2018.

[MW98] I. Mizera and J. A. Wellner. "Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables". In: *The Annals of Statistics* 26.2 (1998), pp. 672–691.

[MZ20] S. Mendelson and N. Zhivotovskiy. "Robust Covariance Estimation under $L_{4}-L_{2}$ Norm Equivalence". In: *The Annals of Statistics* 48.3 (June 2020), pp. 1648–1664.

[Nel73] E. Nelson. "The free Markoff field". In: *Journal of Functional Analysis* 12.2 (1973), pp. 211–227.

[Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Vol. 87. Applied Optimization. Boston, MA: Springer US, 2004.

[NP33] J. Neyman and E. S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231 (1933), pp. 289–337.

[NT13] N. H. Nguyen and T. D. Tran. "Exact Recoverability From Dense Corrupted Observations via $\ell _{1}$-Minimization". In: *IEEE Transactions on Information Theory* 59.4 (2013), pp. 2017–2035.

[NTN11] N. M. Nasrabadi, T. D. Tran, and N. H. Nguyen. "Robust Lasso with Missing and Grossly Corrupted Observations". In: *Advances in Neural Information Processing Systems 24 (NeurIPS)*. 2011.

[NY83] A. S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[ODo14] R. O'Donnell. *Analysis of Boolean Functions*. New York, NY: Cambridge University Press, 2014.

[OO19] R. I. Oliveira and P. Orenstein. "The sub-gaussian property of trimmed means estimators". In: *Technical report, IMPA* (2019).

[PAJL23] A. Pensia, A. A. Asadi, V. Jog, and P. Loh. "Simple Binary Hypothesis Testing under Local Differential Privacy and Communication Constraints". In: *CoRR* arXiv:2301.03566 (2023).

[PBR19] A. Prasad, S. Balakrishnan, and P. Ravikumar. "A Unified Approach to Robust Mean Estimation". In: *CoRR* abs/1907.00927 (July 2019).

[PF20] S. Pesme and N. Flammarion. "Online Robust Regression via SGD on the l1 loss". In: *Advances in Neural Information Processing Systems 33 (NeurIPS)*. 2020.

[PJL18] A. Pensia, V. Jog, and P. Loh. "Generalization Error Bounds for Noisy, Iterative Algorithms". In: *Proc. 2018 IEEE International Symposium on Information Theory*. 2018, pp. 546–550.

[PJL19a] A. Pensia, V. Jog, and P. Loh. "Estimating Location Parameters in Entangled Single-Sample Distributions". In: *CoRR* abs/1907.03087 (2019).

[PJL19b] A. Pensia, V. Jog, and P. Loh. "Mean Estimation for Entangled Single-Sample Distributions". In: *Proc. 2019 IEEE International Symposium on Information Theory*. 2019, pp. 3052–3056.

[PJL20a] A. Pensia, V. Jog, and P. Loh. "Extracting Robust and Accurate Features via a Robust Information Bottleneck". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 131–144.

[PJL20b] A. Pensia, V. Jog, and P. Loh. "Robust Regression with Covariate Filtering: Heavy Tails and Adversarial Contamination". In: *CoRR* abs/2009.12976 (Sept. 2020).

[PJL22] A. Pensia, V. Jog, and P. Loh. "Communication-constrained hypothesis testing: Optimality, robustness, and reverse data processing inequalities". In: *CoRR* arXiv:2206.02765 (2022).

[PLJ22] A. Pensia, P. Loh, and V. Jog. "Simple Binary Hypothesis Testing under Communication Constraints". In: *Proc. 2022 IEEE International Symposium on Information Theory*. 2022.

[PLJD10] P. Paschou, J. Lewis, A. Javed, and P. Drineas. "Ancestry Informative Markers for Fine-Scale Individual Assignment to Worldwide Populations". In: *Journal of Medical Genetics* 47 (12 2010), pp. 835–847.

[PPMZR19] R. Panda, A. Pensia, N. Mehta, M. Zhou, and P. Rai. "Deep Topic Models for Multi-label Learning". In: *Proc. 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2019.

[PSBR20] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar. "Robust Estimation via Robust Gradient Estimation". In: *Journal of the Royal Statistical Society Series B* 82.3 (July 2020), pp. 601–627.

[RL87] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, 1987.

[Rou84] P. J. Rousseeuw. "Least Median of Squares Regression". In: *Journal of the American Statistical Association* 79.388 (1984), pp. 871–880.

[RPWCKZF02] N. Rosenberg, J. Pritchard, J. Weber, H. Cann, K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. "Genetic structure of human populations". In: *Science* 298 (5602 2002), pp. 2381–2385.

[RT70] R. Roberts and J. Tooley. "Estimation with finite memory". In: *IEEE Transactions on Information Theory* 16.6 (1970), pp. 685–691.

[RV06] P. J. Rousseeuw and K. Van Driessen. "Computing LTS Regression for Large Data Sets". In: *Data Mining and Knowledge Discovery* 12.1 (2006), pp. 29–45.

[RWY10] G. Raskutti, M. J. Wainwright, and B. Yu. "Restricted eigenvalue properties for correlated Gaussian designs". In: *Journal of Machine Learning Research* 11.Aug (2010), pp. 2241–2259.

[RY20a] P. Raghavendra and M. Yau. "List Decodable Learning via Sum of Squares". In: *Proc. 31st Annual Symposium on Discrete Algorithms (SODA)*. SIAM, 2020.

[RY20b] P. Raghavendra and M. Yau. "List Decodable Subspace Recovery". In: *Proc. 33rd Annual Conference on Learning Theory (COLT)*. 2020.

[RY84] P. Rousseeuw and V. Yohai. "Robust Regression by Means of S-Estimators". In: *Robust and Nonlinear Time Series Analysis*. Vol. 26. New York, NY: Springer US, 1984, pp. 256–272.

[Sas15] I. Sason. "Tight Bounds for Symmetric Divergence Measures and a New Inequality Relating F-Divergences". In: *2015 IEEE Information Theory Workshop (ITW)*. 2015, pp. 1–5.

[Sas18] I. Sason. "On f-Divergences: Integral Representations, Local Behavior, and Inequalities". In: *Entropy* 20.5 (2018), p. 383.

[SB14] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[SC09] I. Steinwart and A. Christmann. "Fast learning from non-iid observations". In: *Advances in Neural Information Processing Systems 22 (NeurIPS)*. 2009.

[SCV18] J. Steinhardt, M. Charikar, and G. Valiant. "Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers". In: *Proc. 9th Innovations in Theoretical Computer Science Conference (ITCS)*. 2018.

[Sen68] P. K. Sen. "Asymptotic normality of sample quantiles for m-dependent processes". In: *The Annals of Mathematical Statistics* 39.5 (Oct. 1968), pp. 1724–1730.

[Sen70] P. K. Sen. "A note on order statistics for heterogeneous distributions". In: *The Annals of Mathematical Statistics* 41.6 (Dec. 1970), pp. 2137–2139.

[SF20] T. Sasai and H. Fujisawa. "Robust Estimation with Lasso When Outputs Are Adversarially Contaminated". In: *CoRR* abs/2004.05990 (2020).

[Sha14] O. Shamir. "Fundamental Limits of Online and Distributed Algorithms for Statistical Learning and Estimation". In: *Advances in Neural Information Processing Systems 27 (NeurIPS)*. 2014.

[She18] Or Sheffet. "Locally Private Hypothesis Testing". In: *Proc. 35th International Conference on Machine Learning (ICML)*. 2018.

[Sio58] M. Sion. "On General Minimax Theorems." In: *Pacific Journal of Mathematics* 8.1 (1958), pp. 171–176.

[SKL17] J. Steinhardt, P. W. Koh, and P. Liang. “Certified Defenses for Data Poisoning Attacks”. In: *Advances in Neural Information Processing Systems 30* (*NeurIPS*). 2017.

[SO11] Y. She and A. B. Owen. “Outlier Detection Using Nonconvex Penalized Regression”. In: *Journal of the American Statistical Association* 106.494 (2011), pp. 626–639.

[Sti76] S. M Stigler. “The effect of sample heterogeneity on linear functions of order statistics, with applications to robust estimation”. In: *Journal of the American Statistical Association* 71.356 (1976), pp. 956–960.

[Sur21] A. T. Suresh. “Robust Hypothesis Testing and Distribution Estimation in Hellinger Distance”. In: *Proc. 24th International Conference on Artificial Intelligence and Statistics* (*AISTATS*). 2021.

[SVC16] J. Steinhardt, G. Valiant, and M. Charikar. “Avoiding Imposters and Delinquents: Adversarial Crowdsourcing and Peer Prediction”. In: *Advances in Neural Information Processing Systems 29* (*NeurIPS*). 2016.

[SVW16] J. Steinhardt, G. Valiant, and S. Wager. “Memory, Communication, and Statistical Queries”. In: *Proc. 29th Annual Conference on Learning Theory* (*COLT*). 2016.

[SW09] G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics, 2009.

[SWS20] V. Shah, X. Wu, and S. Sanghavi. “Choosing the Sample with Lowest Loss makes SGD Robust”. In: *Proc. 23rd International Conference on Artificial Intelligence and Statistics* (*AISTATS*). 2020.

[Sze89] G. Szegö. *Orthogonal Polynomials*. Vol. XXIII. American Mathematical Society Colloquium Publications. American Mathematical Society, 1989.

[SZF20] Q. Sun, W. Zhou, and J. Fan. “Adaptive Huber Regression”. In: *Journal of the American Statistical Association* 115.529 (2020), pp. 254–265.

[Tal96a] M. Talagrand. "A New Look at Independence". In: *The Annals of Probability* 24.1 (Jan. 1996), pp. 1–34.

[Tal96b] M. Talagrand. "New Concentration Inequalities in Product Spaces". In: *Inventiones Mathematicae* 126.3 (Nov. 1996), pp. 505–563.

[TLM18] B. Tran, J. Li, and A. Madry. "Spectral Signatures in Backdoor Attacks". In: *Advances in Neural Information Processing Systems 31* (*NeurIPS*). 2018.

[TP14] A. M. Tillmann and M. E. Pfetsch. "The Computational Complexity of the Restricted Isometry Property, the Nullspace Property, and Related Concepts in Compressed Sensing". In: *IEEE Transactions on Information Theory* 60.2 (2014), pp. 1248–1259.

[TPBR21] C. Tsai, A. Prasad, S. Balakrishnan, and P. Ravikumar. "Heavy-tailed Streaming Statistical Estimation". In: *Proc. 25th International Conference on Artificial Intelligence and Statistics* (*AISTATS*). 2021.

[Tro15] Joel A. Tropp. "An Introduction to Matrix Concentration Inequalities". In: *Foundations and Trends® in Machine Learning* 8.1-2 (2015), pp. 1–230.

[Tsi88] J. N. Tsitsiklis. "Decentralized Detection by a Large Number of Sensors". In: *Mathematics of Control, Signals, and Systems* 1.2 (1988), pp. 167–182.

[Tsi93] J. N. Tsitsiklis. "Decentralized Detection". In: *Advances in Statistical Signal Processing*. 1993, pp. 297–344.

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2009.

[Tuk60] J. W. Tukey. "A survey of sampling from contaminated distributions". In: *Contributions to probability and statistics* 2 (1960), pp. 448–485.

[Tuk75] J.W. Tukey. "Mathematics and picturing of data". In: *Proceedings of the International Congress of Mathematicians* (*ICM*). Vol. 6. 1975, pp. 523–531.

[Val84] L. Valiant. "A theory of the learnable". In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[Van00] S. A. Van de Geer. *Empirical Processes in M-Estimation*. Vol. 6. Cambridge University Press, 2000.

[van16] S. van de Geer. *Estimation and Testing Under Sparsity*. 1st ed. 2016. École d'Été de Probabilités de Saint-Flour 2159. Springer, 2016.

[VBP94] V.V. Veeravalli, T. Basar, and H.V. Poor. "Minimax Robust Decentralized Detection". In: *IEEE Transactions on Information Theory* 40.1 (1994), pp. 35–40.

[Ver12] R. Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *Compressed Sensing: Theory and Applications*. Ed. by Yonina C. Eldar and GittaEditors Kutyniok. Cambridge University Press, 2012, pp. 210–268.

[Ver18] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[VW09] A. Van Der Vaart and J. A. Wellner. "A note on bounds for VC dimensions". In: *Institute of Mathematical Statistics Collections* 5 (2009), p. 103.

[Wai19] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[Wal45] A. Wald. "Sequential Tests of Statistical Hypotheses". In: *The Annals of Mathematical Statistics* 16.2 (1945), pp. 117–186.

[War65] S. L. Warner. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias". In: *Journal of the American Statistical Association* 60.309 (1965), pp. 63–69.

[WD81] R. S. Wenocur and R. M. Dudley. "Some special Vapnik-Chervonenkis classes". In: *Discrete Mathematics* 33.3 (1981), pp. 313–318.

[Wei69] L. Weiss. "The Asymptotic Distribution of Quantiles from Mixed Samples". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 31.3 (1969), pp. 313–318.

[WLJ07] H. Wang, G. Li, and G. Jiang. "Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso". In: *Journal of Business & Economic Statistics* 25.3 (2007), pp. 347–355.

[Yat85] Y. G. Yatracos. “Rates of Convergence of Minimum Distance Estimators and Kolmogorov’s Entropy”. In: *The Annals of Statistics* 13.2 (1985).

[YCRB19] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett. “Defending against saddle point attack in Byzantine-robust distributed learning”. In: *Proc. 36th International Conference on Machine Learning (ICML)*. 2019.

[Yoh87] V. J. Yohai. “High Breakdown-Point and High Efficiency Robust Estimates for Regression”. In: *The Annals of Statistics* 15.2 (1987), pp. 642–656.

[ZJD16] K. Zhong, P. Jain, and I. S. Dhillon. “Mixed Linear Regression with Multiple Components”. In: *Advances in Neural Information Processing Systems 29 (NeurIPS)*. 2016.

[ZJS22a] B. Zhu, J. Jiao, and J. Steinhardt. “Generalized Resilience and Robust Statistics”. In: *The Annals of Statistics* 50.4 (2022), pp. 2256–2283.

[ZJS22b] B. Zhu, J. Jiao, and J. Steinhardt. “Robust Estimation via Generalized Quasi-Gradients”. In: *Information and Inference: A Journal of the IMA* 11.2 (2022), pp. 581–636.

[ZXLZ15] T. Zhu, P. Xiong, G. Li, and W. Zhou. “Correlated differential privacy: Hiding information in non-IID data set”. In: *IEEE Transactions on Information Forensics and Security* 10.2 (2015), pp. 229–242.

[ZZ73] J. Ziv and M. Zakai. “On Functionals Satisfying a Data-Processing Theorem”. In: *IEEE Transactions on Information Theory* 19.3 (1973), pp. 275–283.

A APPENDIX TO CHAPTER 3

A.1 Robust Mean Estimation and Stability

A.1.1 Robust Mean Estimation from Subset Stability

The theorem statement in [DK19, Theorem 2.7] requires that the input multiset S is stable. We note that the arguments straightforwardly go through when S contains a large stable subset $S' \subseteq S$ (see, e.g., [DKKLMS16; DKKLMS17; DHL19]).

For concreteness, we describe a simple pre-processing of the data, that ensures that the data follows the definition as is: simply throw away points so that the cardinality of the corrupted set matches the cardinality of the stable subset.

Proposition A.1.1. *Let S be a set such that $\exists S' \subseteq S$ such that $|S'| \geq (1-\epsilon)|S|$ and S' is $(C\epsilon, \delta)$ for some $C > 0$. Let T be an ϵ-corrupted version of S. Let T' be the multiset obtained by removing ϵn points of T. Let $\epsilon' = \frac{2\epsilon}{1-\epsilon}$. Then T' is an ϵ'-corrupted version of a $((C-1)\epsilon'/2, \delta)$ stable set.*

Proof. Let T be an ϵ-corrupted version of S. That is, $T = S \cup A \setminus R$. We now remove ϵn points arbitrarily from T to obtain the multiset T' of cardinality $(1-\epsilon)n$.

Let S_2 be any subset of S' such that $|S_2| = |T_1| = (1-\epsilon)n$. Therefore, T' is at most $(2\epsilon)/(1-\epsilon)$-corrupted version of S_2. As S' is $(C\epsilon, \delta)$ stable and S_2 is a large subset of S', Claim A.1.2 states that S_2 is (ϵ_2, δ) stable where $\epsilon_2 \geq 1-(1-C\epsilon)/(1-\epsilon) = (C-1)\epsilon'/2$. □

Claim A.1.2. *If a set S is (ϵ, δ) stable, then its subset S' of cardinality $m > (1-\epsilon)n$ is $(1-(1-\epsilon)\frac{n}{m}, \delta)$ stable.*

Proof. To show that S' is (ϵ', δ) stable, it suffices to ensure that $\epsilon' \leq \epsilon$ and $(1-\epsilon')|S'| \geq (1-\epsilon)|S|$. Therefore, we require that

$$(1-\epsilon')m \geq (1-\epsilon)n \implies \epsilon' \leq 1 - \frac{(1-\epsilon)n}{m}.$$

The upper bound is always less than ϵ for $m \leq n$. □

A.1.2 Adapting to Unknown Upper Bound on Covariance

As stated, the stability-based algorithms in [DKKLMS17; DK19] assume that the inliers are drawn from a distribution with unknown bounded covariance $\Sigma \preceq \sigma^2 I$, where the

parameter $\sigma > 0$ is known. Here we note that essentially the same algorithms work even if the parameter $\sigma > 0$ is unknown. For this, we establish the following simple modification of standard results, see, e.g., [DK19].

Theorem A.1.3. *Let $T \subset \mathbb{R}^d$ be an ϵ-corrupted version of a set S, where S is $(C\epsilon, \delta)$-stable with respect to μ_S and σ^2, where $C > 0$ is a sufficiently large constant. There exists a polynomial time algorithm that given T and ϵ (but not σ or δ) returns a vector $\widehat{\mu}$ so that $\|\mu_S - \widehat{\mu}\|_2 = O(\sigma\delta)$.*

Proof. The algorithm is very similar to the algorithm from [DK19] except for the stopping condition. We define a weight function $w : T \to \mathbb{R}_{\geq 0}$ initialized so that $w(x) = 1/|T|$ for all $x \in T$. We iteratively do the following:

- Compute $\mu(w) = \frac{1}{\|w\|_1} \sum_{x \in T} w(x) x$.
- Compute $\Sigma(w) = \frac{1}{\|w\|_1} \sum_{x \in T} w(x)(x - \mu(w))(x - \mu(w))^\top$.
- Compute an approximate largest eigenvector v of $\Sigma(w)$.
- Define $g(x)$ for $x \in T$ as $g(x) = |v \cdot (x - \mu(w))|^2$.
- Find the largest t so that $\sum_{x \in T : g(x) \geq t} w(x) \geq \epsilon$.
- Define $f(x) = \begin{cases} g(x) & \text{if } g(x) \geq t \\ 0 & \text{otherwise} \end{cases}$.
- Let m be the largest value of $f(x)$ for any $x \in T$ with $w(x) \neq 0$.
- Set $w(x)$ to $w(x)(1 - f(x)/m)$ for all $x \in T$.

We then repeat this loop unless $\|w\|_1 < 1 - 2\epsilon$, in which case we return $\mu(w)$.

Note that if S is (ϵ, δ)-stable with respect to μ_S and σ^2, then S/σ is (ϵ, δ) with respect to μ_S/σ and 1. We note that if σ was known, the weighted universal filter algorithm of [DK19] could be applied to T/σ in order to learn μ_S/σ to error $O(\delta)$. Multiplying the result by σ would yield an approximation to μ_S with error $O(\sigma\delta)$. We note that this algorithm is equivalent to the one provided above, except that we would stop the loop as soon as $\Sigma(w) \leq \sigma(1 + O(\delta^2/\epsilon))$ rather than waiting until $\|w\|_1 \leq 1 - 2\epsilon$.

However, we note that by the analysis in [DK19] of this algorithm, that at each iteration until it stops, $\sum_{x \in S} w(x)$ decreases by less than $\sum_{x \in T \setminus S} w(x)$ does. Since the latter cannot decrease by more than ϵ, this means that the algorithm of [DK19] would stop before ours does. Our algorithm then continues to remove an additional $O(\epsilon)$ mass

from the weight function w (but only this much since f has support on points of mass only a bit more than ϵ). It is easy to see that these extra removals do not increase $\Sigma(w)$ by more than a factor of $1 + O(\epsilon)$. This means that when our algorithm terminates $\Sigma(w)/\sigma \preceq I + O(\delta^2/\epsilon)$. Thus, by the weighted version of Lemma 2.4 of [DK19], we have that

$$\|\mu_S - \mu(w)\|_2 = \sigma\|\mu_S/\sigma - \mu(w)/\sigma\|_2 \leq \sigma O(\delta + \sqrt{\epsilon(\delta^2/\epsilon)}) = O(\sigma\delta) \ .$$

This completes the proof. □

A.2 Tools from Concentration and Truncation

Organization In Section A.2.1, we state the concentration results that we will use repeatedly in the following sections. Section A.2.2 contains some well-known results regarding the properties of the truncated distribution.

A.2.1 Concentration Results

We first state Talagrand's concentration inequality for bounded empirical processes.

Theorem A.2.1 ([BLM13, Theorem 12.5]). *Let $Y_1, \ldots, Y_n$ be independent identically distributed random vectors. Assume that $\mathbb{E}\, Y_{i,s} = 0$, and that $Y_{i,s} \leq L$ for all $s \in \mathcal{T}$. Define*

$$Z = \sup_{s\in\mathcal{T}} \sum_{i=1}^{n} Y_{i,s}, \qquad \sigma^2 = \sup_{s\in\mathcal{T}} \sum_{i=1}^{n} \mathbb{E}\, Y_{i,s}^2.$$

Then, with probability at least $1 - \exp(-t)$, we have that

$$Z = O(\mathbb{E}\, Z + \sigma\sqrt{t} + Lt). \tag{A.1}$$

See [BLM13, Exercise 12.15] for explicit constants.

We will also repeatedly use the following version of Matrix Bernstein inequality [Tro15; Min17].

Theorem A.2.2 ([Tro15, Corollary 7.3.2]). *Let $S_1, \ldots, S_n$ be n independent symmetric matrices such that $\mathbb{E}\, S_i = 0$ and $\|S_i\|_{\mathrm{op}} \leq L$ a.s. for each index i. Let $Z = \sum_{i=1}^n S_i$ and let V be any PSD matrix such that $\sum_{i=1}^n \mathbb{E}\, S_i S_i^\top \preceq V$. Let $\nu = \|V\|_{\mathrm{op}}$ and $r = \mathrm{r}(V)$. Then, we have that*

$$\mathbb{E}\, \|Z\|_{\mathrm{op}} = O(\sqrt{\nu \log r} + L \log r). \tag{A.2}$$

In particular, if $S_i = \xi_i x_i x_i^\top$, where ξ_i is a Rademacher random variable, and x_i is sampled independently from a distribution with zero mean, covariance Σ, and bounded support $\sqrt{L}$, i.e., $\|x_i\|_2 \le \sqrt{L}$ almost surely. Then $\mathbb{E}\,\|Z\|_{\mathrm{op}} = O(\sqrt{nL\|\Sigma\|_{\mathrm{op}} \log \mathrm{r}(\Sigma)} + L \log \mathrm{r}(\Sigma))$.

A.2.2 Properties under Truncation

We state some basic results regarding truncation of a distribution in this subsection. These results are well-known in literature and are included here for completeness (see, e.g., [DKKLMS17; LRV16]).

Proposition A.2.3 (Shift in mean by truncation)**.** *Let X be sampled from a distribution with mean 0 and covariance $\Sigma \preceq I$. For a $t \ge 0$, let $g(\cdot)$ be defined as*

$$g(x) = \begin{cases} x, & \text{if } x \in [-t, t], \\ t, & \text{if } x > t, \\ -t, & \text{if } x < -t. \end{cases}$$

If $t \ge C\epsilon^{-\frac{1}{2}}$, then for all $v \in \mathcal{S}^{d-1}$, $|\,\mathbb{E}\, g(x^\top v)| \le C^{-1}\sqrt{\epsilon}$.

Proof. Let $Z = x^\top v$. By Markov's inequality,

$$\mathbb{P}(Z \ge t) \le \mathbb{P}(Z^2 \ge C^2\epsilon^{-1}) \le \frac{1}{C^2\epsilon^{-1}} = C^{-2}\epsilon.$$

We get that

$$|\,\mathbb{E}\, g(Z)| = |\,\mathbb{E}\, Z - g(Z)| \le \mathbb{E}\,|Z - g(Z)| \le \mathbb{E}\,|Z|\, \mathbb{I}_{|Z| \ge t} \le \sqrt{\epsilon} C^{-1}. \tag{A.3}$$

□

Proposition A.2.4 (Shift in mean by truncation under higher moments)**.** *Let X be sampled from a distribution with mean 0 and covariance $(1 - \sigma_k^2 \epsilon^{1-\frac{2}{k}})I \preceq \Sigma \preceq I$. Moreover, assume that the distribution has bounded moments, i.e., for a $k \ge 4$:*

$$\forall v \in \mathcal{S}^{d-1}, \quad (\mathbb{E}(v^\top X)^k)^{\frac{1}{k}} \le \sigma_k. \tag{A.4}$$

Note that $\sigma_2 \le 1$. Let $T_k = \sigma_k \epsilon^{-\frac{1}{k}}$. Then

1. *For all $M \in \mathcal{M}$, $\mathbb{E}(x^\top M x)^{\frac{k}{2}} \le \sigma_k^k$.*

2. *For all* $M \in \mathcal{M}$ *and* $t \geq CT_k^2$, $\mathbb{E}\, x^\top Mx\, \mathbb{I}_{x^\top Mx \geq t} \leq \sigma_k^2 C^{\frac{2}{k}-1}\epsilon^{1-\frac{2}{k}}$.

3. *Let* $f(\cdot)$ *be defined as* $f(x) = \min(x, t)$. *For a* $t \geq CT_k^2$, $|\,\mathbb{E}\, f(x^\top Mx) - 1| \leq \sigma_k^2\epsilon^{1-\frac{2}{k}}(1 + C^{1-\frac{k}{2}})$.

4. *Let* $t \geq CT_k$. *For all* $v \in \mathcal{S}^{d-1}$, $|\,\mathbb{E}\, x^\top v\, \mathbb{I}_{|x^\top v| \leq t}\,| \leq \sigma_k\epsilon^{1-\frac{1}{k}}C^{1-k}$.

5. *Let* $g(\cdot)$ *be defined as* $g(x) = \mathit{sign}(x)\min(|x|, t)$. *For* $t \geq CT_k$ *and all* $v \in \mathcal{S}^{d-1}$, $|\,\mathbb{E}\, g(x^\top v)| \leq \sigma_k C^{1-k}\epsilon^{1-\frac{1}{k}}$.

6. $\mathbb{E}\,\|X\|_2^k \leq d^{\frac{k}{2}}\sigma_k^k$.

7. $\mathbb{P}(\|X\|_2 \geq \sigma_k\sqrt{d}\epsilon^{-1/k}) \leq \epsilon$.

Proof. We prove each statement in turn.

1. We use the spectral decomposition of M, to write $M = U^\top \Delta U$, where U is a rotation matrix, Δ is a non-negative diagonal matrix with diagonal entries λ_i and trace 1. Observe that if the random variable X satisfies Equation (A.4), then the random variable $Z := UX$ also satisfies Equation (A.4).

   We use the aforementioned observation and apply Jensen's inequality to get:

$$\mathbb{E}(x^\top Mx)^{\frac{k}{2}} = \mathbb{E}(Z^\top \Delta Z)^{\frac{k}{2}} = \mathbb{E}(\sum_{i=1}^d \lambda_i z_i^2)^{\frac{k}{2}} \leq \sum_{i=1}^d \lambda_i\, \mathbb{E}\, z_i^k \leq \sum_{i=1} \lambda_i \sigma_k^k \leq \sigma_k^k.$$

2. Let $Z = x^\top Mx$. From the first part, we have that $\frac{k}{2}$-th moment of Z is bounded by σ_k^2. By Markov's inequality, we get that

$$\mathbb{P}\{Z \geq t\} \leq \mathbb{P}\left\{Z \geq CT_k^2\right\} \leq \mathbb{P}\left\{Z \geq C\frac{\sigma_k^2}{\epsilon^{\frac{2}{k}}}\right\} \leq \frac{\epsilon}{C^{\frac{k}{2}}\sigma_k^k}\,\mathbb{E}\, Z^{\frac{k}{2}} \leq \frac{\epsilon}{C^{\frac{k}{2}}}.$$

   We can now apply Hölder's inequality to get

$$\mathbb{E}\left[Z \underset{Z \geq CT_k^2}{\mathbb{I}}\right] \leq \sigma_k^2 C^{\frac{2}{k}-1}\epsilon^{1-\frac{2}{k}}.$$

3. As above, let $Z = x^\top Mx$. It follows that $f(x) \leq x$. Therefore, we get that

$$\mathbb{E}\, f(x^\top Mx) \leq \mathbb{E}\, x^\top Mx \leq 1.$$

For the lower bound, we get that

$$\begin{aligned}\mathbb{E}\, f(x^\top Mx) &\geq \mathbb{E}\, x^\top Mx \mathop{\mathbb{I}}_{x^\top Mx \leq CT_k^2} = \mathbb{E}\, x^\top Mx 1 - \mathbb{E}\, x^\top Mx \mathop{\mathbb{I}}_{x^\top Mx > CT_k^2} \\ &\geq 1 - \sigma_k^2 \epsilon^{1-\frac{2}{k}} - \sigma_k^2 \epsilon^{1-\frac{2}{k}} C^{1-\frac{k}{2}}.\end{aligned}$$

4. Let $Z = x^\top v$. We note that

$$\mathbb{P}(Z \geq t) \geq \mathbb{P}(Z \geq CT_k) \leq \mathbb{P}(Z^k \geq C^k T_k^k) \leq \frac{\sigma_k^k}{\sigma_k^k \epsilon^{-1} C^k} \leq C^{-k}\epsilon.$$

We now bound the deviation in mean by truncation:

$$\begin{aligned}\mathbb{E}\, Z = \mathbb{E}\, Z \mathop{\mathbb{I}}_{|Z|\leq t} + \mathbb{E}\, Z \mathop{\mathbb{I}}_{|Z|>t} &= 0 \\ \implies |\,\mathbb{E}\, Z \mathop{\mathbb{I}}_{|Z|\leq t}| &= |\,\mathbb{E}\, Z \mathop{\mathbb{I}}_{Z>t}| \\ &\leq (\mathbb{E}\, Z^k)^{\frac{1}{k}} (\mathbb{P}\{Z > t\})^{1-\frac{1}{k}} \\ &= \sigma_k C^{1-k} \epsilon^{1-\frac{1}{k}}.\end{aligned}$$

5. Let $Z = x^\top v$. We get that

$$|\,\mathbb{E}\, g(Z)| = |\,\mathbb{E}\, Z - g(Z)| \leq \mathbb{E}\, |Z - g(Z)| \leq \mathbb{E}\, |Z| \mathop{\mathbb{I}}_{|Z|\geq CT_k} \leq \sigma_k \epsilon^{1-\frac{1}{k}} C^{1-k}.$$

6. It follows by taking $M = \frac{1}{d} I$ in the first part.

7. This follows by Markov's inequality and the previous part.

□

Lemma A.2.5. *Let P be a distribution with mean μ and covariance I. Let $X \sim P$. For $k > 2$, let its k-th central moment be bounded as*

$$\textit{for all } v \in \mathcal{S}^{d-1}: \quad (\mathbb{E}\, |v^\top X|^k)^{\frac{1}{k}} \leq \sigma_k.$$

For $\epsilon \leq 0.5$, let E be the event

$$E = \{\|X - \mu\|_2 \leq T\},$$

where T is such that $\mathbb{P}(E) \geq 1 - \epsilon$. Let Z be the random variable $X|E$, that is X conditioned on $X \in E$. Then, we have that

1. $\|\mu - \mathbb{E}\, Z\|_2 \leq \frac{1}{1-\epsilon}\sigma_k \epsilon^{1-\frac{1}{k}} \leq 2\sigma_k \epsilon^{1-\frac{1}{k}}$.
2. $(1 - 3\sigma_k^2 \epsilon^{1-\frac{2}{k}})I \preceq \mathrm{Cov}(Z) \preceq \frac{1}{1-\epsilon} I$.

Proof. We prove each statement in turn.

1. Let Q be the distribution of Z. We will assume that $\mathbb{P}(E^c) > 0$, otherwise the results hold trivially. Let R be the distribution of X conditioned on $X \in E^c$ and let $Y \sim R$. Note that P can be written as the convex combination of Q and R.

$$P = (\mathbb{P}(E))Q + (1 - \mathbb{P}(E))R. \tag{A.5}$$

   Using this decomposition, we can calculate the shift in mean along any direction $v \in \mathcal{S}^{d-1}$:

$$\begin{aligned}
\mathbb{P}(E) v^\top \mathbb{E}\, Z + (1 - \mathbb{P}\, E)\, \mathbb{E}\, v^\top Y = v^\top \mathbb{E}\, X = \mu \\
\implies v^\top (\mathbb{E}\, Z - \mu) &= \frac{1}{\mathbb{P}(E)} \mathbb{E}\left[-v^\top (X - \mu) \underset{X \notin E}{\mathbb{I}} \right] \\
&\leq \frac{1}{\mathbb{P}(E)} (\mathbb{E}\, |v^\top (X - \mu)|^k)^{\frac{1}{k}} (\mathbb{P}(E^c))^{1-\frac{1}{k}} \\
&\leq \frac{1}{\mathbb{P}(E)} \sigma_k \epsilon^{1-\frac{1}{k}},
\end{aligned}$$

   where the first inequality uses Hölder's inequality. Therefore, $\|\mathbb{E}\, Z - \mu\|_2 \leq \sigma_k \epsilon^{1-1/k}/(1-\epsilon)$.

2. We will follow the notations from the previous part. Note that for all $v \in \mathcal{S}^{d-1}$, the mean minimizes the quadratic loss. In particular,

$$\mathbb{E}(v^\top (Z - \mathbb{E}\, Z))^2 \leq \mathbb{E}(v^\top (Z - \mu))^2.$$

   Using (A.5), we have that

$$\mathbb{E}(v^\top (Z - \mu))^2 \leq \frac{1}{\mathbb{P}(E)} \mathbb{E}(v^\top (X - \mu))^2 \leq \frac{1}{1-\epsilon}.$$

Therefore, we obtain the following upper bound:

$$\mathbb{E}\, v^\top (Z - \mathbb{E}\, Z)^2 \leq \mathbb{E}(v^\top (Z - \mu))^2 \leq \frac{1}{1-\epsilon}.$$

We now turn our attention to lower bound. We first note that

$$\begin{aligned}(1 - \mathbb{P}(E))\, \mathbb{E}(v^\top (Y - \mu))^2 &= \mathbb{E}(v^\top (X - \mu))^2\, \mathbb{I}\{X \in E^c\} \\ &\leq (\mathbb{E}(v^\top (X - \mu))^k)^{\frac{2}{k}} (\mathbb{P}(E))^{1-\frac{2}{k}} \leq \sigma_k^2 \epsilon^{1-\frac{2}{k}}.\end{aligned}$$

Using (A.5), we get

$$\begin{aligned}\mathbb{E}(v^\top (Z - \mu))^2 &= \frac{1}{\mathbb{P}(E)} (\mathbb{E}(v^\top (X - \mu))^2 - (1 - \mathbb{P}(E))\, \mathbb{E}(v^\top (Y - \mu))^2) \\ &\geq (1 - (1 - \mathbb{P}(E))\, \mathbb{E}(v^\top (Y - \mu))^2) \geq 1 - \sigma_k^2 \epsilon^{1-\frac{2}{k}}.\end{aligned}$$

We are now ready to bound from below the deviation from mean:

$$\begin{aligned}\mathbb{E}(v^\top (Z - \mathbb{E}\, Z))^2 &= \mathbb{E}(v^\top (Z - \mu))^2 - (v^\top (\mathbb{E}\, Z - \mu))^2 \\ &\geq 1 - \sigma_k^2 \epsilon^{1-\frac{2}{k}} - (\frac{\sigma_k \epsilon^{1-\frac{1}{k}}}{1-\epsilon})^2 \\ &\geq 1 - \sigma_k^2 \epsilon^{1-\frac{2}{k}} - \frac{\sigma_k^2 \epsilon^{1-\frac{2}{k}}}{1-\epsilon} \geq 1 - 3\sigma_k^2 \epsilon^{1-\frac{2}{k}}.\end{aligned}$$

□

A.3 Bounds on the Number of Points with Large Projections

Organization This section contains the proofs of Lemma 3.2.3 and Lemma 3.4.2 from the main paper. In Section A.3.1, we prove the results controlling the number of outliers uniformly along all directions $v \in \mathcal{S}^{d-1}$. We then generalize these results to projections along PSD matrices in Section A.3.2.

A.3.1 Linear Projections

We state Lemma 1 from Lugosi and Mendelson [LM21b]. We will use this result for distributions with bounded covariance.

Lemma A.3.1 ([LM21b, Lemma 1])**.** *Let $x_1, \dots, x_n$ be n i.i.d. points from a distribution with mean zero and covariance $\Sigma \preceq I$. Let Q_2 be defined as follows:*

$$Q_2 = \frac{256}{\epsilon}\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \frac{16}{\sqrt{\epsilon}}.$$

Then, for a constant $c > 0$, with probability at least $1 - \exp(-c\epsilon n)$,

$$\sup_{v \in \mathcal{S}^{d-1}} \left|\left\{i : |v^\top x_i| \geq Q_2\right\}\right| \leq 0.25\epsilon n \ .$$

We state the following straightforward generalization of Lemma A.3.1 for distributions with bounded central moments. We give the proof for completeness.

Lemma A.3.2. *Let $x_1, \dots, x_n$ be n i.i.d. points from a distribution with mean zero and covariance $\Sigma \preceq I$. Further assume that for all $v \in \mathcal{S}^{d-1}$:*

$$(\mathbb{E}(v^\top X)^k)^{\frac{1}{k}} \leq \sigma_k. \tag{A.6}$$

Let Q_k be defined as follows:

$$Q_k = \Theta\left(\frac{1}{\epsilon}\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \sigma_k \epsilon^{-\frac{1}{k}}\right).$$

Then, there exists a $c > 0$, such that with probability at least $1 - \exp(-cn\epsilon)$,

$$\sup_{v \in \mathcal{S}^{d-1}} \left|\left\{i : |x_i^\top v| \geq Q_k\right\}\right| = O(n\epsilon). \tag{A.7}$$

Proof. We follow the same strategy as in Lugosi and Mendelson [LM21b]. We first set Q_k as follows:

$$Q_k = C\left(\frac{1}{\epsilon}\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \sigma_k \epsilon^{-\frac{1}{k}}\right),$$

for a large enough constant C to be determined later. Consider the function $\chi : \mathbb{R} \to \mathbb{R}$

defined by

$$\chi(x) = \begin{cases} 0, & \text{if } x \le \frac{Q_k}{2}, \\ \frac{2x}{Q_k} - 1, & \text{if } x \in \left[\frac{Q_k}{2}, Q_k\right], \\ 1, & \text{if } x \ge Q_k. \end{cases} \tag{A.8}$$

Therefore, $\mathbb{I}_{x^\top v \ge Q_k} \le \chi(x_i^\top v) \le \mathbb{I}_{x^\top v \ge Q_k/2}$ and note that $\chi(\cdot)$ is a $\frac{2}{Q_k}$ Lipschitz. We first bound the number of points violating the upper tail bounds. The random quantity of interest is the following:

$$Z = \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n \mathbb{I}_{x_i^\top v \ge Q_k} \quad . \tag{A.9}$$

We first calculate its expectation using the symmetrization principle [LT91; BLM13]. We have that

$$\begin{aligned} \mathbb{E}\, Z &= \mathbb{E} \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n \mathbb{I}_{x_i^\top v \ge Q_k} \\ &\le \mathbb{E} \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n \chi(x_i^\top v) \\ &\le \mathbb{E} \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n (\chi(x_i^\top v) - \mathbb{E}\,\chi(x_i^\top v)) + \sup_{v \in \mathcal{S}^{d-1}} \mathbb{E} \sum_{i=1}^n \chi(x_i^\top v) \\ &\le 2\,\mathbb{E} \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n \epsilon_i \chi(x_i^\top v) + \sup_{v \in \mathcal{S}^{d-1}} \mathbb{E} \sum_{i=1}^n \chi(x_i^\top v). \end{aligned} \tag{A.10}$$

We bound the second term in Eq. (A.10) by

$$\mathbb{E} \sum_{i=1}^n \chi(x_i^\top v) \le \mathbb{E} \sum_{i=1}^n \mathbb{I}_{x_i^\top v \ge Q_k/2} = n\,\mathbb{P}(x_i^\top v \ge Q_k/2) \le n\,\mathbb{P}(x_i^\top v \ge C\sigma_k \epsilon^{-\frac{1}{k}}) = O(n\epsilon),$$

by applying Markov inequality and choosing a large enough constant C for Q_k. For the first term in Eq. (A.10), we upper bound $\chi(\cdot)$ using contraction principle for Rademacher averages and independence of x_i:

$$\begin{aligned} \mathbb{E} \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n \epsilon_i \chi(x_i^\top v) &\le \frac{2}{Q_k}\, \mathbb{E} \sup_{v \in \mathcal{S}^{d-1}} \sum_{i=1}^n \epsilon_i x_i^\top v \\ &= \frac{2}{Q_k}\, \mathbb{E} \left\| \sum_i \epsilon_i x_i \right\|_2 \le n \frac{2}{Q_k} \sqrt{n \operatorname{tr}(\Sigma)} = O(n\epsilon), \end{aligned}$$

where we use the covariance bound on x_i and a large enough constant for $Q_k \geq (C/\epsilon)\sqrt{\mathrm{tr}(\Sigma)/n}$. Therefore, we get that $\mathbb{E}\, Z = O(n\epsilon)$. We can upper bound the wimpy variance, i.e., the quantity σ^2 in Theorem A.2.1, by $O(\epsilon n)$. By Talagrand's concentration A.2.1, we get that probability $1 - \exp(-cn\epsilon)$,

$$Z = O(n\epsilon + \sqrt{n\sigma}\sqrt{cn\epsilon}\sqrt{n\gamma} + cn\epsilon) = O(n\epsilon). \tag{A.11}$$

□

A.3.2 Matrix Projections

We will now use the results from the previous section to prove Lemma 3.2.3 and Lemma 3.4.2. The proof follows the ideas from [DL22b, Proposition 1].

Lemma A.3.3. *Suppose that the event $\mathcal{E}_1$ holds, where $\mathcal{E}_1$ is the following*

$$\mathcal{E}_1 := \left\{ \sup_{v \in \mathcal{S}^{d-1}} |\{i : |x_i^\top v| \geq Q_0\}| \leq 0.25\epsilon n \right\}.$$

Let $Q = 8Q_0$. Then the event $\mathcal{E}$ also holds, where $\mathcal{E}$ is defined as follows:

$$\mathcal{E} := \left\{ \sup_{M \in \mathcal{M}} |\{i : x_i^\top M x_i \geq Q^2\}| \leq \epsilon n \right\}.$$

Proof. We follow the same proof strategy as Depersin and Lecué [DL22b]. We reproduce the proof here for completeness.

Suppose that $\mathcal{E}_1$ holds but the desired event $\mathcal{E}$ does not hold. Let M be such that $|\{i : x_i^\top M x_i \geq Q^2\}| > \epsilon n$. Let G be the Gaussian vector in $\mathbb{R}^d$ independent of $x_1, \ldots, x_n$ with distribution $\mathcal{N}(0, M)$. We will work conditionally on $x_1, \ldots, x_n$ in the remaining of the proof. By Gaussian concentration (see, e.g., [BLM13]) we have that with probability at least 0.999: $\|G\|_2 \leq 5$. Let Z be the following random variable

$$Z = \sum_{i=1}^{n} \mathbb{I}_{|x_i^\top G|^2 \geq 25Q_0^2, \|G\|_2 \leq 5}.$$

Under $\mathcal{E}_1$, we have that $Z \leq 0.25\epsilon n$, implying $\mathbb{E}[Z] \leq 0.25\epsilon n$. Moreover, we have that

$x_i^\top G \sim \mathcal{N}(0, x_i^\top M x_i)$. For i such that $x_i^\top M x_i \geq Q^2$, we have that

$$\begin{aligned}\mathbb{P}(|x_i^\top G|^2 > 25Q_0^2, \|G\|_2 \leq 5) &\geq \mathbb{P}(|x_i^\top G|^2 > 25Q_0^2) - \mathbb{P}(\|G\|_2 > 5) \\ &\geq 2\,\mathbb{P}\left(g \geq \frac{5}{8}\right) - 0.001 > 0.528 - 0.001 > 0.527,\end{aligned}$$

where g is a standard Gaussian random variable. Therefore,

$$\mathbb{E}\, Z = \sum_{i=1}^{n} \mathbb{P}(|x_i^\top G|^2 > 25Q_0^2, \|G\|_2 \leq 5) \geq \epsilon n(0.527),$$

which is a contradiction as $\mathbb{E}[Z] \leq 0.25\epsilon n$. □

We are now ready to prove Lemma 3.2.3 and 3.4.2.

Proof. (Proof of Lemma 3.2.3) The result now follows from Lemma A.3.1, due to Lugosi and Mendelson [LM21b, Lemma 1], and Lemma A.3.3. □

Proof. (Proof of Lemma 3.4.2) The result now follows from Lemma A.3.2, which might require a change of variables, and Lemma A.3.3. □

A.4 Stability for Distributions with Bounded Covariance

Organization Section A.4.1 contains the proof of the sufficient conditions for stability under bounded covariance assumption (Claim 3.2.1). Section A.4.2 contains the arguments for deterministic rounding (Lemma A.4.2).

A.4.1 Sufficient Conditions for Stability

The following claim simplifies the stability condition for the bounded covariance case.

Claim A.4.1 (Claim 3.2.1)**.** *Let S be a set such that $\|\mu_S - \mu\|_2 \leq \sigma\delta$, and $\|\overline{\Sigma}_S - \sigma^2 I\|_{\text{op}} \leq \sigma^2\delta^2/\epsilon$ for some $0 \leq \epsilon \leq \delta$. Let $\epsilon' < 0.5$. Then S is (ϵ', δ') stable with respect to μ and σ^2, where $\delta' = 2\sqrt{\epsilon'} + 2\delta\sqrt{\epsilon'/\epsilon}$.*

Proof. Let $\epsilon' < 0.5$. Without loss of generality, we can assume that $\sigma = 1$. For $S' \subseteq S$: $|S'| \geq (1-\epsilon')|S|$,

$$\frac{1}{|S'|}\sum_{i\in S'}(x_i^\top v)^2 - 1 \leq \frac{1}{|S'|}\sum_{i\in S}(x_i^\top v)^2 - 1 \leq \frac{1}{1-\epsilon'}\left(1 + \frac{\delta^2}{\epsilon}\right) - 1$$

$$= \frac{\frac{\delta^2}{\epsilon} + \epsilon'}{1 - \epsilon'} \leq \frac{1}{\epsilon'} \left(2\epsilon' + 2\delta \sqrt{\frac{\epsilon'}{\epsilon}} \right)^2 \leq \frac{(\delta')^2}{\epsilon'}.$$

As $\delta' \geq \sqrt{\epsilon'}$, the lower bound on eigenvalues of $\overline{\Sigma}_{S'}$ is trivially satisfied. We now bound the deviation in mean. Observe that the uniform distribution on S' can be obtained by conditioning the uniform distribution on S on an event E, such that $\mathbb{P}(E) \geq 1 - \epsilon'$. Using this observation in conjunction with Hölder's inequality gives us that for any v, the shift in mean is at most

$$\left| \frac{1}{|S'|} \sum_{i \in S'} v^\top x_i - \frac{1}{|S|} \sum_{i \in S'} v^\top x_i \right| \leq 2\sqrt{1 + \frac{\delta^2}{\epsilon}} \sqrt{\epsilon'} \leq 2\sqrt{\epsilon'} + 2\delta \sqrt{\frac{\epsilon'}{\epsilon}} \leq \delta'. \tag{A.12}$$

□

A.4.2 Deterministic Rounding of the Weight Function

The next lemma states that it suffices to find a distribution $w \in \Delta_{n,\epsilon}$ for stability.

Lemma A.4.2 (Lemma 3.2.9)**.** *For $\epsilon \leq \frac{1}{3}$, let $w \in \Delta_{n,\epsilon}$ be such that for $\epsilon \leq \delta$, we have*

1. $\|\mu_w - \mu\|_2 \leq \sigma\delta$.
2. $\|\overline{\Sigma}_w - \sigma^2 I\|_{\text{op}} \leq \sigma^2\delta^2/\epsilon$.

Then there exists a subset $S_1 \subseteq S$ such that

1. $|S_1| \geq (1 - 2\epsilon)|S|$.
2. *S_1 is (ϵ', δ') stable with respect to μ and σ^2, where $\delta' = O(\delta + \sqrt{\epsilon} + \sqrt{\epsilon'})$.*

Proof. Without loss of generality, we will assume that $\sigma^2 = 1$. We will use Claim A.4.1 to prove this result by first showing that there exists a subset $S' \subseteq [n]$ with bounded covariance and good sample mean.

Without loss of generality, we will assume that ϵn is an integer and $\mu = 0$. We will also assume that $\frac{1}{(1-\epsilon)n} \geq w_1 \geq w_2 \geq \cdots \geq w_n \geq 0$. For any $k \in [n]$, we have that

$$1 = \sum_i w_i \leq \frac{n - k}{(1 - \epsilon)n} + k w_k,$$

which implies that

$$w_k \geq \frac{1}{k}\frac{(1-\epsilon)n-(n-k)}{(1-\epsilon)n} = \frac{k-\epsilon n}{(1-\epsilon)nk}.$$

Setting $k = 2\epsilon n$, we have that

$$w_k \geq \frac{2\epsilon n}{2n(1-\epsilon)} = \frac{1}{2(1-\epsilon)n}. \tag{A.13}$$

We now have a lower bound on w_i for all $i \leq (1-2\epsilon)n$. Now let S_1 be the set of the $n-k$ points with the largest w_i. In particular, for each $i \in S_1$, $w_i \geq \frac{1}{2(1-\epsilon)n}$. We have that,

$$\begin{aligned}
\sum_{i\in S_1}\frac{1}{|S_1|}(x_i^\top v)^2 &= \sum_{i\in S_1}\frac{1}{(1-2\epsilon)n}(x_i^\top v)^2 \\
&\leq \sum_{i\in S_1}\frac{1}{(1-2\epsilon)}2w_i(1-\epsilon)(x_i^\top v)^2 && \text{(Using Eq. (A.13))} \\
&\leq \frac{2(1-\epsilon)}{(1-2\epsilon)}\sum_{i\in S} w_i(x_i^\top v)^2 \\
&\leq 9\left(1+\frac{\delta^2}{\epsilon}\right). && \text{(A.14)}
\end{aligned}$$

Let the uniform distribution on S_1 be $u^{(1)}$ and the uniform distribution on S be u. We now calculate the total variation distance between w and $u^{(1)}$.

$$d_{\mathrm{TV}}(w, u^{(1)}) \leq d_{\mathrm{TV}}(w, u) + d_{\mathrm{TV}}(u, u^{(1)}) \leq \epsilon + 2\epsilon = 3\epsilon.$$

Therefore, there exist distributions $p^{(1)}, p^{(2)}, p^{(3)}$ such that

$$w = (1-3\epsilon)p^{(1)} + 3\epsilon p^{(2)}, \qquad u^{(1)} = (1-3\epsilon)p^{(1)} + 3\epsilon p^{(3)}. \tag{A.15}$$

This decomposition follows from an alternate characterization of total variation distance(see, e.g., [Tsy09, Lemma 2.1]). The decomposition in (A.15) implies that for any unit vector v, the following holds:

$$3\epsilon\sum_i p_i^{(2)}(x_i^\top v)^2 \leq \sum_i w_i(x_i^\top v)^2 \leq 1+\frac{\delta^2}{\epsilon} \tag{A.16}$$

$$3\epsilon\sum_i p_i^{(3)}(x_i^\top v)^2 \leq \sum_i u_i^{(1)}(x_i^\top v)^2 \leq 9\left(1+\frac{\delta^2}{\epsilon}\right). \tag{A.17}$$

Let v be an arbitrary unit vector. Then using (A.15), we get the following:

$$\begin{aligned}
\left|\sum_{i=1}^{n}(1-3\epsilon)p_i^{(1)}x_i^\top v\right| &\le \left|\sum_{i=1}^{n} w_i x_i^\top v\right| + \left|3\epsilon\sum_i p_i^{(2)}x_i^\top v\right| \\
&\le \delta + 3\epsilon\sqrt{\sum_{i=1}^{n} p_i^{(2)}(x_i^\top v)^2} && \text{(Stability of } w\text{)} \\
&= \delta + \sqrt{3\epsilon}\sqrt{3\epsilon\sum_{i=1}^{n} p_i^{(2)}(x_i^\top v)^2} \\
&\le \delta + \sqrt{3\epsilon}\sqrt{\left(1+\frac{\delta^2}{\epsilon}\right)} && \text{(Using (A.16))} \\
&\le \delta + \sqrt{3\epsilon} + \sqrt{3}\delta \le 3\delta + 2\sqrt{\epsilon}.
\end{aligned}$$

We will now combine this result with (A.16) and (A.15). Starting with the decomposition in (A.15), we have the following:

$$\begin{aligned}
\left|\sum_{i=1}^{n} u_i^{(1)}x_i^\top v\right| &\le \left|\sum_{i=1}^{n}(1-3\epsilon)p_i^{(1)}x_i^\top v\right| + \left|\sum_{i=1}^{n} 3\epsilon p_i^{(3)}x_i^\top v\right| \\
&\le 3\delta + 2\sqrt{\epsilon} + \sqrt{3\epsilon}\sqrt{3\epsilon\sum_i p_i^{(3)}(x_i^\top v)^2} \\
&\le 3\delta + 2\sqrt{\epsilon} + \sqrt{27}\sqrt{\epsilon+\delta^2} && \text{(Using (A.17))} \\
&\le 10\delta + 10\sqrt{\epsilon}. && \text{(A.18)}
\end{aligned}$$

Therefore using Equations (A.14) and (A.18), we have a set S_1 that satisfies the conditions in Claim A.4.1 with $\delta'' = 10\delta + 10\sqrt{\epsilon}$. Using Claim A.4.1, we get that S_1 is (ϵ', δ') stable. □

A.5 Stability for Distributions with Bounded Central Moments

Organization In this section, we provide the detailed arguments regarding the proof of Theorem 3.1.8 that were omitted from the main text. We start with a simplified stability condition in Section A.5.1. Section A.5.2 contains the argument for rounding a good distribution $w \in \Delta_{n,\epsilon}$ to a subset. Section A.5.3 contains the arguments for controlling the second moment matrix from above and below respectively. Sections A.5.3 and A.5.4 contain the arguments for concentration of the second moment matrix and

mean respectively.

A.5.1 Sufficient Conditions for Stability

We will prove the existence of a stable set with high probability using the following claim. This is analogous to Claim A.4.1 in the bounded covariance setting, but we also need a lower bound on the minimum eigenvalue of $\overline{\Sigma}_{S'}$ for all large subsets S'.

Claim 3.4.1. *Let $0 \leq \epsilon \leq \delta$ and $\epsilon \leq 0.5$. A set S is $(\epsilon, O(\delta))$ stable with respect to μ and $\sigma^2 = 1$, if it satisfies the following for all unit vectors v.*

1. $\|\mu_S - \mu\|_2 \leq \delta$.
2. $v^\top \overline{\Sigma}_S v \leq 1 + \delta^2/\epsilon$.
3. *For all subsets $S' \subseteq S : |S'| \geq (1-\epsilon)|S|$, $v^\top \overline{\Sigma}_{S'} v \geq (1 - \delta^2/\epsilon)$.*

Proof. Without loss of generality, we will assume that $\mu = 0$. We first show the second condition in the definition of stability. Let S' be any proper subset of S, such that $|S'| \geq (1-\epsilon)|S|$. Note that the minimum eigenvalue of S' is lower-bounded by the assumption:

$$v^\top \Sigma_{S'} v = \frac{1}{|S \setminus S_\epsilon|} \sum_{i \in S \setminus S_\epsilon} (v^\top x)^2 \geq 1 - \frac{\delta^2}{\epsilon}. \tag{A.19}$$

We now look at the largest eigenvalue of S':

$$\begin{aligned} v^\top \Sigma_S v - 1 = \frac{1}{|S'|} \sum_{i \in S'} (v^\top x)^2 - 1 &\leq \frac{|S|}{|S'|} \frac{1}{|S|} \sum_{i \in S} (v^\top x)^2 - 1 \\ &\leq \frac{1}{1-\epsilon} \left(1 + \frac{\delta^2}{\epsilon}\right) - 1 \leq \frac{1}{1-\epsilon} \left(\frac{\delta^2}{\epsilon} + \epsilon\right) \leq \frac{2\delta^2}{\epsilon} + 2\epsilon \leq 4\frac{\delta^2}{\epsilon}. \end{aligned}$$

We now need to show that the mean of S' is also good. In order to do that, we first control the deviation due to a small set $S \setminus S'$.

$$\begin{aligned} \frac{1}{|S|} \sum_{i \in S \setminus S'} (v^\top x_i)^2 &= \frac{1}{|S|} \sum_{i \in S} (v^\top x_i)^2 - \frac{1}{|S|} \left(\sum_{i \in S'} \left(v^\top x_i\right)^2 \right) \\ &\leq \left(1 + \frac{\delta^2}{\epsilon}\right) - \frac{|S'|}{|S|} \left(1 - \frac{\delta^2}{\epsilon}\right) \\ &\leq \left(1 + \frac{\delta^2}{\epsilon}\right) - (1-\epsilon) \left(1 - \frac{\delta^2}{\epsilon}\right) \leq \frac{2\delta^2}{\epsilon} + \epsilon. \end{aligned} \tag{A.20}$$

We first break the deviation in mean into two terms, and control each individually:

$$\left|\frac{1}{|S'|}\sum_{i\in S'}(v^\top x_i)\right| = \frac{|S|}{|S'|}\left|\frac{1}{|S|}\sum_{i\in S\setminus S_\epsilon}(v^\top x_i)\right| \le \frac{|S|}{|S'|}\left|\frac{1}{|S|}\sum_{i\in S}(v^\top x_i)\right| + \frac{|S|}{|S'|}\left|\frac{1}{|S|}\sum_{i\in S\setminus S'}(v^\top x_i)\right|.$$

We can upper bound the first term by $\|\mu_S\|/(1-\epsilon) \le \delta/(1-\epsilon)$. We bound the second term using the Cauchy-Schwarz inequality and Eq. (A.20):

$$\begin{aligned}\frac{|S|}{|S'|}\left|\frac{1}{|S|}\sum_{i\in S\setminus S'}(v^\top x_i)\right| &\le \frac{|S\setminus S'|}{|S'|}\cdot\left|\frac{1}{|S\setminus S'|}\sum_{i\in S\setminus S'}(v^\top x_i)\right| \\ &\le \frac{|S\setminus S'|}{|S'|}\cdot\sqrt{\frac{1}{|S\setminus S'|}\sum_{i\in S\setminus S'}(v^\top x_i)^2} \\ &= \frac{\sqrt{|S\setminus S'||S|}}{|S'|}\cdot\sqrt{\frac{1}{|S|}\sum_{i\in S\setminus S'}(v^\top x_i)^2} \le \frac{\sqrt{\epsilon}}{1-\epsilon}\sqrt{\frac{2\delta^2}{\epsilon}+\epsilon}.\end{aligned}$$

Overall, we get that

$$|v^\top \mu_{S'}| \le \frac{1}{1-\epsilon}(\delta + \sqrt{2}\delta + \epsilon) \le 5\delta + 2\epsilon \le 7\delta.$$

□

A.5.2 Randomized Rounding of Weight Function

In this section, we show how to recover a subset from a $w \in \Delta_{n,\epsilon}$. Unlike the deterministic rounding in Section A.4.2, we do a randomized rounding in Lemma A.5.1 to get a better dependence on ϵ. For the second condition ($\delta^2 = O(\epsilon)$) in Lemma A.5.1 to hold, it is necessary that $n = \Omega(d)$. If $n = O(d)$, it is not a problem because, in this regime, the bounded covariance assumption already leads to optimal error.

Lemma A.5.1. *Let $k \ge 4$. Let $w \in \Delta_{n,\epsilon}$, for $\epsilon \le \frac{1}{3}$, be a distribution on the set of points S such that*

1. $\|\mu_w - \mu\|_2 \le \delta$.
2. $\|\overline{\Sigma}_w\|_{\mathrm{op}} - 1 \le \frac{\delta^2}{\epsilon} \le r_1$, *for some* $r_1 > 1$.
3. *Let $C \ge 4$. For all subsets S': $|S'| \ge (1-C\epsilon)n$ and $v \in \mathcal{S}^{d-1}$: $v^\top \overline{\Sigma}_{S'} v \ge 1 - \delta^2/(C\epsilon)$.*
4. $w_i > 0$ *implies that* $\|x_i\|_2 \le r_2\sigma_k\sqrt{d}\gamma^{-1/k}$ *for some* $r_2 \ge 1$.

Then, there exists a subset $S_1 \subseteq [n]$ such that

1. $|S_1| \geq (1-2\epsilon)n$.
2. S_1 *is* (ϵ', δ') *stable, where*

$$\epsilon' = (C-2)\epsilon, \qquad \delta' = O\Big(\delta + \sqrt{\frac{r_1 d \log d}{n}} + r_2 \sigma_k \epsilon^{\frac{1}{2}-\frac{1}{k}} \sqrt{\frac{d \log d}{n}} + r_2 r_1 \sigma_k \epsilon^{1-\frac{1}{k}}\Big). \tag{A.21}$$

Proof. We will use Claim 3.4.1 to prove this result. Without loss of generality, let $\mu = 0$. Therefore, it suffices to find a subset such that both the mean and the largest eigenvalue are controlled. Let $Y_i \sim \text{Bernoulli}(w_i(1-\epsilon)n)$. We have that $\sum_{i=1}^n \mathbb{E}\, Y_i = (1-\epsilon)n$. Let S_1 be the (random) set:

$$S_1 = \{i : Y_i = 1\}. \tag{A.22}$$

By a Chernoff bound, we have that for some constant $c' > 0$,

$$\mathbb{P}(|S_1| \geq (1-2\epsilon)n) \leq \exp(-c'n\epsilon). \tag{A.23}$$

Let E be the event $E = \{|S_1| \geq (1-2\epsilon)n\}$. We now bound the mean of the set S_1. Consider the following random variable Z:

$$Z = \sum_i (Y_i - (1-\epsilon)w_i n)x_i. \tag{A.24}$$

The random variable Z satisfies $\mathbb{E}\, Z = 0$. Moreover, its covariance can be bounded using the assumption as follows:

$$\begin{aligned} v^\top \Sigma_Z v &= \sum_{i=1}^n w_i(1-\epsilon)n(1-w_i(1-\epsilon)n)(v^\top x_i)^2 \\ &\leq (1-\epsilon)n \sum_{i=1}^n w_i (x_i^\top v)^2 \leq (1-\epsilon)n\left(1 + \frac{\delta^2}{\epsilon}\right) \preceq 2r_1 n. \end{aligned}$$

Therefore, with probability at least 0.8, we have that

$$\begin{aligned} &\|Z\|_2 \leq 10\sqrt{r_1 n d} \\ \implies &\|\sum Y_i x_i\|_2 \leq (1-\epsilon)n \|\sum_i w_i X_i\|_2 + 10\sqrt{r_1 n d}. \end{aligned}$$

Let E_2 be the event that $E_2 = \{\|\sum Y_i x_i\|_2 \leq (1-\epsilon)n\sigma + 10\sqrt{r_1 nd}\}$. This implies that on the event $E \cap E_1$,

$$\|\mu_{S_1}\|_2 \leq \frac{1-\epsilon}{1-2\epsilon}\delta + 10\frac{c_5}{1-2\epsilon}\sqrt{\frac{d}{n}} \leq 2\delta + 30\sqrt{\frac{r_1 d}{n}}. \tag{A.25}$$

We now focus our attention on upper bounding the eigenvalue. Define the symmetric random matrix, Z_i as $Z_i := Y_i x_i x_i^\top - w_i(1-\epsilon)nx_i x_i^\top$. We have that $\mathbb{E}\, Z_i = 0$ and $\|Z_i\|_{\text{op}} \leq r_2^2 d\sigma_k \epsilon^{1-\frac{1}{k}}$ almost surely. We now bound the matrix variance statistic (used in Theorem A.2.2):

$$\begin{aligned}
\nu(Z) &= \left\|\sum_{i=1}^n w_i(1-\epsilon)n(1-w_i(1-\epsilon)n)\|x_i\|^2 x_i x_i^\top\right\|_{\text{op}} \\
&\leq \left\|\sum_{i=1}^n w_i(1-\epsilon)n\frac{r_2^2\sigma_k^2 d}{\epsilon^{\frac{2}{k}}} x_i x_i^\top\right\|_{\text{op}} \\
&\leq (1-\epsilon)\frac{r_2^2\sigma_k^2 nd}{\epsilon^{\frac{2}{k}}}\left\|\sum_{i=1}^n w_i x_i x_i^\top\right\|_{\text{op}} \\
&\leq (1-\epsilon)\frac{r_2^2\sigma_k^2 nd}{\epsilon^{\frac{2}{k}}}\left\|\overline{\Sigma}_w\right\|_{\text{op}} \leq 2\frac{r_1 r_2^2\sigma_k^2 nd}{\epsilon^{\frac{2}{k}}}.
\end{aligned}$$

By the Matrix-Chernoff concentration result (Theorem A.2.2), we get that with probability at least 0.8, we have that

$$\left\|\sum_{i=1}^n Y_i x_i x_i^\top - w_i(1-\epsilon)nx_i x_i^\top\right\|_{\text{op}} = O\left(\sqrt{\frac{r_1 r_2^2\sigma_k^2 nd\log d}{\epsilon^{\frac{2}{k}}}} + \frac{r_2^2\sigma_k^2 d\log d}{\epsilon^{\frac{2}{k}}}\right). \tag{A.26}$$

Let E_3 be the event above, which happens with probability at least 0.8. Under the event $E \cap E_3$, we get that

$$\begin{aligned}
v^\top\overline{\Sigma}_{S_1} v &\leq \frac{1-\epsilon}{1-2\epsilon}w_i(x_i^\top v)^2 + \frac{1}{1-2\epsilon}O\left(\sqrt{\frac{r_1 r_2^2\sigma_k^2 d\log d}{n\epsilon^{\frac{2}{k}}}} + \frac{r_2^2\sigma_k^2 d\log d}{n\epsilon^{\frac{2}{k}}}\right) \\
&\leq \frac{1-\epsilon}{1-2\epsilon}\left(1+\frac{\delta^2}{\epsilon}\right) + O\left(\sqrt{\frac{r_1 r_2^2\sigma_k^2 d\log d}{n\epsilon^{\frac{2}{k}}}} + \frac{r_2^2\sigma_k^2 d\log d}{n\epsilon^{\frac{2}{k}}}\right) \\
&\leq 1 + \frac{1}{\epsilon}O\left(\epsilon^2 + \delta^2 + \sqrt{\frac{d\log d}{n}}r_1 r_2\sigma_k\epsilon^{1-\frac{1}{k}} + r_2^2\sigma_k^2\epsilon^{1-\frac{2}{k}}\frac{d\log d}{n}\right) \\
&\leq 1 + \frac{1}{\epsilon}\left(O\left(\delta + r_1 r_2\sigma_k\epsilon^{1-\frac{1}{k}} + \sqrt{\frac{d\log d}{n}} + r_2\sigma_k\epsilon^{\frac{1}{2}-\frac{1}{k}}\sqrt{\frac{d\log d}{n}}\right)\right)^2.
\end{aligned} \tag{A.27}$$

Let $\epsilon' = (C-2)\epsilon$. Note that if $|S_1| \geq (1-2\epsilon)|S|$, then $|S'| \geq (1-\epsilon')|S_1|$ implies that $|S'| \geq (1-C\epsilon)|S|$, which leads to a lower bound on the minimum eigenvalue. This follows from the following elementary calculations:

$$\frac{|S'|}{|S|} \geq (1-2\epsilon)\frac{|S_1|}{|S|} \geq (1-2\epsilon)(1-(C-2)\epsilon) \geq 1 - C\epsilon. \tag{A.28}$$

Using Equations (A.23), (A.25) and (A.27), we get that there exists a subset S_1 such that for all $v \in \mathcal{S}^{d-1}$ and $\delta' = O(\delta + \sqrt{r_1 d \log d/n} + r_1 r_2 \sigma_k \epsilon^{1/2-1/k}\sqrt{d\log d/n} + r_1 r_2 \sigma_k \epsilon^{1-\frac{1}{k}})$:

1. $|S_1| \geq (1-2\epsilon)n \geq (1-\epsilon')n$.
2. $\|\mu_{S_1}\|_2 \leq \delta'$.
3. $v^\top \overline{\Sigma}_{S_1} v \leq 1 + \frac{\delta'^2}{\epsilon'}$.
4. For all subsets $S' \subseteq S_1 : |S'| \geq (1-\epsilon')|S_1|$, $v^\top \overline{\Sigma}_{S'} v \geq 1 - \frac{\delta'^2}{\epsilon'}$.

We now invoke Claim 3.4.1 to conclude that S' is $(\epsilon', 7\delta')$-stable. □

A.5.3 Upper Bound on the Second Moment Matrix

Lemma A.5.2. *Consider the conditions in Lemma 3.4.3. Then, with probability* $1-\tau$, $R'/n \leq \delta^2/\epsilon$ *, where* $\delta = O(\sqrt{d\log d/n} + \sigma_k \epsilon^{1-\frac{1}{k}} + \sigma_4\sqrt{\log(1/\tau)/n})$.

Proof. (Proof of Lemma A.5.2) We first calculate the wimpy variance required for Theorem A.2.1,

$$\begin{aligned}\sigma^2 &= \sup_{M\in M} \sum_{i=1}^n \mathbf{Var}(f(x_i^\top M x_i)) \leq \sup_{M\in M}\sum_{i=1}^n \mathbb{E}\, f(x_i^\top M x_i)^2 \\ &\leq n \sup_{M\in M} \mathbb{E}(x_i^\top M x_i)^4 \leq n\sigma_4^4.\end{aligned}$$

We use symmetrization, contraction, and matrix concentration (Theorem A.2.2) to bound $\mathbb{E}\, R'$ as follows:

$$\begin{aligned}\mathbb{E}\, R' &= \mathbb{E} \sup_{M\in\mathcal{M}} \sum_{i=1}^n f(x_i^\top M x_i) - \mathbb{E}\, f(x_i^\top M x_i) \leq 2\,\mathbb{E} \sup_{M\in\mathcal{M}} \sum_{i=1}^n \epsilon_i f(x_i^\top M x_i) \\ &\leq 2\,\mathbb{E} \sup_{M\in\mathcal{M}} \sum_{i=1}^n \epsilon_i x_i^\top M x_i = 2\,\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i x_i x_i^\top\right\|_{\text{op}} \\ &= O\left(\sqrt{\frac{\sigma_k^2 nd\log(d)}{\epsilon^{\frac{2}{k}}}} + \frac{\sigma_k^2 d\log d}{\epsilon^{\frac{2}{k}}}\right),\end{aligned}$$

where we use Theorem A.2.2, with $\nu = O(\sigma_k^2 n d \epsilon^{-\frac{2}{k}})$ and $L = O(\sigma_k^2 d \epsilon^{-\frac{2}{k}})$.

Note that $Q_k = O(\sigma_k \epsilon^{-\frac{1}{k}} + (1/\epsilon)\sqrt{d/n}$. As R' is bounded by Q_k^2, we can apply Theorem A.2.1 to get that with probability at least $1-\tau$, R'/n is bounded as follows:

$$\begin{aligned}
\frac{R'}{n} &= O\left(\sqrt{\frac{\sigma_k^2 d \log d}{n\epsilon^{\frac{2}{k}}}} + \frac{\sigma_k^2 d \log d}{n\epsilon^{\frac{2}{k}}} + \sigma_4^2\sqrt{\frac{\log(\frac{1}{\tau})}{n}} + \frac{\sigma_k^2}{\epsilon^{\frac{2}{k}}}\frac{\log(\frac{1}{\tau})}{n} + \frac{1}{\epsilon^2}\frac{d}{n}\frac{\log(\frac{1}{\tau})}{n}\right) \\
&= \frac{1}{\epsilon} O\left(\sqrt{\frac{d\log d}{n}}\sigma_k\epsilon^{1-\frac{1}{k}} + \frac{d\log d}{n}\sigma_k^2\epsilon^{1-\frac{2}{k}} + \sigma_4\epsilon\sigma_4\sqrt{\frac{\log(\frac{1}{\tau})}{n}} + \sigma_k^2\epsilon\epsilon^{1-\frac{2}{k}} + \frac{d}{n}\right) \\
&\qquad\qquad (\text{ Using } \tfrac{\log(\frac{1}{\tau})}{n} = O(\epsilon)) \\
&= \frac{1}{\epsilon} O\left(\left(\sqrt{\frac{d\log d}{n}} + \sigma_k\epsilon^{1-\frac{1}{k}} + \sigma_k\epsilon^{\frac{1}{2}-\frac{1}{k}}\sqrt{\frac{d\log d}{n}} + \sigma_4\epsilon + \sigma_4\sqrt{\frac{\log(\frac{1}{\tau})}{n}}\right)\right)^2 \\
&= \frac{1}{\epsilon} O\left(\left(\sqrt{\frac{d\log d}{n}} + \sigma_k\epsilon^{1-\frac{1}{k}} + \sigma_4\sqrt{\frac{\log(\frac{1}{\tau})}{n}}\right)\right)^2 \\
&\qquad\qquad (\text{Using } \sigma_4\epsilon \le \sigma_k\epsilon^{1-\frac{1}{k}} \text{ and } \sigma_k\epsilon^{\frac{1}{2}-\frac{1}{k}} = O(1)) \\
&\le \frac{\delta^2}{\epsilon},
\end{aligned}$$

where we use the parameter regime stated in Lemma 3.4.3. □

A.5.4 Controlling the Mean

Lemma A.5.3. *Consider the setting in Lemma 3.4.5. Then, with probability,* $1 - \tau - \exp(-n\epsilon)$,

$$\frac{R'}{n} = O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \sigma_k\epsilon^{1-\frac{1}{k}}\right).$$

Proof. We first calculate the wimpy variance required for Theorem A.2.1,

$$\sigma^2 = \sup_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n \mathbf{Var}(g(x_i^\top v)) \le \sup_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n \mathbb{E}\, g(v^\top x_i)^2 \le \sup_{v\in\mathcal{S}^{d-1}} n\,\mathbb{E}(v^\top x_i)^2 \le n.$$

We use symmetrization, contraction of Rademacher averages to bound $\mathbb{E}\, R'$.

$$\mathbb{E}\, R' = \mathbb{E} \sup_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n g(v^\top x_i) - \mathbb{E}\, g(v^\top x_i)$$

$$\leq 2\,\mathbb{E}\sup_{v\in\mathcal{S}^{d-1}}\sum_{i=1}^{n}\epsilon_i g(v^\top x_i)$$
$$\leq 2\,\mathbb{E}\sup_{v\in\mathcal{S}^{d-1}}\sum_{i=1}^{n}\epsilon_i v^\top x_i = 2\,\mathbb{E}\left\|\sum_{i=1}^{n}\epsilon_i x_i\right\|_2 \leq 2\sqrt{\frac{d}{n}}.$$

By applying Theorem A.2.1, we get that with probability at least $1-\tau$,

$$\begin{aligned}\frac{R'}{n} &= O\left(\frac{\mathbb{E}\,R'}{n} + \sqrt{\frac{\log(1/\tau)}{n}} + Q_k\frac{\log(1/\tau)}{n}\right)\\ &= O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \sigma_k\epsilon^{-\frac{1}{k}}\frac{\log(\frac{1}{\tau})}{n} + \frac{1}{\epsilon}\sqrt{\frac{d}{n}}\frac{\log(1/\tau)}{n}\right)\\ &= O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \sigma_k\epsilon^{1-\frac{1}{k}}\right),\end{aligned}$$

where the last inequality uses the assumption that $\frac{\log(1/\tau)}{n} = O(\epsilon)$. □

B APPENDIX TO CHAPTER 4

Additional Notation For a vector $x \in \mathbb{R}^d$ and $H \subset [d]$, we denote v_H to denote the vector that is equal to v on $i \in H$, and zero otherwise. For a real-valued random variable X and $m \in \mathbb{N}$, we use $\|X\|_{L_m}$ to denote $(\mathbb{E}\,|X|^m)^{1/m}$.

B.1 Miscellaneous Lemmas and Facts

B.1.1 Finding a Stable Subset from a Stable Weighted Subset

For a set S on n points, we define $\Delta_{n,\epsilon}$ as the set of weights $w \in \mathbb{R}^n$ such that $w_i \in [0, 1/((1-\epsilon)n]$ for all $i \in [n]$ and $\sum_i w_i = 1$. For a fixed vector $\mu \in \mathbb{R}^d$ that will be clear from context, a set of n points $S = \{x_1, \ldots, x_n\}$, and weights $w \in \Delta_{n,\epsilon}$ over S, we use $\overline{\Sigma}_w$ to denote $\sum_i w_i(x_i - \mu)(x_i - \mu)^\top$.

The goal of this section is to show Proposition B.1.1, which states that if we have a weight w over S such that $\overline{\Sigma}_w$ (with respect to some vector μ) has bounded $\mathcal{X}_k$ norm proportional to σ^2 for some $\sigma > 0$, then there must exist some large subset $S' \subseteq S$ that is stable with respect to μ and σ.

Proposition B.1.1. *Let S be a set of n points in $\mathbb{R}^d$. Let $\Delta_{n,\epsilon}$ be the set of weights defined above, and define the notation $\overline{\Sigma}_w = \sum_{x_i \in S} w_i(x_i - \mu)(x_i - \mu)^\top$ for some given vector $\mu \in \mathbb{R}^d$. Suppose that there exists a $w \in \Delta_{n,\epsilon}$ such that $\|\overline{\Sigma}_w\|_{\mathcal{X}_k} \le B\sigma^2$ for some vector μ. Then there exists a subset $S' \subseteq S$ such that $(i)|S'| \ge (1-2\epsilon)n$ and (ii) S' is (ϵ, δ, k)-stable with respect to μ and σ, where $\delta = O(\sqrt{B} + 1)$.*

Observe that $\|\overline{\Sigma_w}\|_{\mathcal{X}_k} \le B\sigma^2$ implies $\|\overline{\Sigma_w} - \sigma^2 I\|_{\mathcal{X}_k} \le (B+1)\sigma^2$ by the triangle inequality. In order to show Proposition B.1.1, we show Lemma B.1.2, which is a weakening of Proposition B.1.1 where we additionally assume that $\mu_w = \sum_i w_i x_i$ is close to μ, where μ is the vector we use to define $\overline{\Sigma_w}$ as well as the vector that we want to find a large sample subset S' to be stable with respect to. To use Lemma B.1.2, we additionally show Proposition B.1.4, which states that $\|\overline{\Sigma_w}\|_{\mathcal{X}_k} \le B\sigma^2$ is enough to imply that μ_w is close to μ. We combine Lemma B.1.2 and Proposition B.1.4 to prove Proposition B.1.1 at the end of Appendix B.1.1.

Lemma B.1.2. *Suppose, for some $\epsilon \le \frac{1}{3}$ and for some $\delta \ge \sqrt{\epsilon}$, there exist a $w \in \Delta_{n,\epsilon}$ over a set of n samples $S = \{x_1, \ldots, x_n\}$, a $\mu \in \mathbb{R}^d$ and a $\sigma > 0$ such that*

- $\|\mu_w - \mu\|_{2,k} \leq \delta\sigma$,
- $\|\sum_{i\in[n]} w_i(x_i - \mu)(x_i - \mu)^\top - \sigma^2 I\|_{\mathcal{X}_k} \leq \sigma^2\frac{\delta^2}{\epsilon}$.

Then, there exists a subset $S' \subseteq S$ of samples such that

- $|S'| \geq (1 - 2\epsilon)|S|$,
- *S' is (ϵ, δ', k)-stable with respect to μ and σ, where $\delta' = O(\delta + \sqrt{\epsilon})$.*

Proof. Without loss of generality, we will only handle the $\sigma = 1$ case to simplify notation.

The main step is to show the existence of a large subset S' whose mean is within $10\delta + 10\sqrt{\epsilon}$ of μ and whose variance is at most $9(1 + \delta^2/\epsilon)$. In fact, we can simply choose S' to be the subset whose weights w_i are the largest.

Without loss of generality, assume $\mu = 0$ and that ϵn is an integer. We also order the samples in decreasing order of weight in w, namely, $1/((1 - \epsilon)n) \geq w_1 \geq w_2 \geq \ldots \geq w_n$.

First, we will lower bound each w_i. We have that for each $k \in [n]$,

$$1 = \sum_i w_i \leq \frac{k}{(1 - \epsilon)n} + (n - k)w_k,$$

which upon rearranging implies that

$$w_k \geq \frac{(1 - \epsilon)n - k}{(1 - \epsilon)n(n - k)}.$$

In particular, for $k = (1 - 2\epsilon)n$, we have

$$w_{(1-2\epsilon)n} \geq \frac{1}{2(1 - \epsilon)n}.$$

Letting S' to be the $(1 - 2\epsilon)n$ points with the largest weight, we have that for all $i \in S'$, $w_i \geq \frac{1}{2(1-\epsilon)n}$. We will use this to now bound the $\mathcal{X}_k$ norm of $\Sigma_{S'} = \frac{1}{|S'|}\sum_{i\in S'} x_i x_i^\top$. Consider an arbitrary $M \in \mathcal{X}_k$, we have

$$\begin{aligned}
\sum_{i\in S'} \frac{1}{|S'|}\langle x_i x_i^\top, M\rangle &= \sum_{i\in S'} \frac{1}{(1 - 2\epsilon)n}\langle x_i x_i^\top, M\rangle \\
&\leq \sum_{i\in S'} \frac{2(1 - \epsilon)}{1 - 2\epsilon} w_i\langle x_i x_i^\top, M\rangle \\
&\leq \sum_{i\in S} \frac{2(1 - \epsilon)}{1 - 2\epsilon} w_i\langle x_i x_i^\top, M\rangle
\end{aligned}$$

$$\leq 9\left(1+\frac{\delta^2}{\epsilon}\right).$$

Since $\delta \geq \sqrt{\epsilon}$, this in turn implies the (rather loose in constants) inequality that $\|\Sigma_{S'} - I\|_{\mathcal{X}_k} \leq 20(\delta^2/\epsilon)$.

Next, we show that the mean $\mu_{S'}$ of S' is $10\delta+10\sqrt{\epsilon}$-close to $\mu = 0$. This will essentially follow from 1) the uniform distribution $U_{S'}$ over S' is close in total variation distance to w and 2) the contribution of the tail to the mean of a bounded-covariance distribution is small.

For 1), using the notation that U_S is the uniform distribution over S (analogous to the S' notation just before), it is immediate that by the triangle inequality,

$$d_{\mathrm{TV}}(w, U_{S'}) \leq d_{\mathrm{TV}}(w, U_S) + d_{\mathrm{TV}}(U_S, U_{S'}) \leq \epsilon + 2\epsilon = 3\epsilon.$$

A standard consequence is that there exists distributions $p^{(1)}$, $p^{(2)}$ and $p^{(3)}$ such that

$$w = (1-3\epsilon)p^{(1)} + 3\epsilon p^{(2)} \quad \text{and} \quad U_{S'} = (1-3\epsilon)p^{(1)} + 3\epsilon p^{(3)}.$$

Intuitively, treating $p^{(2)}$ and $p^{(3)}$ as the "tails", we will bound their contributions to the mean under the boundedness of the covariance of w and $U_{S'}$.

Take any k-sparse unit vector direction $v \in \mathcal{U}_k$, we can bound the following variances in the direction of v:

$$3\epsilon \sum_i p_i^{(2)} \langle x_i, v\rangle^2 \leq \sum_i w_i \langle x_i, v\rangle^2 \leq 1 + \frac{\delta^2}{\epsilon},$$

$$3\epsilon \sum_i p_i^{(3)} \langle x_i, v\rangle^2 \leq \sum_i U_{S',i} \langle x_i, v\rangle^2 \leq 9\left(1+\frac{\delta^2}{\epsilon}\right),$$

where we used the fact that $vv^\top$ is in $\mathcal{X}_k$ for a k-sparse unit vector v.

By Jensen's inequality, we can then conclude that

$$\left|3\epsilon \sum_i p_i^{(2)} \langle x_i, v\rangle\right| \leq \sqrt{3\epsilon}\sqrt{3\epsilon \sum_i p_i^{(2)} \langle x_i, v\rangle^2} \leq \sqrt{3\epsilon}\sqrt{1+\frac{\delta^2}{\epsilon}} \leq \sqrt{3}(\sqrt{\epsilon}+\delta),$$

$$\left|3\epsilon \sum_i p_i^{(3)} \langle x_i, v\rangle\right| \leq \sqrt{3\epsilon}\sqrt{3\epsilon \sum_i p_i^{(3)} \langle x_i, v\rangle^2} \leq 3\sqrt{3\epsilon}\sqrt{1+\frac{\delta^2}{\epsilon}} \leq 3\sqrt{3}(\sqrt{\epsilon}+\delta).$$

Finally, since $U_{S'} = w - 3\epsilon p^{(2)} + 3\epsilon p^{(3)}$, by the triangle inequality, we have

$$\begin{aligned}
|\langle \mu_{S'} - \mu, v\rangle| &= \left|\sum_i U_{S',i}\langle x_i, v\rangle\right| \\
&\le \left|\sum_i w_i\langle x_i, v\rangle\right| + \left|3\epsilon\sum_i p_i^{(2)}\langle x_i, v\rangle\right| + \left|3\epsilon\sum_i p_i^{(3)}\langle x_i, v\rangle\right| \\
&\le \delta + \sqrt{3}(\sqrt{\epsilon} + \delta) + 3\sqrt{3}(\sqrt{\epsilon} + \delta) \\
&\le 10\delta + 10\sqrt{\epsilon},
\end{aligned}$$

where the second inequality uses the above bounds as well as the assumption that $\|\mu_w - \mu\|_{2,k} \le \delta$.

Now that we have shown that $\mu_{S'}$ is close to μ in $\ell_{2,k}$ norm and $\Sigma_{S'}$ is small in the $\mathcal{X}_k$ norm, we will use the following lemma (Lemma B.1.3) to show that the set S' is $(\epsilon, O(\delta + \sqrt{\epsilon})$-stable with respect to μ. □

Lemma B.1.3 (Bounded Mean and Covariance implies $O(\sqrt{\epsilon})$ stability)**.** *Let $\mu \in \mathbb{R}^d$ and let S' be a set of samples such that $\|\mu_{S'} - \mu\|_{2,k} \le \delta$ and $\left\|\frac{1}{|S'|}\sum_{x\in S'}(x-\mu)(x-\mu)^\top - I\right\|_{\mathcal{X}_k} \le \frac{\delta^2}{\epsilon}$ for some $0 \le \epsilon \le \delta$ and $\epsilon \le 0.5$. Then S' is (ϵ, δ', k)-stable with respect to μ where $\delta' = O(\delta + \sqrt{\epsilon})$ and $\delta' \ge \sqrt{\epsilon}$.*

Proof. Consider an arbitrary large subset $S'' \subseteq S'$ where $|S''| \ge (1-\epsilon)|S'|$. Without loss of generality, take $\mu = 0$. Then, for an arbitrary $M \in \mathcal{X}_k$,

$$\langle \overline{\Sigma}_{S''} - I, M\rangle = \frac{1}{S''}\sum_{i\in S''}\langle x_i x_i^\top, M\rangle - 1,$$

which is trivially at least $-1 \ge -(\delta'^2)/\epsilon$ for $\delta' \ge \sqrt{\epsilon}$. As for the upper bound, we have

$$\begin{aligned}
\langle \overline{\Sigma} - I, M\rangle &= \frac{1}{S''}\sum_{i\in S''}\langle x_i x_i^\top, M\rangle - 1 \\
&\le \left(\frac{1}{S''}\sum_{i\in S'}\langle x_i x_i^\top, M\rangle\right) - 1 \\
&\le \frac{1}{1-\epsilon}\left(1 + \frac{\delta^2}{\epsilon}\right) - 1 \\
&= \frac{\frac{\delta^2}{\epsilon} + \epsilon}{1-\epsilon} \\
&\le \frac{2}{\epsilon}(\delta^2 + \epsilon^2)
\end{aligned}$$

$$\leq \frac{\delta'^2}{\epsilon} ,$$

for some $\delta' = \Theta(\delta + \sqrt{\epsilon})$.

We now bound the error in the mean of S'' in $\ell_{2,k}$ norm. First, observe that, for an arbitrary k-sparse unit vector v,

$$\begin{aligned}
\left| \frac{1}{|S'|} \sum_{i \in S' \setminus S''} \langle x_i, v \rangle \right| &= \left| \frac{1}{|S'|} \sum_{i \in S'} \mathbb{I}[x_i \in S' \setminus S''] \langle x_i, v \rangle \right| \\
&\leq \frac{1}{|S'|} \sum_{i \in S'} |\mathbb{I}[x_i \in S' \setminus S''] \langle x_i, v \rangle| \\
&\leq \sqrt{\epsilon} \sqrt{\frac{1}{|S'|} \sum_{i \in S'} \langle x_i, v \rangle^2} \\
&\leq \sqrt{\epsilon} \sqrt{1 + \frac{\delta^2}{\epsilon}} \\
&= \sqrt{\epsilon + \delta^2} ,
\end{aligned}$$

where the second inequality is an application of H$\acute{\text{o}}$lder's inequality, and the third inequality uses the fact that for a unit k-sparse vector v, $vv^\top$ is in $\mathcal{X}_k$.

Thus, again for an arbitrary k-sparse unit vector v,

$$\begin{aligned}
|\langle \mu_{S''}, v \rangle| &= \left| \frac{1}{|S''|} \sum_{i \in S''} \langle x_i, v \rangle \right| \\
&\leq \frac{1}{1 - \epsilon} \left| \frac{1}{|S'|} \sum_{i \in S''} \langle x_i, v \rangle \right| \\
&\leq 2 \left(\left| \frac{1}{|S'|} \sum_{i \in S'} \langle x_i, v \rangle \right| + \left| \frac{1}{|S'|} \sum_{i \in S' \setminus S''} \langle x_i, v \rangle \right| \right) \\
&\leq 2(\delta + \sqrt{\epsilon + \delta^2}) = O(\delta + \sqrt{\epsilon}) = \delta'.
\end{aligned}$$

□

Proposition B.1.4 (Bounded Covariance and Stability)**.** *Let $\mu \in \mathbb{R}^d$ and let S be a set of n samples. Let $w \in \Delta_{n,\epsilon}$ over the set of samples S such that $\| \sum_i w_i (x_i - \mu)(x_i - \mu)^\top \|_{\chi_k} \leq r$ for some $r > 0$. Then $\|\mu_w - \mu\|_{2,k} \leq \sqrt{r}$.*

Proof. For every k-sparse unit vector v, $vv^\top$ is in $\mathcal{X}_k$, and thus for every sparse unit vector v, we have that $\sum_i w_i \langle x_i - \mu, v \rangle^2 \leq r$. Applying Cauchy-Schwarz inequality, we get that for

any sparse unit vector v, it follows that $\sum_i w_i \langle x_i - \mu, v\rangle \le \sqrt{\sum_i w_i \langle x_i - \mu, v\rangle^2} \le \sqrt{r}$. □

With Proposition B.1.4 and Lemma B.1.2, we can prove Proposition B.1.1.

Proof of Proposition B.1.1. Without loss of generality, we will assume that $\sigma = 1$. By Proposition B.1.4, we have that $\|\mu_w - \mu\|_{2,k} \le \sqrt{B}$. We thus have a weighting $w \in \Delta_{n,\epsilon}$, where $\|\mu_w - \mu\|_{2,k} \le \delta_0$ and $\|\overline{\Sigma}_w - I\|_{\mathcal{X}_k} \le \delta_0^2/\epsilon$ for $\delta_0 = \sqrt{B} + 1$, where we use triangle inequality on the $\|\cdot\|_{\mathcal{X}_k}$ norm. By Lemma B.1.2, we know that there exists a set S' such that $|S'| \ge (1-2\epsilon)n$ and S' is (ϵ, δ, k)-stable with respect to μ and σ, where $\delta = O(\delta_0 + \sqrt{\epsilon}) = O(\sqrt{\epsilon} + \sqrt{B} + 1) = O(\sqrt{B} + 1)$. □

B.1.2 Median-of-Means Pre-Processing

This section shows Fact 4.2.2, which states that the median-of-means pre-processing technique allows us to reduce to the constant-corruption case.

Fact 4.2.2 (Median-of-Means Pre-Processing)**.** *Suppose there is an efficient algorithm such that, on input $\sigma \in \mathbb{R}_+$ and a 0.01-corrupted set of $n \gg k^2 \log d + \log(1/\tau)$ samples from a distribution D with mean μ and covariance Σ with $\|\Sigma\|_{\mathcal{X}_k} \le \sigma^2$ and $\mathbb{E}_{X\sim D}[(X_j - \mu_j)^4] = O(\sigma^4)$ for each coordinate $j \in [d]$, returns $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_{2,k} \le O(\sigma)$ with probability at least $1 - \tau$.*

Then, there is an efficient algorithm such that, on input $\epsilon \in (0, 0.01]$ and an ϵ-corrupted set of $n \gg (k^2 \log d + \log(1/\tau))/\epsilon$ samples from a distribution with mean μ and covariance Σ, satisfying (i) $\|\Sigma\|_{\mathcal{X}_k} \le 1$ and (ii) $\mathbb{E}_{X\sim D}[(X_j - \mu_j)^4] = O(1)$ for every coordinate $j \in [d]$, returns a mean estimate $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_{2,k} \le O(\sqrt{\epsilon})$ with probability at least $1 - \tau$.

Proof. The new algorithm simply performs median-of-means preprocessing as defined in Section 4.2 before the fact statement, yielding g new samples that are fed into the algorithm that works with constant corruption. The uncorrupted new samples, namely the ones that are the sample mean of groups containing no originally corrupted samples, are distributed i.i.d. according to the distribution D' which has mean μ, and covariance $\Sigma' = (g/n)\Sigma$, with axis-wise fourth moment $\mathbb{E}_{Y\sim D'}[(Y_j - \mu_j)^4]$ being bounded by $C(g^2/n^2)\,\mathbb{E}_{X\sim D}[(X_j - \mu_j)^4]$ for every $j \in [d]$ for some constant $C > 0$, obtained by the following fact:

Fact B.1.5. (*Marcinkiewicz-Zygmund inequality*) *Recall the notation $\|X\|_{L_s}$ for a centered random variable X, defined as $\mathbb{E}[|X|^s]^{1/s}$. Let $W_1, \ldots, W_m, W$ be identical and independent*

centered random variables on $\mathbb{R}$ with a finite $\|W\|_{L_s}$ norm for $s \geq 2$. Then,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} W_i \right\|_{L_s} \leq \frac{3\sqrt{s}}{\sqrt{m}} \|W\|_{L_s}.$$

First note that we give g samples to the original algorithm, and $g = \Omega(\epsilon n) = \Omega(k^2 \log d + \log(1/\tau))$ by definition. Next, we need to check that the *normalized* axis-wise 4th moment of D' is $O(1)$ times the (bound on the) $\mathcal{X}_k$-norm of the covariance matrix, that is, for all $j \in [d]$, it holds that $(\mathbb{E}_{X \sim D'}[(X_j - \mu_j)^4])^{1/4} \leq O(\sigma^4)$ and $\|\Sigma'\|_{\mathcal{X}_k} = O(\sigma^2)$. By the calculations at the end of the previous paragraph and the assumptions in the statement, we note that this is true for $\sigma = O(\sqrt{g/n})$.

Lastly, we check that, by the scale-invariance of the original algorithm that works with constant corruption, the estimation error of the final algorithm is upper bounded by $O(\sigma \|\Sigma\|_{\mathcal{X}_k}) = O(\sqrt{(g/n)\|\Sigma\|_{\mathcal{X}_k}}) = O(\sqrt{g/n}) = O(\sqrt{\epsilon})$ as desired. □

B.1.3 Basic Properties of $\mathcal{X}_k$-Norm

The following is a straightforward bound on the $\mathcal{X}_k$-norm of a matrix based on entry-wise bounds.

Lemma 4.6.2. *Let $A \in \mathbb{R}^{d \times d}$ be a symmetric matrix such that $|A_{i,i}| \leq \eta_1$ for each $i \in [d]$, and $|A_{i,j}| \leq \eta_2$ for each $i \neq j \in [d] \times [d]$. Then $\|A\|_{\mathcal{X}_k} \leq \eta_1 + k\eta_2$.*

Proof. Let $A = B + C$, where B is a diagonal matrix and C is diagonal-free. Then we have the following using triangle inequality: $\|A\|_{\mathcal{X}_k} \leq \|B\|_{\mathcal{X}_k} + \|C\|_{\mathcal{X}_k}$. Thus it suffices to bound each of these terms by 1.

$$\|B\|_{\mathcal{X}_k} \leq \sup_{M : \sum_{i=1}^{d} |M_{i,i}| \leq 1} \langle B, M \rangle = \|B\|_\infty \leq \eta_1,$$

where we use that B is a diagonal matrix with entry at most η_1.

$$\|C\|_{\mathcal{X}_k} \leq \sup_{M : \|M\|_1 \leq k} \langle C, M \rangle = \sup_{M : \|M\|_1 \leq k} \|C\|_\infty \|M\|_1 \leq k\eta_2.$$

□

B.2 Concentration and Truncation

B.2.1 Truncation Can Increase Spectral Norm of Covariance

We show how truncation can increase the spectral norm of covariance from 1 to $\omega(1)$.

Consider the distribution which, with probability $1/(2k)$, returns a vector where each coordinate is independent $-\sqrt{k}$ with probability $2/3$ and $2\sqrt{k}$ with probability $1/3$. Otherwise, with probability $1 - 1/(2\sqrt{k})$, the distribution returns the origin. The mean of the distribution is the origin, and the covariance is I.

Now consider the truncation $h_{0,\sqrt{k}}$, which truncates at distance $\sqrt{k}$ from the origin. Let Y be the resulting random variable. The mean of Y, μ', is thus equal to $(1/2k)(-\sqrt{k}/3, \ldots, -\sqrt{k}/3) = -1/(6\sqrt{k})\mathbf{v}$, where $\mathbf{v}$ is the all ones vector. The norm of μ' is $\Theta(\sqrt{d/k})$. Since the distribution returns the origin with constant probability (asymptotically tending to 1), the variance of Y along the direction of μ', which is $\mathbf{v}/\sqrt{d}$, is at least $\Omega(d/k) = \omega(1)$.

B.2.2 Preserving Moments under Truncation

Lemma 4.3.2 shows that truncation (mostly) preserves the mean, covariance and axis-wise fourth moments of a distribution under axis-wise fourth moment assumptions on the input distribution.

Lemma 4.3.2 (Truncation in ℓ_∞)**.** *Let P be a distribution over $\mathbb{R}^d$ with mean μ_P and covariance Σ_P, with $\|\Sigma\|_{\mathcal{X}_k} \le \sigma^2$ for some $\sigma^2 > 0$. Let $X \sim P$ and assume that for all $j \in [d]$, $\mathbb{E}[(X - \mu_P)_j^4] \le \sigma^4\nu^4$ for some $\nu \ge 1$. Let $b \in \mathbb{R}^d$ be such that $\|b - \mu\|_\infty \le a/2$ and $a := 2\sigma\sqrt{k/\epsilon}$ for some $\epsilon \in (0, 1)$. Define Q to be the distribution of $Y := h_{a,b}(X)$. Let the mean and covariance of Q be μ_Q and Σ_Q respectively. Then the following hold:*

(1) $\|\mu_P - \mu_Q\|_\infty \le \sigma\sqrt{\epsilon/k}$

(2) $\|\mu_P - \mu_Q\|_{2,k} \le \sigma\sqrt{\epsilon}$

(3) $\|\Sigma_P - \Sigma_Q\|_{\mathcal{X}_k} \le 3\sigma^2\epsilon\nu^4$

(4) *For all $i \in [d]$, $\mathbb{E}[(Y - \mu_Q)_i^4] \le 8\nu^4\sigma^4$*

(5) $\|Y - \mu_Q\|_\infty \le 2a = 4\sigma\sqrt{k/\epsilon}$ *almost surely.*

Proof. Let $Y := h_{a,b}(X)$ and denote $\mu := \mu_P$. Fix an $i \in [d]$. Since $|\mu_i - b_i| \le a/2$ and we truncate at radius a, we have the following:

$$|Y_i - \mu_i| \le |X_i - \mu_i|, \quad \text{and} \quad |X_i - Y_i| \le |X_i - \mu_i|. \tag{B.1}$$

Let $\mathcal{E}_i$ be the event that $Y_i \neq X_i$. We get the following by Markov's inequality and moment bounds:

$$\mathcal{P}(\mathcal{E}_i) = \mathbb{P}(|X_i - b_i| > a) \le \mathbb{P}(|X_i - \mu_i| \ge a/2) \le \min\left(4\frac{\sigma^2}{a^2}, 16\frac{\sigma^4\nu^4}{a^4}\right) = \min\left(\frac{\epsilon}{k}, \frac{\epsilon^2\nu^4}{k^2}\right). \tag{B.2}$$

1. We can verify the following relation using Equation (B.1):

$$|Y_i - X_i| \le \mathop{\mathbb{I}}_{\mathcal{E}_i} \cdot (|Y_i - X_i|) \le \mathop{\mathbb{I}}_{\mathcal{E}_i} \cdot (|X_i - \mu_i|). \tag{B.3}$$

   Applying Cauchy-Schwarz on the above inequality gives the desired conclusion:

$$|\,\mathbb{E}[Y_i] - \mu_i| = |\,\mathbb{E}[Y_i - X_i]| \le \mathbb{E}\left[\mathop{\mathbb{I}}_{\mathcal{E}} \cdot (|X_i - \mu_i|)\right] \le \sqrt{\mathbb{P}(\mathcal{E})}\sqrt{\mathbb{E}\left[|X_i - \mu_i|^2\right]} \le \sigma\sqrt{\frac{\epsilon}{k}},$$

   where we use that variance of X_i is at most σ^2 and use Equation (B.2).

2. This follows directly from above.

3. By Lemma 4.6.2, it suffices to show that $\|\Sigma_Q - \Sigma_P\|_\infty \le 3\sigma^2\epsilon\nu^4/k$. Using triangle inequality, we obtain the following:

$$\begin{aligned}\|\Sigma_P - \Sigma_Q\|_\infty &= \Big\| \mathbb{E}[(X - \mu_P)(X - \mu_P)^\top] - \mathbb{E}[(Y - \mu_P)(Y - \mu_P)^\top] \\ &\qquad + (\mu_Q - \mu_P)(\mu_Q - \mu_P)^\top \Big\|_\infty \\ &\le \Big\| \mathbb{E}[(X - \mu_P)(X - \mu_P)^\top] - \mathbb{E}[(Y - \mu_P)(Y - \mu_P)^\top] \Big\|_\infty \\ &\qquad + \Big\| (\mu_Q - \mu_P)(\mu_Q - \mu_P)^\top \Big\|_\infty.\end{aligned}$$

   By the first part above, we have that $\|(\mu_Q - \mu_P)(\mu_Q - \mu_P)^\top\|_\infty \le \sigma^2\epsilon/k \le \sigma^2\nu^4\epsilon/k$, where we use that $\nu \ge 1$. We will thus focus on the first term. Without loss of generality, we will assume that $\mu_P = 0$ for the remainder of this proof. Thus for

any $i, j \in [d]$, we thus need to upper bound $\mathbb{E}[|X_i X_j - Y_i Y_j|]$.

$$\begin{aligned}
\mathbb{E}[|X_i X_j - Y_i Y_j|] &\leq \mathbb{E}[|X_i||X_j - Y_j|] + \mathbb{E}[|Y_j||X_i - Y_i|] \\
&\leq \mathbb{E}[|X_i||X_j| \cdot \mathbb{I}_{\mathcal{E}_j}] + \mathbb{E}[|X_i||X_j| \cdot \mathbb{I}_{\mathcal{E}_i}] \qquad \text{(Using Equation (B.3))} \\
&\leq \sqrt{\mathbb{E}[|X_i X_j|^2]} \left(\sqrt{\mathbb{P}(\mathcal{E}_i)} + \sqrt{\mathbb{P}(\mathcal{E}_j)} \right) \\
&\leq (\mathbb{E}[X_i^4])^{1/4} (\mathbb{E}[X_j^4])^{1/4} \left(\sqrt{\mathbb{P}(\mathcal{E}_i)} + \sqrt{\mathbb{P}(\mathcal{E}_j)} \right) \\
&= \sigma^2 \nu^2 \left(2 \frac{\epsilon \nu^2}{k} \right) \\
&= \frac{2\sigma^2 \nu^4 \epsilon}{k}.
\end{aligned}$$

Combining the above with Lemma 4.6.2, we get that the $\|\Sigma_P - \Sigma_Q\|_{\mathcal{X}_k} \leq 3\sigma^2 \epsilon \nu^4$.

4. Fix an $i \in [d]$. We use the triangle inequality and Equation (B.3) to get the following:

$$\mathbb{E}[(Y - \mu_Q)_i^4] \leq 4(\mathbb{E}[(Y - \mu_P)_i^4]) + 4\|\mu_P - \mu_Q\|_\infty^4 \leq 4\sigma^4 \nu^4 + 4\sigma^4 \epsilon^2 / k^2 \leq 8\sigma^4 \nu^4,$$

where the last inequality uses that $\nu \geq 1$ and $\epsilon \leq 1$.

5. This follows by definition of the random variable Y, the function $h_{a,b}$, and the parameter a.

□

B.2.3 Standard Concentration Tools

Fact B.2.1 (VC inequality). *Let $\mathcal{F}$ be a family of boolean functions over $\mathcal{X}$ with VC dimension r and let $S = \{x_1, \ldots, x_n\}$ be a set of n i.i.d. data points from a distribution P over $\mathcal{X}$. If $n \gg c(r + \log(1/\tau))/\gamma^2$, then with probability $1 - \tau$, for all $f \in \mathcal{F}$, we have that*

$$\left| \sum_{i=1}^{n} \frac{f(x_i)}{n} - \mathop{\mathbb{E}}_{P}[f(x)] \right| \leq \gamma.$$

Lemma B.2.2 (Uniform concentration over $\mathcal{A}_{k,P}$). *Let S be a set of n i.i.d. data points from a distribution P, and let $\mathcal{A}_{k,P}$ be as defined in Equation (4.4). There exists a constant $c > 0$ such that if $n \geq c(k^2 \log d + \log(1/\tau))/(q^2)$, then Equation (4.5) holds with probability at least $1 - \tau$ over the set S of n i.i.d. points from distribution P.*

Proof. Let Q be the distribution of $y := xx^\top$. Let $\mathcal{F} := \{\mathbb{I}_{y \cdot A > s_1} : A \in \mathcal{A}_k\}$. Suppose for now that VC dimension of $\mathcal{F}$ is less than $Ck^2 \log(d)$. Then the standard VC inequality (Fact B.2.1) implies that if $n \geq c(k^2 \log d + \log(1/\tau))/(q^2)$, then Equation (4.5) holds because under $y \sim Q$, $\mathbb{P}(y \cdot A > s_1) \leq q$ for all $A \in \mathcal{A}_{k,P}$. Thus it remains to show an upper bound on the VC dimension of $\mathcal{F}$. Since $\mathcal{F}$ corresponds to a family of linear functions that are k^2-sparse in d^2 dimensional space, [AV19, Theorem 6] implies that the VC dimension is at most $4k^2 \log(3d)$. This completes the proof. □

B.3 Choice of Numerical Constants

This section shows how to pick the numerical constants q, s_1, s_2, s_3, V_Z and B. In the proof of Theorem 4.5.1, these constants need to satisfy the following constraints:

1. $s_3 \geq 2$.
2. q is at least a small constant since the sample complexity is inversely proportional to $1/q^2$.
3. See (4.8):
$$\frac{1}{s_2} + \frac{s_2}{s_3^2} \leq 10^{-6}.$$
4. See (4.10):
$$\frac{\sigma^2}{s_1} + \frac{4}{s_3} + \frac{s_2 \times \nu^4}{s_3 \times s_1^2} \leq 10^{-6} \times q.$$
5. See (4.12): $B \geq s_3 \times s_1 + 10\sqrt{V_Z}$.
6. See (4.13): $\frac{0.57(\sigma^2 \times r^2 \times s_2)}{\sqrt{V_Z}} \leq 0.1$.
7. Paley-Zygmund: $0.004 \geq 4 \times q$.

Therefore, we pick the constants as follows:

1. ν, σ and r are the numbers we get from the ℓ_∞ truncation, and thus there is nothing to choose here.
2. $q = 0.001$.

3. $s_2 = 10^7$.

4. $s_3 = 10^{10}$.

5. Solve for s_1 in terms of above in Constraint 4. It suffices to take $s_1 = \max(\sigma^2, \nu^2) \times 10^{10}$.

6. Solve for $\sqrt{V_Z}$ using Constraint 6. It suffices to take $V_Z = 10^{16}\sigma^4 r^4$.

7. Solve for B using Constraint 5. It suffices to take $B = \max(\sigma^2, \sigma^2 r^2, \nu^2) \times 10^{20}$.

C APPENDIX TO CHAPTER 5

C.1 List of Algorithms

C.1.1 Iterative Filtering Algorithm

A recent line of work in the robust mean estimation literature has led to various algorithms that succeed when stability holds (see Diakonikolas and Kane [DK19] for a recent survey). We choose to work with the iterative filtering algorithm with independent removal [DK19]:

Theorem C.1.1. (*Diakonikolas and Kane [DK19]*) *Let $\epsilon < 1/2$, and suppose $S \subseteq \mathbb{R}^p$ is a multiset such that there exists a subset $S' \subseteq S$ such that (i) $|S'| \geq (1-\epsilon)|S|$ and (ii) S' is $(C\epsilon, \delta)$-stable with respect to μ and σ^2 for a large enough constant $C > 1$. Let T be an ϵ-corrupted version of S. Then there exists a computationally efficient algorithm that, given T and ϵ as inputs, with probability at least $1 - O(\exp(-\Omega(n\epsilon)))$, outputs a multiset $T' \subseteq T$ such that (i) $|T'| \geq (1 - c_1\epsilon)|T|$ and (ii) T' is $(c_2 C\epsilon, c_3\delta)$-stable with respect to μ and σ^2.*

Remark C.1.2. *Note that by the definition of stability, the empirical mean of an (ϵ, δ)-stable set lies within $\sigma\delta$ of μ. Thus, Theorem C.1.1 provides a high-probability error bound on the empirical mean of the filtered data points, when the original data set is an ϵ-corrupted version of a data set containing a large stable subset.*

Stability-based algorithms use the fact that if the empirical covariance matrix has a small spectral norm, the empirical mean is a good estimate of μ. The algorithm mentioned in Theorem C.1.1 uses this insight to obtain a subset of cardinality $(1-O(\epsilon))n$ such that the resulting empirical covariance matrix has a small spectral norm. At a high level, the algorithm iteratively uses the projection of the points along the leading eigenvector of the empirical covariance matrix (of the remaining points) to define a distribution over the (remaining) points such that the probability mass over the outliers is greater than the mass over the inliers. This distribution is then used to remove points stochastically, so that at each iteration, the algorithm is more likely to remove outliers than inliers. Since the number of outliers is at most ϵn, it does not remove too many inliers. Whereas prior work has focused on using the filtering algorithm mentioned in Theorem C.1.1 as a subroutine to find an estimate $\widehat{\mu}$ for μ (or, more generally, to robustly estimate the gradient of a function), we emphasize that our motivation in applying the filtering algorithm is to *identify a subset* T' that satisfies weak stability—indeed, mean

estimation is unnecessary because we already know the covariate distribution is centered around 0.

The probability of success of our preprocessing step will depend on the probability of success of Theorem C.1.1 applied to i.i.d. data from a distribution satisfying Assumption 5.2.1. We will use the following recent result from Diakonikolas et al. [DKP20], which provides a useful guarantee for when the condition of Theorem C.1.1 is satisfied with high probability:

Theorem C.1.3. (*Diakonikolas et al. [DKP20]*) *Let S be a set of n i.i.d. points from a distribution in $\mathbb{R}^p$ with mean μ and covariance I. Suppose the distribution satisfies $(k, 2)$-hypercontractivity with parameter σ_k, for some $k \geq 4$. Suppose $\epsilon' := C\left(\epsilon + \frac{\log(1/\tau)}{n}\right) = O(1)$, for a large enough constant C. Then with probability at least $1 - \tau$, there exists a subset $S' \subseteq S$ such that $|S'| \geq (1 - \epsilon')|S|$ and S' is $(C_1\epsilon', \delta)$-stable, where $C_1 > 2$ is any large constant and $\delta = O\left(\sqrt{\frac{p \log p}{n}} + \sigma_k \epsilon^{1-\frac{1}{k}} + \sigma_4\sqrt{\frac{\log(1/\tau)}{n}}\right)$, with prefactor depending on C_1.*

Combining the two theorems above, we see that with probability $1 - \tau$, we can identify a large subset $S' \subseteq S$, in a computationally efficient manner, such that S' is $(O(\epsilon), \delta)$-stable for an appropriate choice of ϵ and δ as specified by Theorem C.1.3. This rather technical conclusion is the starting point of our work.

C.1.2 Additional Algorithms

Algorithm 11 Huber Regression Asymmetric Noise

function HUBER_REGRESSION_WITH_FILTERING($(x_i, y_i)_{i\in[2n]}, \gamma, \epsilon'$)
  for $i \leftarrow 1$ to n **do**
    $(x_i', y_i') \leftarrow \left(\frac{x_i - x_{n+i}}{\sqrt{2}}, \frac{y_i - y_{n+i}}{\sqrt{2}}\right)$
  end for
  $S_1 \leftarrow$ FilteredCovariates$((x_i')_{i\in[n]}, \epsilon')$
  $\widehat{\beta} \leftarrow$ HuberRegression$((x_i', y_i')_{i\in S_1}, \gamma)$
  return $\widehat{\beta}$
end function

C.2 Notation and Definitions

Here, we list some notation and basic definitions used in the paper. For a real-valued random variable z, let $\|z\|_{\psi_2}$ denote the sub-Gaussian norm of z. We use $[n]$ as a shorthand for $\{1, \ldots, n\}$. For a vector $b \in \mathbb{R}^n$ and $m \in [n]$, we say that b is m-sparse if at most

Algorithm 12 Alternating minimization algorithm

function ALTERNATING_MINIMIZATION($(x_i, y_i)_{i\in[n]}, m, J$)
  $b^0 \leftarrow 0$
  for $j \leftarrow 1$ to J **do**
    $b^j \leftarrow \mathrm{HT}_m(P_X b^{j-1} + (I - P_X)y)$
  end for
  $\widehat{\beta}_J \leftarrow (X^\top X)^{-1} X^\top (y - b^j)$
  return $\widehat{\beta}_J$
end function

Algorithm 13 Alternating minimization algorithm

function ALTERNATING_MINIMIZATION_WITH_FILTERING($(x'_i, y'_i)_{i\in[n]}, \epsilon', m, J$)
  $T_1 \leftarrow$FilteredCovariates$((x_i)_{i\in[n]}, \epsilon')$
  $\widehat{\beta}_J \leftarrow$ ALTERNATING_MINIMIZATION$((x'_i, y'_i)_{i\in T_1}, m, J)$
  return $\widehat{\beta}_J$
end function

Algorithm 14 LAD with filtered covariates

function LAD_WITH_FILTERING($(x'_i, y'_i)_{i\in[n]}, \epsilon'$)
  $T_1 \leftarrow$FilteredCovariates$((x_i)_{i\in[n]}, \epsilon')$
  $\widehat{\beta}_{\mathrm{LAD}} \leftarrow \mathrm{LAD}((x'_i, y'_i)_{i\in T_1})$
  return $\widehat{\beta}_{\mathrm{LAD}}$
end function

m entries of b are nonzero, and we also write $\|b\|_0 = m$. For $1 \le i \le n$, we write $|b|_{(i)}$ to denote the i^{th} smallest component of b according to magnitude. Let $\mathcal{S}^{n-1}$ denote the unit sphere in n dimensions. For a square matrix M, we use $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ to denote the largest and smallest eigenvalues, respectively. We use $\|M\|_2$ to denote the spectral norm. For two matrices M_1, M_2, we write $M_1 \succeq M_2$ to denote the fact that $M_1 - M_2$ is positive semidefinite.

For a differentiable function f, we use ∇f to denote its gradient. For a scalar $x \in \mathbb{R}$, we use $\mathrm{sgn}(x)$ to denote the sign of x, i.e., $\mathrm{sgn}(x) = 0$ for $x = 0$; $\mathrm{sgn}(x) = 1$ for $x > 0$; and $\mathrm{sgn}(x) = -1$ for $x < 0$. For two sets A and B, let $A \setminus B$ denote the set difference and let $A \triangle B$ denote the symmetric difference. Let $\mathbb{I}(A)$ denote the indicator function over a set A.

We use $c, C, c_1, C_1, \ldots$ to denote absolute positive constants with values that might change from line to line. We also use the standard big-O notation to simplify the expressions in two regimes: For two nonnegative functions f and g with domain D,

we say that $f = O(g)$ when one of the following is true: (i) $D = \mathbb{N}$, and there exists constants C and n_0 such that $f(n) \leq Cg(n)$ for all $n \geq n_0$; or (ii) $D = [0, 1]$, and there exists constants C and $\epsilon_0 \in (0, 1)$ such that $f(\epsilon) \leq Cg(\epsilon)$ for $\epsilon \leq \epsilon_0$. The setting will be clear from context. We say that $f = \Omega(g)$ if $g = O(f)$, and we say that $f = \Theta(g)$ when $f = O(g)$ and $f = \Omega(g)$. We also use $\lesssim$ and $\gtrsim$ to hide constants.

We also recall the following definitions:

Definition C.2.1. (*Hypercontractivity*) *A random vector $X \in \mathbb{R}^p$* satisfies $(k, 2)$-hypercontractivity with parameter σ_k *if for all unit vectors $v \in \mathbb{R}^p$, we have* $\left(\mathbb{E}\,|v^\top X|^k\right)^{1/k} \leq \sigma_k \left(\mathbb{E}(v^\top X)^2\right)^{1/2}$.

We note that $(k, 2)$-hypercontractivity is also referred to as k^{th} bounded moments or $L_k - L_2$ norm equivalence.

Definition C.2.2. (*Strong convexity*) *For a convex set $\mathcal{X} \subseteq \mathbb{R}^n$, we say that a continuously differentiable function $f : \mathcal{X} \to \mathbb{R}$ is* α-strongly convex *if for any $x, y \in \mathcal{X}$, we have*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|_2^2.$$

C.3 Contributions and Related Work

In this appendix, we further discuss our contributions in the context of previous work.

C.3.1 Contributions

We will assume throughout our paper that prior to contamination, the covariates are drawn from a distribution with mean 0 and covariance I, and also satisfies a property known as hypercontractivity (bounded fourth moments). We will also assume that the additive noise in the linear model is independent of the covariates and (in most cases) has finite first or second moments. Note that these assumptions are significantly less restrictive than the usual assumptions of sub-Gaussianity. Under these assumptions, we can show that the Huber estimator after filtering achieves the optimal ℓ_2-error rate of $O\left(\sigma\sqrt{\frac{p}{n}} + \sigma\sqrt{\frac{\log(1/\tau)}{n}}\right)$, provided the sample size satisfies $n = \Omega(p \log p)$. Furthermore, our method is computationally feasible, since we simply need to perform the iterative filtering algorithm, followed by optimization of a convex function. If adversarial contamination is introduced to the covariates and/or response variables, the error bound of the filtered Huber estimator becomes $O\left(\sigma\left(\sqrt{\frac{p \log p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \epsilon^{1-1/k}\right)\right)$,

provided $n = \Omega(p \log p)$ and the covariates satisfy an additional k^{th} moment bound, for $k \geq 4$. Note that the dependence on ϵ matches the lower bound derived in Bakshi and Prasad [BP21].[35] When the covariates are drawn from a Gaussian distribution with identity covariance, the error of the filtered Huber estimator improves to $O\left(\sigma\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \epsilon\sqrt{\log(1/\epsilon)}\right)\right)$, provided $n = \Omega(p)$. The dependence on p, n, and τ is optimal, while the dependence on ϵ is nearly-optimal up to a $\sqrt{\log(1/\epsilon)}$ factor [CGR16]. (This rate also shaves off the additional $\sqrt{\log(1/\epsilon)}$ factor achieved in previous works [DKS19; CATJFB20], which obtained the rate $O(\epsilon \log(1/\epsilon))$ in terms of ϵ.) Going back to the heavy-tailed setting, i.e., when the covariates are drawn from a distribution with mean zero and bounded fourth moments, we extend our analysis to the setting where the covariance matrix Σ of the covariates is unknown but satisfies $(1/2)I \preceq \Sigma \preceq 2I$. We show that the filtered Huber estimator achieves the error rate $O\left(\sigma\left(\sqrt{\frac{p \log p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \sqrt{\epsilon}\right)\right)$, provided $n = \Omega(p \log p)$. The SQ lower bound of [DKS19] suggests that such a dependence on ϵ is essentially optimal for computationally-efficient algorithms when $n = o(p^2)$ even when the uncontaminated distribution is Gaussian.

We derive error bounds for the LTS and LAD estimators under slightly different assumptions: When the noise distribution has bounded $(k')^{\text{th}}$ moments, for some $k' \geq 2$, we obtain an error rate of the form $O\left(\sigma\left(\frac{p \log p}{n} + \epsilon + \frac{\log(1/\tau)}{n}\right)^{1/2 - 1/k'}\right)$ for the LTS estimator, provided $n = \Omega(p \log p)$. Assuming a first moment bound of κ on the noise distribution, we can show that the LAD estimator has ℓ_2-error $O(\kappa)$, provided $n = \Omega(p \log p)$. Although the error bounds for the LTS and LAD estimators are somewhat weaker than the bounds we obtain for the Huber regression estimator, we note that the LTS estimator is extremely quick to compute in practice [BJK15; BJKK17], and the LAD estimator does not involve any tuning parameters, unlike the Huber estimator (which requires a tuning parameter for the loss) and the LTS estimator (which requires a tuning parameter specifying the degree of trimming). Furthermore, we show that a simple postprocessing step involving applying the robust multivariate mean algorithm to a shifted data set can be used to obtain near-optimal error guarantees in terms of τ and p. Lastly, we note that the LTS or LAD estimators may be practically useful for initializing a gradient descent algorithm when optimizing the Huber regression objective in order

[35] However, the k^{th} moment of the covariates in the lower bound instance of [BP21, Theorem 6.1] increases as ϵ decreases, so the optimal rate (in terms of ϵ) could possibly be better. On the other hand, no polynomial-time estimators with better rates are currently known for our setting. We also note that the lower bound of $\Omega(\sqrt{\epsilon})$ [CATJFB20, Theorem D.1] for the case of bounded fourth moments is not applicable here due to different assumptions: the covariance is degenerate in [CATJFB20, Theorem D.1].

to save on computation.

C.3.2 Related Work

Several recent works have highlighted significant challenges that appear in the presence of heavy-tailed responses and/or adversarial contamination in responses [LDB09; NTN11; NT13; BJK15; MGJK19; SF20; WLJ07]. In all of these works, the covariates are assumed to satisfy strong assumptions: sub-Gaussian tails and no contamination. The preceding works can be loosely categorized into two categories: (i) regularization-based estimators and (ii) thresholding-based estimators. In the first category, a popular choice is a penalized Lasso-type estimator that solves $\min_{\beta,z} \left\{ \frac{1}{n}\|y - X\beta - z\|_2^2 + \lambda\|z\|_1 \right\}$, where the variable z accounts for outliers in the response variables. Several works have shown that Lasso-type estimators can handle contamination or heavy-tailed noise in responses [NT13; SF20]—indeed, Huber regression is closely related to penalized Lasso-type estimators [SO11; SF20]. The idea of using the Huber loss for estimation under heavy-tailed error distributions has recently been studied in the context of mean estimation [Cat12; Min19] and regression [FLW17; SZF20]. Our work on Huber regression is closely related to Sun et al. [SZF20], and we roughly follow their proof structure. However, we establish significantly tighter results for heavy-tailed covariates (see Section 5.3 for more details).

Another popular convex estimator is the LAD estimator with a Lasso penalty [WLJ07; KP19]. In the dense setting, Karmalkar and Price [KP19] (see also [DMT07]) studied the LAD estimator $\min_\beta \|y - X\beta\|_1$, and showed its robustness to adversarial contamination in the responses. However, their theory imposes a deterministic condition on the covariates that holds with high probability for sub-Gaussian distributions, but not necessarily for heavy-tailed or corrupted covariates. As opposed to convex relaxation-based estimators, several recent works have studied alternating minimization algorithms for robust regression [JK17; BJK15; BJKK17; JTK14]. These algorithms were developed to optimize the nonconvex objective function corresponding to the LTS estimator [Rou84]. In our paper, we critically leverage the aforementioned results on LAD [KP19] and LTS [BJK15; BJKK17] estimation by showing that the deterministic conditions under which the respective algorithms are guaranteed to succeed are satisfied with high probability by our preprocessed covariates.

Turning to papers which analyze corruption in both covariates and responses, a general framework for robust convex optimization was considered in Diakonikolas et al. [DKKLSS19] and Prasad et al. [PSBR20] using the robust mean estimation algorithm

on gradients of the loss function. Although these results lead to polynomial-time estimators for several tasks, the resulting rates are suboptimal for linear regression. In the Gaussian setting, Diakonikolas et al. [DKS19] proposed computationally efficient estimator with near-optimal error (as a function of ϵ) guarantees under adversarial contamination in both covariates and responses. In fact, our result improves the dependence from $\epsilon \log(1/\epsilon)$ to $\epsilon\sqrt{\log(1/\epsilon)}$.

C.4 Auxiliary Results

We recall the Chernoff bound below [Ver18; BLM13]:

Lemma C.4.1. *Let $X_1, \dots, X_n$ be independent $\{0,1\}$-valued random variables. Let $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ be the empirical mean and let μ denote its expectation, i.e., $\mu = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\, X_i$. Then with probability at least $1-\tau$, we have*

$$\widehat{\mu} \lesssim \mu + \frac{\log(1/\tau)}{n}.$$

In particular, for $\kappa \geq 1$, we have $\widehat{\mu} \leq 2\kappa\mu$, with probability at least $1 - \exp(-c\kappa n\mu)$.

We will use the following version of Talagrand's concentration inequality regarding bounded empirical processes [Tal96b]:

Lemma C.4.2. (*Theorem 12.5 of Boucheron et al. [BLM13]*) *Let $X_1, \dots, X_n$ be n i.i.d. vectors such that for each $s \in \mathcal{T}$, we have $\mathbb{E}\, X_{i,s} = 0$ and $X_{i,s} \leq L$. Define $Z := \sup_{s\in\mathcal{T}} \sum_{i=1}^n X_{i,s}$, and define σ^2 (the wimpy variance) to be $\sigma^2 := \sup_{s\in\mathcal{T}} \mathbb{E}\sum_{i=1}^n X_{i,s}^2$. Then with probability at least $1-\tau$, we have*

$$Z \lesssim \mathbb{E}\, Z + \sigma\sqrt{\log(1/\tau)} + L\log(1/\tau).$$

We recall the following lemma from Lugosi and Mendelson [LM21b]:

Lemma C.4.3. (*Lugosi and Mendelson [LM21b]*) *Let $X_1, \dots, X_n$ be n i.i.d. points from a distribution over $\mathbb{R}^p$ with mean zero and covariance Σ. For an $\epsilon > 0$ such that $\epsilon = O(1)$, let $Q := C\left(\sqrt{\frac{\|\Sigma\|_2}{\epsilon}} + \frac{1}{\epsilon}\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}\right)$ for a large enough constant C. For a unit vector v, define the set $S_v := \left\{i : |X_i^\top v| \geq Q\right\}$. Let $\mathcal{E}$ be the event $\mathcal{E} = \{\sup_v |S_v| \leq \epsilon n\}$. Then with probability at least $1 - \exp(-nc\epsilon)$, the event $\mathcal{E}$ holds.*

We will also require the following generalization of the result above from Diakonikolas et al. [DKP20, Lemma C.1]:

Lemma C.4.4. (*Diakonikolas et al. [DKP20]*) *Let $X_1, \dots, X_n$ be n i.i.d. points from a distribution over $\mathbb{R}^p$ with mean zero and covariance Σ. Suppose that for some $k \geq 2$, the inequality $\mathbb{E}\left((v^\top X_i)^k\right)^{1/k} \leq \sigma_{x,k}\,\mathbb{E}\left((v^\top X_i)^2\right)^{1/2}$ holds for all $v \in \mathcal{S}^{p-1}$. For some $\epsilon > 0$ such that $\epsilon = O(1)$, define $Q := C\left(\sigma_{x,k}\sqrt{\|\Sigma\|_2}\epsilon^{-1/k} + \frac{1}{\epsilon}\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}\right)$ for a large enough constant C. For a unit vector v, define the set $S_v := \left\{i : |X_i^\top v| \geq Q\right\}$. Let $\mathcal{E}$ be the event $\mathcal{E} = \{\sup_v |S_v| \leq \epsilon n\}$. Then with probability at least $1 - \exp(-nc\epsilon)$, the event $\mathcal{E}$ holds.*

We also need the following version of the matrix Bernstein inequality:

Lemma C.4.5. (*Corollary 7.3.2 of Tropp [Tro15]*) *Let $S_1, \dots, S_n$ be n independent symmetric matrices such that $\mathbb{E}[S_i] = 0$ and $\|S_i\|_2 \leq L$ a.s., for each index i. Let $Z = \sum_{i=1}^n S_i$, and let V be any positive semidefinite matrix such that $\sum_{i=1}^n \mathbb{E}[S_i S_i^\top] \preceq V$. Let $\nu = \|V\|_2$ and $r = \mathrm{rank}(V)$. Then*

$$\mathbb{E}[\|Z\|_2] \lesssim \sqrt{\nu \log r} + L \log r.$$

In particular, if $S_i = \xi_i x_i x_i^\top$, where ξ_i is a Rademacher random variable and x_i is sampled independently from a distribution with $\mathbb{E}[x_i x_i^\top] = \Sigma$ and bounded support $\sqrt{L}$, i.e., $\|x_i\|_2 \leq \sqrt{L}$ a.s. for each index i, we have $\mathbb{E}[\|Z\|_2] \lesssim \sqrt{nL\|\Sigma\|_2 \log(\mathrm{rank}(\Sigma))} + L\log(\mathrm{rank}(\Sigma))$.

We will also use the following results:

Lemma C.4.6. (*Lemma 6.1.2 of Vershynin [Ver18]*) *Let Y and Z be independent random variables such that $\mathbb{E}(Z) = 0$. Then for every convex function f, one has*

$$\mathbb{E}(f(Y)) \leq \mathbb{E}(f(Y + Z)).$$

Lemma C.4.7. *Let W and Z be two independent symmetric random variables. Let $Y := W + Z$. Then for any $r \geq 0$, we have $\mathbb{P}(|Z| \geq r) \leq 2\,\mathbb{P}(|Y| \geq r)$.*

Proof. Note that

$$\{Z \geq r, W \geq 0\} \cup \{Z \leq -r, W \leq 0\} \subseteq \{|Y| \geq r\}.$$

Thus, by the independence of W and Z and the symmetry of Z, we have

$$\mathbb{P}(|Y| \geq r) \geq \mathbb{P}(Z \geq r, W \geq 0) + \mathbb{P}(Z \leq -r, W \leq 0)$$

$$
\begin{aligned}
&= \mathbb{P}(Z \geq r)\,\mathbb{P}(W \geq 0) + \mathbb{P}(Z \leq -r)\,\mathbb{P}(W \leq 0) \\
&= \mathbb{P}(Z \geq r)\Big(\mathbb{P}(W \geq 0) + \mathbb{P}(W \leq 0)\Big) \\
&\geq \mathbb{P}(Z \geq r) \\
&= \frac{1}{2}\,\mathbb{P}(|Z| \geq r),
\end{aligned}
$$

completing the proof. □

We also recall the following result on convex functions from Sun et al. [SZF20]:

Lemma C.4.8. *Let $\mathcal{L}(\beta) : \mathbb{R}^p \to \mathbb{R}$ be a convex function and let $\beta_1 \in \mathbb{R}^p$. For some $\eta \in (0, 1]$ and $\beta_2 \in \mathbb{R}^p$, let $\beta_\eta = \beta_1 + \eta(\beta_2 - \beta_1)$. Then we have*

$$\langle \nabla \mathcal{L}(\beta_\eta) - \nabla \mathcal{L}(\beta_1), \beta_\eta - \beta_1 \rangle \leq \eta \langle \nabla \mathcal{L}(\beta_2) - \nabla \mathcal{L}(\beta_1), \beta_2 - \beta_1 \rangle.$$

We will use the following standard properties regarding convexity and strong convexity [Nes04; BV04]:

Lemma C.4.9. *For a convex set $\mathcal{X} \subseteq \mathbb{R}^n$, let f be a continuously differentiable function $f : \mathcal{X} \to \mathbb{R}$. Then the following statements hold:*

1. *If f is α-strongly convex and continuously differentiable, then for any two points $x, y \in \mathcal{X}$, we have*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \alpha \|y - x\|_2^2.$$

2. *If f is twice continuously differentiable, then $\nabla^2 f \succeq \alpha I$.*

3. *If f is α_1-strongly convex and g is α_2-strongly convex, then $f + g$ is $(\alpha_1 + \alpha_2)$-strongly convex.*

C.5 Lower bounds for OLS and Multivariate Sample Mean

In this appendix, we derive a lower bound on the ℓ_2-error of the OLS estimator by first proving a lower bound on the estimation error of the empirical mean.

C.5.1 Lower Bound for Mean Estimation

We prove the following result regarding the estimation error of the sample mean. This result generalizes an analogous univariate result of Catoni [Cat12, Proposition 6.2].

Proposition C.5.1. *For any variance of $\sigma^2 > 0$, dimension p, sample size n, and probability $\tau \leq \frac{1}{4}$, there exists a multivariate distribution with mean $\mu \in \mathbb{R}^p$ and covariance $\sigma^2 I$ such that the sample mean $\widehat{\mu}$ on n i.i.d. samples satisfies the bound*

$$\|\widehat{\mu} - \mu\|_2^2 = \Omega\left(\frac{p\sigma^2}{n\tau}\right),$$

with probability at least τ. Moreover, the distribution of the random variable $\mu + ZX$ satisfies the bound, where X is uniform on $\{-1, 1\}^p$ and Z is a univariate random variable supported on $\left\{-\sigma\sqrt{\frac{n}{2\tau}}, 0, \sigma\sqrt{\frac{n}{2\tau}}\right\}$, with

$$\mathbb{P}\left(Z = -\sigma\sqrt{\frac{n}{2\tau}}\right) = \mathbb{P}\left(Z = \sigma\sqrt{\frac{n}{2\tau}}\right) = \frac{\tau}{n},$$

and X and Z are independent.

Proof. Without loss of generality, we will assume that $\mu = 0$. Let $\epsilon = \frac{1}{\sqrt{2n\tau}}$, so

$$\mathbb{P}(Z = -\sigma n\epsilon) = \mathbb{P}(Z = \sigma n\epsilon) = \frac{1}{2n^2\epsilon^2}$$

and $\mathbb{P}(Z = 0) = 1 - \frac{1}{n^2\epsilon^2}$. Note that $\mathrm{Cov}(ZX) = \mathbb{E}(Z^2)I = \sigma^2 I$.

Let $(X_1, \ldots, X_n)$ and $(Z_1, \ldots, Z_n)$ be independent pairs of n i.i.d. random samples drawn from the distributions of X and Z, respectively. Let $W_i := Z_i X_i$, so the W_i's are i.i.d. and $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n W_i$. Now note that for all i, we have $\|X_i\|_2 = \sqrt{p}$. Hence, we can write

$$\begin{aligned}
\mathbb{P}\left(\|\widehat{\mu} - \mu\|_2 \geq \sigma\sqrt{p}\epsilon\right) &= \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n W_i\right\|_2 \geq \sigma\sqrt{p}\epsilon\right) \\
&\geq \mathbb{P}\left(\exists i : \|W_i\|_2 \geq \sigma n\sqrt{p}\epsilon \text{ and } \forall j \neq i, \|W_j\|_2 = 0\right) \\
&= \mathbb{P}\left(\exists i : \|X_i\|_2|Z_i| \geq \sigma n\sqrt{p}\epsilon \text{ and } \forall j \neq i, \|Z_j X_j\|_2 = 0\right) \\
&= \mathbb{P}\left(\exists i : |Z_i| \geq \sigma n\epsilon \text{ and } \forall j \neq i, Z_j = 0\right) \\
&= n \cdot \frac{1}{n^2\epsilon^2}\left(1 - \frac{1}{n^2\epsilon^2}\right)^{n-1} \\
&\geq \frac{1}{n\epsilon^2}\left(1 - \frac{1}{n^2\epsilon^2}\right)^n.
\end{aligned}$$

We now simplify the last term using two simple observations: (i) $(1+x)^r \geq 1+rx$, for $x \geq -1$ and $r \geq 1$; and (ii) $\frac{1}{n\epsilon^2} = 2\tau \leq \frac{1}{2}$:

$$\frac{1}{n\epsilon^2}\left(1-\frac{1}{n^2\epsilon^2}\right)^n \geq \frac{1}{n\epsilon^2}\left(1-\frac{1}{n\epsilon^2}\right) \geq \frac{1}{2n\epsilon^2} = \tau.$$

Thus, we conclude that

$$\|\widehat{\mu}-\mu\|_2 \geq \sigma\sqrt{d}\epsilon = \sigma\sqrt{\frac{p}{2n\tau}},$$

with probability at least τ. □

C.5.2 Lower Bound for OLS

In this section, we state a lower bound for the OLS estimator using reductions to the sample mean. We consider the following linear model:

$$y_i = x_i^\top \beta^* + z_i, \qquad 1 \leq i \leq n,$$

where x_i and z_i are independent. We also assume that $\mathbb{E}(z_i^2) = \sigma^2$.

Proposition C.5.2. (*Lower bound for OLS for multivariate distributions*) *For every dimension p, sample size $n = \Omega(p)$, and probability $\tau \leq \frac{1}{4}$ such that $\frac{\log(1/\tau)}{n} = O(1)$, there exist covariate and error distributions satisfying Assumptions 5.2.1 and 5.2.3, such that the OLS estimator $\widehat{\beta}_{OLS}$ satisfies the bound*

$$\|\widehat{\beta}_{OLS}-\beta^*\|_2^2 = \Omega\left(\frac{p\sigma^2}{n\tau}\right),$$

with probability at least $\frac{\tau}{2}$. Moreover, the bound is satisfied when the distribution of the covariates is uniform on $\{-1,1\}^p$, and the distribution of the noise is defined as in Proposition C.5.1.

Proof. Suppose the covariates and noise are sampled according to the stated distributions; we will show that the lower bound holds. Let the corresponding sampled points be denoted by $\{(x_i, y_i)\}_{i=1}^n$.

Note that the distribution of the covariates is $O(1)$-sub-Gaussian; i.e., for any unit vector v, we have $\|v^\top x\|_{\psi_2} = O(1)$. Thus, Assumption 5.2.1 holds. Furthermore, the covariance matrix of the covariates has exponential concentration near the true covariance

I, so if we denote $\Sigma_n = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and define the event

$$\mathcal{E}_1 := \left\{x_1, \dots, x_n : \|\Sigma_n^{-1} - I\|_2 \le 0.1\right\},$$

then $\mathbb{P}(\mathcal{E}_1) \ge 1 - \exp(-cn)$ when $n = \Omega(p)$ (cf. Exercise 4.7.3 of Vershynin [Ver18]).

Define $\widehat{W} := \frac{1}{n}\sum_{i=1}^n x_i z_i$, and note that the OLS estimator satisfies $\widehat{\beta} - \beta^* = \Sigma_n^{-1}\widehat{W}$. Thus,

$$\|\widehat{\beta}_{OLS} - \beta^*\|_2 \ge \|\widehat{W}\|_2 - \|(\Sigma_n^{-1} - I)\widehat{W}\|_2 \ge \|\widehat{W}\|_2 - \|\Sigma_n^{-1} - I\|_2\|\widehat{W}\|_2.$$

Let $\mathcal{E}_2$ be the event

$$\mathcal{E}_2 := \left\{\|\widehat{W}\|_2 = \Omega\left(\sqrt{\frac{p\sigma^2}{n\tau}}\right)\right\}.$$

Then on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\|\widehat{\beta}_{OLS} - \beta^*\|_2 \ge 0.9\|\widehat{W}\|_2 = \Omega\left(\sqrt{\frac{p\sigma^2}{n\tau}}\right).$$

Finally, note that $\mathbb{P}(\mathcal{E}_2) \ge \tau$ by Proposition C.5.1, so $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \ge \tau - \exp(-cn) \ge \frac{\tau}{2}$, and the desired result follows. □

C.6 Results Regarding Stability

In this appendix, we state and prove several results stemming from our notions of stability.

Proposition C.6.1. *Let $S = \{x_1, \dots, x_n\}$ be an (ϵ, δ)-stable set with respect to $\mu = 0$ and $\sigma^2 = 1$, such that $\frac{\delta^2}{\epsilon} < 1$. Then S is also (ϵ, L, U)-weakly stable with $L = (1-\epsilon)\left(1 - \frac{\delta^2}{\epsilon}\right)$ and $U = 1 + \frac{\delta^2}{\epsilon}$. In particular, if $\frac{\delta^2}{\epsilon} < 0.5$, we have $L = \Omega(1)$ and $U = O(1)$.*

Proof. By the definition of strong stability and the triangle inequality, we clearly have

$$\left\|\frac{1}{n}\sum_{i\in[n]} x_i x_i^\top\right\|_2 \le 1 + \frac{\delta^2}{\epsilon},$$

showing that we can take $U = 1 + \frac{\delta^2}{\epsilon}$.

For the lower bound, consider a subset $S \subseteq [n]$ such that $|S| \geq (1-\epsilon)n$. By the stability condition, we know that for any unit vector v, we have

$$v^\top \left(I - \frac{1}{|S|} \sum_{i \in S} x_i x_i^\top \right) v \leq \frac{\delta^2}{\epsilon},$$

implying that

$$\frac{n}{|S|} \cdot v^\top \left(\frac{1}{n} \sum_{i \in S} x_i x_i^\top \right) v \geq 1 - \frac{\delta^2}{\epsilon}.$$

Hence,

$$\lambda_{\min} \left(\frac{1}{n} \sum_{i \in S} x_i x_i^\top \right) \geq \frac{|S|}{n} \left(1 - \frac{\delta^2}{\epsilon} \right) \geq (1-\epsilon) \left(1 - \frac{\delta^2}{\epsilon} \right),$$

giving the desired result. The second result follows by noting that $\epsilon < 1/2$. □

Proposition C.6.2. *Let $S = \{x_1, \ldots, x_n\}$ be a set of n i.i.d. points in $\mathbb{R}^p$ from a distribution P with mean 0 and covariance Σ. Suppose the following holds:*

1. *$\kappa_l I \preceq \Sigma \preceq \kappa_u I$, where $\kappa_l \in (0, 1]$ and $\kappa_u \geq 1$ are constants.*
2. *The distribution P satisfies $(4, 2)$-hypercontractivity with parameter $\sigma_{x,4}$.*

Let $\epsilon < c^*$, where c^* is a small enough constant depending on $\sigma_{x,4}$ and $\frac{\kappa_l}{\kappa_u}$. Suppose $n \gtrsim \frac{\kappa_u^2}{\kappa_l^2} \cdot \frac{(p \log p)\sigma_{x,4}^2}{\sqrt{\epsilon}} + \frac{\kappa_u}{\kappa_l} \cdot \frac{p}{\epsilon}$. Then with probability at least $1 - O(\exp(-\Omega(n\epsilon)))$, for every subset $S' \subseteq S$ such that $|S'| \geq (1-\epsilon)n$, we have $\lambda_{\min}\left(\frac{1}{n}\sum_{i \in S'} x_i x_i^\top\right) \geq 0.8\kappa_l$.

Proof. The proof follows the same principle as the references [KM15; DKP20]. In particular, the proof is similar to Diakonikolas et al. [DKP20, Lemma 4.3] who consider the case when $\kappa_l = \kappa_u = 1$. For completeness, we provide a full proof here for the general case.

Let $r \geq 2$ denote a large enough constant to be specified later. First, we only consider distributions which are supported on a ball of radius at most $r\sigma_{x,4}\sqrt{\kappa_u}\epsilon^{-1/4}\sqrt{p}$. (This is because a standard argument shows that we can simply ignore the points that do not satisfy this condition, since $(\mathbb{E}\,\|X\|_2^4)^{1/4} \leq \sigma_{x,4}\sqrt{\kappa_u}\epsilon^{-1/4}\sqrt{p}$ for $X \sim P$, as outlined at the end of the proof.) We will allow P to have a nonzero mean μ, as long as $\|\mu\|_2 \leq \sigma_{x,4}\sqrt{\kappa_u}\epsilon^{-1/4}$.

We will now apply Lemma C.4.4, which establishes a bound for an $(1-\epsilon)$-fraction of points when projected along any unit vector. Let $Q = C\left(\sigma_{x,4}\sqrt{\kappa_u}\epsilon^{-1/4} + \frac{1}{\epsilon}\sqrt{\frac{p\kappa_u}{n}}\right) + \|\mu\|_2$, which is greater than the threshold from Lemma C.4.4 applied to the recentered distribution P. Using the bound on $\|\mu\|_2$, we have $Q \lesssim \left(\sigma_{x,4}\sqrt{\kappa_u}\epsilon^{-1/4} + \frac{1}{\epsilon}\sqrt{\frac{p\kappa_u}{n}}\right)$. Let $\mathcal{E}$ denote the

event from Lemma C.4.4, stating that for any unit vector v, we have $\left|\{i : |x_i^\top v| \geq Q\}\right| \leq \epsilon n$. By Lemma C.4.4, we know that $\mathbb{P}(\mathcal{E}) \geq 1 - \exp(-cn\epsilon)$.

We will now assume that the event $\mathcal{E}$ holds and incur an additional failure probability of $\exp(-cn\epsilon)$ by a union bound. Define the function $f : \mathbb{R}_+ \to \mathbb{R}_+$, as follows:

$$f(x) = \begin{cases} x, & \text{if } x \in [0, Q^2], \\ Q^2, & \text{otherwise,} \end{cases},$$

and let $g(x) = -f(x)$. For any $v \in \mathcal{S}^{p-1}$, on the event $\mathcal{E}$, we have the following bound:

$$\begin{aligned} \min_{S' : |S'| \geq (1-\epsilon)n} \sum_{i \in S'} (x_i^\top v)^2 &\geq \sum_{i=1}^n f((x_i^\top v)^2) - \epsilon Q^2 n \\ &= -\left(\sum_{i=1}^n g((x_i^\top v)^2) - \mathbb{E}\, g((x_i^\top v)^2)\right) + n\,\mathbb{E}\, f((x_i^\top v)^2) - \epsilon Q^2 n. \end{aligned}$$

Taking an infimum over $v \in \mathcal{S}^{p-1}$, we then have

$$\begin{aligned} \inf_{v \in \mathcal{S}^{p-1}} \min_{S' : |S'| \geq (1-\epsilon)n} \sum_{i \in S'} (x_i^\top v)^2 \geq -\epsilon Q^2 n - \sup_{v \in \mathcal{S}^{p-1}} \left(\sum_{i=1}^n g((x_i^\top v)^2) - \mathbb{E}\, g((x_i^\top v)^2)\right) \\ + n\left(\inf_{v \in \mathcal{S}^{p-1}} \mathbb{E}\, f((x_i^\top v)^2)\right). \quad \text{(C.1)} \end{aligned}$$

Now define the random variable

$$N := \sup_{v \in \mathcal{S}^{p-1}} \sum_{i=1}^n g((x_i^\top v)^2) - \mathbb{E}\, g((x_i^\top v)^2).$$

Let $\xi_1, \ldots, \xi_n$ be n i.i.d. Rademacher random variables. We first bound the expectation of N using symmetrization and contraction of Rademacher averages [LT91; BLM13]:

$$\begin{aligned} \mathbb{E}\, N &\leq 2\,\mathbb{E} \sup_{v \in \mathcal{S}^{p-1}} \left|\sum_{i=1}^n \xi_i g((x_i^\top v)^2)\right| \leq 4\,\mathbb{E} \sup_{v \in \mathcal{S}^{p-1}} \left|\sum_{i=1}^n \xi_i (x_i^\top v)^2\right| \\ &\leq 4\,\mathbb{E}\left(\left\|\sum_{i=1}^n \xi_i x_i x_i^\top\right\|_2\right) \\ &\lesssim \frac{r^2 \sigma_{x,4}^2 \kappa_u p \log p}{\sqrt{\epsilon}} + \sqrt{\frac{n r^2 \sigma_{x,4}^2 \kappa_u^2 p \log p}{\sqrt{\epsilon}}}, \end{aligned}$$

where the last step uses the matrix Bernstein inequality (Lemma C.4.5) with L equal to $(r\sigma_{x,4}\sqrt{\kappa_u}\epsilon^{-1/4}\sqrt{p})^2$ and $\nu = nL\kappa_u$, because $\|x_i\|_2 \leq r\sigma_{x,4}\sqrt{\kappa_u}\epsilon^{-1/4}\sqrt{p}$ and $\mathbb{E}\, x_i x_i^\top \preceq \kappa_u I$.

We now bound the following term (which is usually called the *wimpy variance* [BLM13]):

$$\sigma^2 := \sup_{v\in\mathcal{S}^{p-1}} n\,\mathbf{Var}(g((x_i^\top v)^2)) \le \sup_{v\in\mathcal{S}^{p-1}} n\,\mathbb{E}((x_i^\top v)^2)^2 \le n\sigma_{x,4}^4 (v^\top \Sigma v)^2 \le n\sigma_{x,4}^4\kappa_u^2.$$

Using Talagrand's inequality for bounded empirical processes (cf. Lemma C.4.2), we therefore have that with probability at least $1-\exp(-n\epsilon)$,

$$\begin{aligned}
\frac{N}{n} &\lesssim \frac{r^2\sigma_{x,4}^2\kappa_u p\log p}{n\sqrt{\epsilon}} + \sqrt{\frac{r^2\sigma_{x,4}^2\kappa_u^2 p\log p}{n\sqrt{\epsilon}}} + \sigma_{x,4}^2\kappa_u\sqrt{\epsilon} + \epsilon Q^2 \\
&\lesssim \frac{r^2\sigma_{x,4}^2\kappa_u p\log p}{n\sqrt{\epsilon}} + \sqrt{\frac{r^2\sigma_{x,4}^2\kappa_u^2 p\log p}{n\sqrt{\epsilon}}} + \sigma_{x,4}^2\kappa_u\sqrt{\epsilon} + \sigma_{x,4}^2\kappa_u\sqrt{\epsilon} + \frac{p\kappa_u}{\epsilon n} \\
&\lesssim \frac{r^2\sigma_{x,4}^2\kappa_u p\log p}{n\sqrt{\epsilon}} + \sqrt{\frac{r^2\sigma_{x,4}^2\kappa_u^2 p\log p}{n\sqrt{\epsilon}}} + \sigma_{x,4}^2\kappa_u\sqrt{\epsilon} + \frac{p\kappa_u}{\epsilon n},
\end{aligned}$$

where we use the definition of Q. By taking $\epsilon \lesssim \left(\frac{\kappa_l}{\kappa_u}\right)^2\left(\frac{1}{\sigma_{x,4}}\right)^4$ and $n \gtrsim \frac{\kappa_u^2}{\kappa_l^2}\cdot\frac{r^2\sigma_{x,4}^2(p\log p)}{\sqrt{\epsilon}} + \frac{\kappa_u}{\kappa_l}\cdot\frac{p}{\epsilon}$, we can make the expression above less than $0.05\kappa_l$. These calculations also show that we can upper-bound ϵQ^2 by $0.05\kappa_l$. Thus, we have the following:

$$\max\left\{\frac{N}{n}, \epsilon Q^2\right\} \le 0.05\kappa_l. \tag{C.2}$$

Finally, note that for any $v\in\mathcal{S}^{p-1}$, the Cauchy-Schwarz inequality gives

$$\begin{aligned}
\mathbb{E}\left|f((x_i^\top v)^2) - (x_i^\top v)^2\right| &= \mathbb{E}\left((x_i^\top v)^2\,\mathbb{I}\{(x_i^\top v)^2 > Q^2\}\right) \le \sqrt{\mathbb{E}(x_i^\top v)^4}\sqrt{\mathbb{P}(|x_i^\top v| > Q)} \\
&\le \frac{\mathbb{E}[|x_i^\top v|^4]}{Q^2} \lesssim \frac{\sigma_{x,4}^4\kappa_u^2}{\kappa_u\sigma_{x,4}^2\epsilon^{-1/2}} = \sqrt{\epsilon}\sigma_{x,4}^2\kappa_u,
\end{aligned}$$

implying that there exists a constant $c>0$ such that

$$\mathbb{E}\,f((x_i^\top v)^2) \ge \mathbb{E}(x_i^\top v)^2 - c\sigma_{x,4}^2\sqrt{\epsilon}\kappa_u \ge \kappa_l - c\sigma_{x,4}^2\sqrt{\epsilon}\kappa_u.$$

Taking $\epsilon \lesssim \left(\frac{\kappa_l}{\kappa_u}\right)^2\left(\frac{1}{\sigma_{x,4}}\right)^4$, we have

$$\mathbb{E}\,f((x_i^\top v)^2) \ge 0.95\kappa_l. \tag{C.3}$$

Combining inequalities (C.1), (C.2), and (C.3), we then obtain the bound

$$\frac{1}{n}\inf_{v\in\mathcal{S}^{p-1}}\min_{S':|S'|\geq(1-\epsilon)n}\sum_{i\in S'}(x_i^\top v)^2 \geq \inf_{v\in\mathcal{S}^{p-1}}\mathbb{E}\, f((x_i^\top v)^2) - \epsilon Q^2 - \frac{N}{n}$$
$$\geq 0.95\kappa_l - 0.05\kappa_l - 0.05\kappa_l \geq 0.85\kappa_l.$$

This completes the proof.

Unbounded support: We now outline a general argument for the case when the support of the distribution is unbounded. Let $X \sim P$. By Jensen's inequality and $(4,2)$-hypercontractivity, we have

$$\mathbb{E}\,\|X\|_2^4 = p^2\,\mathbb{E}\left[\left(\sum_{j=1}^p \frac{1}{p}X_j^2\right)^2\right] \leq p^2\,\mathbb{E}\left[\sum_{j=1}^p \frac{1}{p}\left(X_j^2\right)^2\right] = p\,\mathbb{E}\left[\sum_{j=1}^p X_j^4\right] \leq \sigma_{x,4}^4 p^2\kappa_u^2,$$

since for each j, we have $\mathbb{E}[X_j^4] = \mathbb{E}[(e_j^\top X)^4] \leq \sigma_{x,4}^4\|\Sigma\|_2^2$, where e_j is the canonical basis vector. Applying Markov's inequality, we then obtain

$$\mathbb{P}\{\|X\|_2 > r\sigma_{x,4}\sqrt{\kappa_u}\epsilon^{-1/4}\sqrt{p}\} \leq \frac{\mathbb{E}\,\|X\|_2^4}{r^4\sigma_{x,4}^4\kappa_u^2\epsilon^{-1}p^2} \leq \frac{\epsilon}{r^4},$$

where $r \geq 2$ is the constant to be specified below. Let $\mathcal{E}_r = \{x : \|x\|_2 \leq r\sigma_{x,4}\sqrt{\kappa_u}\epsilon^{-1/4}\sqrt{p}\}$. Applying a Chernoff bound, we see that with probability at least $1-\exp(-cn\epsilon)$, at most $\frac{n\epsilon}{2}$ points lie outside $\mathcal{E}_r$, where we take r to be a sufficiently large constant. Let P_r be the distribution of P conditioned on $\mathcal{E}_r$. Simply ignoring the points that lie outside $\mathcal{E}_r$, we will only focus on points that come from the distribution P_r and incur an additional failure probability of $\exp(-cn\epsilon)$.

Let $y_1,\ldots,y_m$ be m i.i.d. points from P_r, where $m \geq n\left(1-\frac{\epsilon}{2}\right)$. It suffices to show that any subset of $\{y_1,\ldots,y_m\}$ of size at least $\left(1-\frac{\epsilon}{2}\right)m$ satisfies the desired conclusion. This is exactly what was considered in the first part of the proof, up to constant factors; thus, it remains to show that the distribution P_r satisfies $(4,2)$-hypercontractivity and has an appropriately bounded second moment matrix.

Let $Z_r \sim P_r$ and $X \sim P$. For any $v \in \mathcal{S}^{p-1}$, we have $\mathbb{E}(v^\top Z)^2 \leq \mathbb{E}(v^\top X)^2$. We now look at the lower bound:

$$\mathbb{P}(X \in \mathcal{E}_r)\,\mathbb{E}[(v^\top Z)^2] = \mathbb{E}\left[(v^\top X)^2\,\mathbb{I}_{X\in\mathcal{E}_r}\right]$$

$$
\begin{aligned}
&= \mathbb{E}(v^\top X)^2 - \mathbb{E}[(v^\top X)^2 \mathop{\mathbb{I}}_{X \in \mathcal{E}_r^c}] \\
&\geq \mathbb{E}(v^\top X)^2 - \sqrt{\mathbb{E}[(v^\top X)^4]}\sqrt{\mathbb{P}(X \notin \mathcal{E}_r)} \\
&\geq \mathbb{E}(v^\top X)^2 - \sigma_{x,4}^2 \mathbb{E}(v^\top X)^2 \sqrt{\epsilon r^{-4}} \\
&\geq \mathbb{E}[(v^\top X)^2](1 - \sigma_{x,4}^2 \sqrt{\epsilon} r^{-2}).
\end{aligned}
$$

This shows that $\mathbb{E}(v^\top Z)^2 \geq 0.99\kappa_l$, when $\epsilon \lesssim \kappa_l^2 r^4 \sigma_{x,4}^{-4}$. It also shows that P_r satisfies $(4,2)$-hypercontractivity, as follows:

$$
\left(\mathbb{E}(v^\top Z_r)^4\right)^{1/4} \leq \left(\mathbb{E}(v^\top X)^4\right)^{1/4} \leq \sigma_{x,4}\left(\mathbb{E}(v^\top X)^2\right)^{1/2} \leq \frac{\sigma_{x,4}}{\left(1 - \sigma_{x,4}^2\sqrt{\epsilon} r^{-2}\right)^{1/2}}\left(\mathbb{E}(v^\top Z)^2\right)^{1/2}.
$$

Thus, when $\epsilon \lesssim r^4 \sigma_{x,4}^{-4}$, we see that P_r satisfies $(4,2)$-hypercontractivity with $\sigma'_{x,4} \leq 2\sigma_{x,4}$. Finally, we note that P_r might not be centered, but the means of P_r and P differ by at most $\sigma_{x,4}\sqrt{\kappa_u}\epsilon^{3/4}$ in the Euclidean norm: for any unit vector $v \in \mathcal{S}^{p-1}$, we have

$$
\begin{aligned}
|\,\mathbb{E}[v^\top Z]| &\leq |2\,\mathbb{P}(X \in \mathcal{E}_r)\,\mathbb{E}[v^\top Z]| \\
&= 2\left|\mathbb{E}\left[v^\top X \mathop{\mathbb{I}}_{X \in \mathcal{E}_r}\right]\right| \\
&= 2\left|\mathbb{E}[v^\top X] - \mathbb{E}[(v^\top X) \mathop{\mathbb{I}}_{X \in \mathcal{E}_r^c}]\right| \\
&= 2\left|\mathbb{E}[(v^\top X) \mathop{\mathbb{I}}_{X \in \mathcal{E}_r^c}]\right| \\
&\leq 2\left(\mathbb{E}[(v^\top X)^4]\right)^{1/4}(\mathbb{P}(X \notin \mathcal{E}_r))^{3/4} \\
&\leq 2\sigma_{x,4}\sqrt{\kappa_u}\epsilon^{3/4} r^{-3},
\end{aligned}
$$

using the facts that $\mathbb{P}\{X \in \mathcal{E}_r\} \geq \frac{1}{2}$ and $\mathbb{P}\{X \notin \mathcal{E}_r\} \leq \frac{\epsilon}{r^4}$. The proof now follows from the bounded support setting considered above, which allows the norm of the mean to be as large as $\sigma_{x,4}\sqrt{\kappa_u}\epsilon^{-1/4}$. □

Proposition C.6.3. *Consider the setting of Theorem C.1.3 with $k = 4$. Let $\epsilon < c^*$, where c^* is a small enough constant. Let C be any large constant. Suppose $n = \Omega\left(\frac{p \log p}{\epsilon}\right)$. Then for any $\tau = O(\exp(-\Omega(n\epsilon)))$, with probability at least $1 - \tau$, there exists a set $S_1 \subseteq S$ such that*

(i) $|S_1| \geq (1 - \epsilon)n$,

(ii) S_1 *is* (ϵ_1, δ_1)*-stable, where* $\epsilon_1 = C\epsilon$ *and* $\delta_1 = O\left(\sqrt{\frac{p \log p}{n}} + \sigma_{x,4}\epsilon^{3/4} + \sigma_{x,4}\sqrt{\frac{\log(1/\tau)}{n}}\right)$*, and*

(iii) $\frac{\delta_1^2}{\epsilon_1} < 0.01$.

Moreover, let T be an ϵ'-corrupted set version of S, where $\epsilon' \le \epsilon$. Let T_1 be the output of the filter algorithm with input T and ϵ. Then with probability at least $1 - 2\tau$, the set T_1 satisfies

(i) $|T_1| \ge (1 - c_1\epsilon)n$,

(ii) T_1 *is* (ϵ_2, δ_2)*-stable, where* $\epsilon_2 = c_2 C\epsilon$ *and* $\delta = O\left(\sqrt{\frac{p \log p}{n}} + \sigma_{x,4}\epsilon^{3/4} + \sigma_{x,4}\sqrt{\frac{\log(1/\tau)}{n}}\right)$, *and*

(iii) $\frac{\delta_2^2}{\epsilon_2} < 0.05$.

Proof. We will show that these statements are consequences of Theorems C.1.1 and C.1.3.

Fix the constant C, the desired premultiplier in the stability results. Let $\epsilon_3 > 0$ be a value to be decided later, and let τ be such that $\frac{\log(1/\tau)}{n} \le c_1\epsilon_3$. Suppose ϵ_3 is such that $\epsilon := C_1\left(\epsilon_3 + \frac{\log(1/\tau)}{n}\right)$ is the parameter in Theorem C.1.3. Applying Theorem C.1.3, we see that with probability $1 - \tau$, there exists a $(C\epsilon, \delta_1)$-stable set $S' \subseteq S$, with $|S'| \ge (1 - \epsilon)|S|$ and $\delta_1 = O\left(\sqrt{\frac{p \log p}{n}} + \sigma_{x,4}\epsilon_3^{3/4} + \sigma_{x,4}\sqrt{\frac{\log(1/\tau)}{n}}\right)$, where the premultiplier depends on C.

Note that

$$\begin{aligned}\frac{\delta_1^2}{\epsilon_1} &\lesssim \frac{p \log p}{n\epsilon} + \sigma_{x,4}^2\epsilon_3^{1/2} + \sigma_{x,4}^2\frac{\log(1/\tau)}{n\epsilon} \\ &\lesssim \frac{p \log p}{n\epsilon} + \sigma_{x,4}^2\sqrt{\epsilon} + \sigma_{x,4}^2\frac{c_1}{C_1}.\end{aligned}$$

The last expression can be made less than 0.01 by choosing $n = \Omega\left(\frac{p \log p}{\epsilon}\right)$, restricting ϵ (and thus ϵ_3) to be less than a small enough constant c^*, and choosing c_1 to be small enough. The last condition yields that the failure probability can be made as small as $\exp(-\Omega(n\epsilon))$. This completes the proof of the first statement. Moreover, the bound 0.01 was arbitrary and can be made as small as required under qualitatively similar constraints.

For the second part, we assume that the constant C is large enough for Theorem C.1.1 to succeed. By the first part, we know that with probability at least $1 - \exp(-\Omega(n\epsilon))$, there exist $S_1 \subseteq S$ such that $|S_1| \ge (1 - \epsilon)|S|$ and S_1 is $(C\epsilon, \delta_1)$-stable. Theorem C.1.1 then implies that with probability at least $1 - O(\exp(-\Omega(n\epsilon)))$, the output of the filter algorithm T_1 satisfies $|T_1| \ge (1 - c_1\epsilon)n$ and is (ϵ_2, δ_2)-stable, where $\epsilon_2 = c_2 C\epsilon$ and $\delta_2 = c_3\delta_1$. It remains to check that $\frac{\delta_2^2}{\epsilon_2} < 0.05$. Note that $\frac{\delta_2^2}{\epsilon_2} = \frac{c_3^2}{c_2 C} \cdot \frac{\delta^2}{\epsilon}$. Since c_3, c_2 and C are constants, we can make $\frac{\delta_2^2}{\epsilon_2} < 0.05$ by taking $\frac{\delta_1^2}{\epsilon} < 0.05 \cdot \frac{c_2 C}{c_3^2}$ in the first part. □

Proposition C.6.4. *Let $\{x_1, \ldots, x_n\}$ be an (ϵ, δ)-stable set with respect to μ and σ^2. Then for any unit vector v and any $S' \subseteq [n]$ such that $|S'| \leq \epsilon n$, we have*

$$\frac{1}{n} \sum_{i \in S'} ((x_i - \mu)^\top v)^2 \leq \frac{3\sigma^2 \delta^2}{\epsilon}. \tag{C.4}$$

Proof. Without loss of generality, we assume that $\mu = 0$ and $\sigma^2 = 1$. By the stability assumption, we have the inequality

$$\frac{1}{n} \sum_{i \in [n]} (x_i^\top v)^2 \leq 1 + \frac{\delta^2}{\epsilon}.$$

Furthermore, using the lower bound on eigenvalues over the set $[n] \setminus S'$, we have

$$\frac{1}{|[n] \setminus S'|} \sum_{i \in [n] \setminus S'} (x_i^\top v)^2 \geq 1 - \frac{\delta^2}{\epsilon}.$$

Combining the inequalities, we obtain

$$\begin{aligned}
\frac{1}{n} \sum_{i \in S'} (x_i^\top v)^2 &= \frac{1}{n} \sum_{i \in [n]} (x_i^\top v)^2 - \frac{|[n] \setminus S'|}{n} \frac{1}{|[n] \setminus S'|} \sum_{i \in [n] \setminus S'} (x_i^\top v)^2 \\
&\leq \left(1 + \frac{\delta^2}{\epsilon}\right) - (1 - \epsilon)\left(1 - \frac{\delta^2}{\epsilon}\right) \\
&= \frac{2\delta^2}{\epsilon} + \epsilon - \delta^2 \leq \frac{3\delta^2}{\epsilon},
\end{aligned}$$

where we use the fact that $\epsilon \leq \delta$. □

Proposition C.6.5. *Let $\{x_1, \ldots, x_n\}$ be an (ϵ, δ)-stable set with respect to μ and σ^2. Then for any unit vector v and any $S' \subseteq [n]$ such that $|S'| \leq \epsilon n$, we have*

$$\frac{1}{n} \sum_{i \in S'} |(x_i - \mu)^\top v| \leq 2\sigma\delta. \tag{C.5}$$

Proof. Without loss of generality, we assume that $\mu = 0$ and $\sigma^2 = 1$. By Proposition C.6.4, we have

$$\frac{1}{n} \sum_{i \in S'} (x_i^\top v)^2 \leq \frac{4\delta^2}{\epsilon}.$$

Applying the Cauchy-Schwarz inequality, we then have

$$\frac{1}{|S'|}\sum_{i\in S'}|x_i^\top v| \le \sqrt{\frac{1}{|S'|}\sum_{i\in S'}|x_i^\top v|^2} \le \sqrt{\frac{n}{|S'|}\frac{4\delta^2}{\epsilon}}.$$

Hence, we obtain

$$\frac{1}{n}\sum_{i\in S'}|x_i^\top v| = \frac{|S'|}{n}\frac{1}{|S'|}\sum_{i\in S'}|x_i^\top v| \le \frac{|S'|}{n}\sqrt{\frac{n}{|S'|}\frac{4\delta^2}{\epsilon}} = \sqrt{\frac{|S'|}{n}\frac{4\delta^2}{\epsilon}} \le 2\delta.$$

□

Proposition C.6.6. *Let $\{x_1,\dots,x_n\}$ be an (ϵ,δ)-stable set with respect to μ and σ^2. Let $a_1,\dots,a_n$ be scalars and suppose $\max_{1\le i\le n}|a_i| \le a$. Then for any $S' \subseteq [n]$ such that $|S'| \le \epsilon n$, we have*

$$\left\|\frac{1}{n}\sum_{i\in S'} a_i(x_i-\mu)\right\|_2 \le 2a\sigma\delta. \tag{C.6}$$

Proof. Without loss of generality, we assume that $\mu = 0$ and $\sigma^2 = 1$. We have

$$\left\|\frac{1}{n}\sum_{i\in S'} a_i x_i\right\|_2 = \frac{1}{n}\sup_{v\in\mathcal{S}^{p-1}}\sum_{i\in S'} a_i x_i^\top v \le \frac{1}{n}\sup_{v\in\mathcal{S}^{p-1}}\sum_{i\in S'}|a_i||x_i^\top v| \le \frac{a}{n}\sup_{v\in\mathcal{S}^{p-1}}\sum_{i\in S'}|x_i^\top v| \le 2a\delta, \tag{C.7}$$

where the last step uses Proposition C.6.5. □

C.7 Huber Regression

In this appendix, we provide additional proof details for the results in Section 5.3.

C.7.1 Selection of γ

In this section, we discuss how to estimate an appropriate tuning parameter γ from the data.

C.7.1.1 Random Design and Asymmetric Noise

We first consider the setting of random design and asymmetric noise, discussed in Section 5.3.2. A natural approach is to estimate the scale of the noise distribution based on residuals $y_i - x_i^\top \widehat{\beta}_0$ calculated from an initial estimate $\widehat{\beta}_0$ of β^*. Indeed, the estimate $\widehat{\beta}_0$ can be quite rough, since only need to estimate the scale of the noise up to a constant factor. Based on these observations, consider the following procedure:

1. Split the sample into two equal parts.
2. Using the first part, compute $\widehat{\beta}_0$ via the LAD estimator (cf. Section 5.5 below).
3. Using the second part, compute the symmetrized data points $\{(x_i', y_i')\}_{i=1}^{\lfloor n/2 \rfloor}$ defined as in the first step of Algorithm 11. Then compute the residuals $w_i' = y_i' - (x_i')^\top \widehat{\beta}_0$.
4. Define $\widehat{\gamma}$ to be twice the $\left(1 - \frac{c^*}{4}\right)^{\text{th}}$ empirical quantile of the $|w_i'|$'s.

Note that by our assumptions on the original data set, the sample-splitting step yields two sets of i.i.d. points. Thus, we may use Theorem 5.5.1 to show that $\|\widehat{\beta}_0 - \beta^*\|_2 = O(\kappa)$ if we assume $\mathbb{E}\,|z_i| = \kappa < \infty$. Altogether, we can show that our procedure yields an estimator $\widehat{\gamma}$ such that $\mathbb{P}\left(|Z_1 - Z_2| \geq \widehat{\gamma}/2\right) \leq c^*$ (where Z_1 and Z_2 are fresh i.i.d. draws from the distribution of the z_i's) and $\widehat{\gamma} = O(\mathbb{E}\,|z_i|)$, with high probability. Although other methods for choosing an initial estimator $\widehat{\beta}_0$ would also work, we suggest using the LAD estimator for initialization, since it is tuning parameter-free.

To prove that this method works, we use the result of Theorem 5.5.1, as well as the following lemma, where we denote $\epsilon = c^*$ for notational brevity.

Lemma C.7.1. *Let $S = \{(x_1, y_1), \ldots, (x_{2n}, y_{2n})\}_{i=1}^{2n}$ be i.i.d. points from the linear model $y_i = x_i^\top \beta^* + z_i$, where the covariates are centered and isotropic, and the noise is independent of the covariates and satisfies $\mathbb{E}\,|z_i| = \kappa < \infty$. Let $\widehat{\beta}_0$ be an estimator independent of S such that $\|\widehat{\beta}_0 - \beta^*\|_2 = O(\kappa)$. Then the sample-splitting estimator $\widehat{\gamma}$ with $\epsilon = c^*$ satisfies*

(i) $\mathbb{P}\left(|Z_1 - Z_2| \geq \frac{\widehat{\gamma}}{\sqrt{2}}\right) < \epsilon$, *and*

(ii) $|\widehat{\gamma}| = O\left(\frac{\kappa}{\epsilon}\right)$,

with probability at least $1 - 2\exp(-\Omega(n\epsilon^2))$.

Proof. Let $\beta_1 = \beta^* - \widehat{\beta}_0$. Note that conditioned on $\widehat{\beta}_0$, the pairs $\{(x_i', w_i')\}_{i=1}^{\lfloor n/2 \rfloor}$ are i.i.d. draws from the linear model

$$w_i' = (x_i')^\top \beta_1 + z_i', \tag{C.8}$$

where $z_i' \stackrel{d}{=} \frac{z_1 - z_2}{\sqrt{2}}$ is the symmetrized version of the error variables.

Let x', w', and z' denote generic random variables with the same distributions as x_i', w_i', and z_i', respectively. Note that x' is centered and isotropic, and z' is symmetric with $\mathbb{E}\,|z'| \le \sqrt{2}\kappa$. By the triangle inequality, we therefore have

$$\mathbb{E}\,|w'| \le \mathbb{E}\,|(x')^\top \beta_1| + \mathbb{E}\,|z'| \le \sqrt{\mathbb{E}\,((x')^\top \beta_1)^2} + \mathbb{E}\,|z'| \le \|\beta_1\|_2 + \sqrt{2}\kappa = O(\kappa),$$

using the fact that x' is isotropic and $\|\beta_1\|_2 = O(\kappa)$ by assumption.

Now let F_n denote the empirical cdf of the $|w_i'|$'s, so $F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(|w_i'| \le t)$. Define the event

$$\mathcal{E} := \left\{ \sup_{t \in \mathbb{R}} |F_n(t) - \mathbb{P}(|w'| \le t)| \le \frac{\epsilon}{8} \right\}.$$

By the Dvoretzky-Kiefer-Wolfowitz inequality [Mas90], we know that $\mathbb{P}(\mathcal{E}) \ge 1 - 2\exp(-n\epsilon^2/32)$. Note that by definition, we have $\frac{\widehat{\gamma}}{2} = \inf\left\{t : F_n(t) \ge 1 - \frac{\epsilon}{4}\right\}$. On the event $\mathcal{E}$, we therefore have

$$\mathbb{P}\left(|w'| \ge \frac{\widehat{\gamma}}{2}\right) \le \frac{3\epsilon}{8}. \tag{C.9}$$

Furthermore, since both z' and $(x')^\top \beta_1$ are symmetric random variables, Lemma C.4.7 applied to the linear model (C.8) gives us

$$\mathbb{P}\left(|z'| \ge \frac{\widehat{\gamma}}{2}\right) \le 2\,\mathbb{P}\left(|w'| \ge \frac{\widehat{\gamma}}{2}\right) \le \frac{3\epsilon}{4} < \epsilon,$$

which is part (i).

We now show that $|\widehat{\gamma}| \le \frac{8\,\mathbb{E}\,|w'|}{\epsilon}$ on the event $\mathcal{E}$. Suppose the contrary. By Markov's inequality, we would have

$$\mathbb{P}\left(|w'| \ge \frac{\widehat{\gamma}}{2}\right) \le \mathbb{P}\left(|w'| \ge \frac{4\,\mathbb{E}\,|w'|}{\epsilon}\right) \le \frac{\epsilon}{4},$$

which contradicts inequality (C.9). Therefore, we must have $\widehat{\gamma} = O\left(\frac{\mathbb{E}\,|w'|}{\epsilon}\right) = O\left(\frac{\kappa}{\epsilon}\right)$, as wanted. □

C.7.1.2 Adversarial Contamination

We now consider the setting of Section 5.3.3. Since the adversarial contamination mechanism might create dependencies between data points, the analysis of a sample-splitting

algorithm to estimate an appropriate parameter γ from the data, as in the previous subsection, becomes more complicated. A covering argument akin to the one employed in the proof of Theorem 5.6.2 below could be used instead, albeit at the price of a slightly worse error rate. Another approach would be to tune the Huber parameter using Lepski's method [Lep91; Bir01b], at the expense of a slightly worse error probability due to a union bound over a grid of parameter values. As noted in Remark 5.3.3, if the $(k')^{\text{th}}$ moment of the noise distribution is finite and known, Markov's inequality implies that we can set $\gamma = \Omega((\mathbb{E}\,|z_1 - z_2|^{k'})^{1/k'})$, for any positive k'.

C.7.2 Proof of Theorem 5.3.1

We will follow the proof structure of Sun et al. [SZF20]. The proof relies on the fact that $\mathcal{L}_\gamma(\beta)$ is a convex function. We first show (Lemma C.7.2) that the gradient at β^* is small, and then show (Lemma C.7.3) that the loss function is strongly convex in a sufficiently large ball around β^*. Combining these two observations, we conclude that β^* is close to the empirical minimizer, $\widehat{\beta}_{H,\gamma}$. Our rates are substantially tighter than those of Sun et al. [SZF20] due to the improved guarantees of Lemmas C.7.2 and C.7.3 in comparison to the results in that paper.

We now state and prove the two supporting lemmas:

Lemma C.7.2. *Consider the setting of Theorem 5.3.1. With probability at least $1 - \tau$, the gradient of the loss function satisfies*

$$\|\nabla \mathcal{L}_\gamma(\beta^*)\|_2 \lesssim \gamma\sqrt{U}\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log 1/\tau}{n}}\right).$$

Proof. We first note that the gradient at β^* has a simple structure:

$$\nabla \mathcal{L}_\gamma(\beta^*) = -\frac{1}{n}\sum_{i=1}^{n} \psi_\gamma(y_i - x_i^\top \beta^*)x_i = -\frac{1}{n}\sum_{i=1}^{n} \psi_\gamma(z_i)x_i.$$

For brevity, we define $W := \nabla \mathcal{L}_\gamma(\beta^*)$ and $W_i = \psi_\gamma(z_i)$. Note that since the z_i's are symmetric, the W_i's are i.i.d. bounded random variables and $\mathbb{E}(W) = 0$.

We will now show that W has sub-Gaussian concentration around 0. Let v be any unit vector. Since the W_i's are bounded by γ, the sub-Gaussian norm of $v^\top Z$ can be

bounded using Proposition 2.6.1 of Vershynin [Ver18]:

$$\|v^\top W\|_{\psi_2} \lesssim \frac{1}{n}\sqrt{\sum_{i\in[n]} \gamma^2 (v^\top x_i)^2} \leq \gamma\sqrt{\frac{U}{n}},$$

where the last step uses weak stability. Therefore, W is an $O\left(\gamma\sqrt{\frac{U}{n}}\right)$-sub-Gaussian vector, so again using the results of Vershynin [Ver18] for the concentration of a sub-Gaussian vector, we have

$$\|W\|_2 = \|W - \mathbb{E}\, W\|_2 \lesssim \gamma\sqrt{\frac{U}{n}}\left(\sqrt{p} + \sqrt{\log\frac{1}{\tau}}\right),$$

with probability at least $1-\tau$. □

Lemma C.7.3. *Consider the setting in Theorem 5.3.1. Let r, U, τ, and γ be such that*

$$C_2\left(\frac{r\sqrt{U}}{\gamma} + \mathbb{P}\left(|z_i| \geq \frac{\gamma}{2}\right) + \frac{\log(1/\tau)}{n}\right) \leq \epsilon,$$

for a constant $C_2 > 0$. Then with probability at least $1-\tau$, the loss function $\mathcal{L}_\gamma(\beta)$ is L-strongly convex in the ball $\{\beta : \|\beta - \beta^*\|_2 \leq r\}$.

Proof. First note that $\mathcal{L}_\gamma(\beta)$ is a convex function. The Hessian of $\mathcal{L}_\gamma$ is not defined due to the fact that the Huber loss is not twice differentiable at γ. However, if we define the matrix

$$H_n(\beta) := \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \mathbb{I}\left(|y_i - x_i^\top\beta| < \gamma\right),$$

it follows that the strong convexity parameter of $\mathcal{L}(\beta)$ is at least $\lambda_{\min}(H_n)$ (see Lemma C.4.9).

Let $W := \sup_{\beta:\|\beta-\beta^*\|_2\leq r} \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left(|y_i - x_i^\top\beta| \geq \gamma\right)$ and define the event $\mathcal{E} := \{W < \epsilon\}$. By the weak stability property, we are guaranteed that on the event $\mathcal{E}$, we have $\lambda_{\min}(H_n(\beta)) \geq L$ for any β such that $\|\beta - \beta^*\|_2 \leq r$.

In the remainder of the proof, we will show that the event $\mathcal{E}$ holds with high probability. We first note that W can be bounded from above, as follows:

$$W = \sup_{\beta:\|\beta-\beta^*\|\leq r} \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left(|y_i - x_i^\top\beta| \geq \gamma\right)$$

$$\leq \sup_{\beta:\|\beta-\beta^*\|\leq r} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(|x_i^\top(\beta-\beta^*)| \geq \frac{\gamma}{2}\right) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(|z_i| \geq \frac{\gamma}{2}\right). \tag{C.10}$$

We can deterministically bound the first term using weak stability. Using the fact that for $x \geq 0$ and $y > 0$, the inequality $\mathbb{I}(x \geq y) \leq \frac{x}{y}$ holds, we obtain the following bound for all β such that $\|\beta - \beta^*\|_2 \leq r$:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(|x_i^\top(\beta-\beta^*)| \geq \frac{\gamma}{2}\right) &\leq \frac{2}{\gamma}\frac{\sum_{i=1}^{n}|x_i^\top(\beta-\beta^*)|}{n} \leq \frac{2}{\gamma}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|x_i^\top(\beta-\beta^*)|^2} \\ &\leq \frac{2}{\gamma}\sqrt{U\|\beta-\beta^*\|_2^2} \leq \frac{2r\sqrt{U}}{\gamma},\end{aligned}$$

where we also use weak stability and the Cauchy-Schwarz inequality. Altogether, we obtain

$$W \leq \frac{2r\sqrt{U}}{\gamma} + \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(|z_i| \geq \frac{\gamma}{2}\right). \tag{C.11}$$

Now let $W' := \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(|z_i| \geq \frac{\gamma}{2}\right)$. Note that

$$\mathbb{E}\,W' = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,\mathbb{I}\left(|z_i| \geq \frac{\gamma}{2}\right) = \mathbb{P}\left(|z_i| \geq \frac{\gamma}{2}\right).$$

Note that W' is an empirical mean of indicator random variables. Thus, applying a Chernoff bound (cf. Lemma C.4.1), we obtain

$$W' \lesssim \mathbb{E}\,W' + \frac{\log(1/\tau)}{n},$$

with probability at least $1-\tau$. Overall, we obtain the following bound on W: with probability at least $1-\tau$,

$$W \lesssim \frac{r\sqrt{U}}{\gamma} + \mathbb{P}\left(|z_1| \geq \frac{\gamma}{2}\right) + \frac{\log(1/\tau)}{n}. \tag{C.12}$$

Therefore, the event $\mathcal{E}$ (and thus, the desired lower bound on H_n) holds with probability $1-\tau$, as long as the right-hand side of inequality (C.12) is less than ϵ. □

With the help of Lemmas C.7.2 and C.7.3, we are ready to prove the theorem. Throughout the remainder of the proof, let $\widehat{\beta} = \widehat{\beta}_{H,\gamma}$.

We first verify the conditions for Lemma C.7.3. By assumption, we have $\frac{C_2\log(1/\tau)}{n} \leq$

$\frac{\epsilon}{3}$ and $C_2\,\mathbb{P}\left(|z_i| \geq \gamma/2\right) \leq \frac{\epsilon}{3}$. Therefore, for all $r \leq \frac{\epsilon\gamma}{3C_2\sqrt{U}} := r^*$, the condition of Lemma C.7.3 is satisfied, and the function $\mathcal{L}_\gamma$ is L-strongly convex in the region $\{\beta : \|\beta - \beta^*\|_2 \leq r^*\}$.

For an $\eta \in (0, 1]$, let $\widehat{\beta}_\eta$ be defined as $\widehat{\beta}_\eta := \beta^* + \eta(\widehat{\beta} - \beta^*)$, and let $\eta_* \in (0, 1]$ be the largest η such that $\|\widehat{\beta}_\eta - \beta^*\|_2 \leq r^*$. Using the convexity of $\mathcal{L}_\gamma(\beta)$ with Lemma C.4.8 and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\langle \widehat{\beta}_{\eta^*} - \beta^*, \nabla\mathcal{L}_\gamma(\widehat{\beta}_{\eta^*}) - \nabla\mathcal{L}_\gamma(\beta^*)\rangle &\leq \eta_*\langle \widehat{\beta} - \beta^*, \nabla\mathcal{L}_\gamma(\widehat{\beta}) - \nabla\mathcal{L}_\gamma(\beta^*)\rangle \\ &\leq \eta_*\|\nabla\mathcal{L}_\gamma(\beta^*)\|_2\|\widehat{\beta} - \beta^*\|_2, \end{aligned} \tag{C.13}$$

where we use the fact that $\nabla\mathcal{L}_\gamma(\widehat{\beta}) = 0$. Using the L-strong convexity of $\mathcal{L}_\gamma$ in the ball of radius r^* (cf. Lemma C.4.9) and inequality (C.13), we obtain

$$\eta_*\|\nabla\mathcal{L}_\gamma(\beta^*)\|_2\|\widehat{\beta} - \beta^*\|_2 \geq \langle \widehat{\beta}_{\eta^*} - \beta^*, \nabla\mathcal{L}_\gamma(\widehat{\beta}_{\eta^*}) - \nabla\mathcal{L}_\gamma(\beta^*)\rangle \geq L\|\widehat{\beta}_{\eta^*} - \beta^*\|_2^2.$$

We now use Lemma C.7.2 and the fact that $\|\widehat{\beta}_{\eta^*} - \beta^*\|_2 = \eta^*\|\widehat{\beta} - \beta^*\|_2$ to obtain the following bound:

$$\|\widehat{\beta}_{\eta^*} - \beta^*\|_2 \leq \frac{1}{L}\|\nabla\mathcal{L}_\gamma(\beta^*)\|_2 \leq C\frac{\gamma\sqrt{U}}{L}\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}}\right) := R_n. \tag{C.14}$$

Note that $\frac{R_n}{r^*} = \frac{3CC_2U}{\epsilon L}\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}}\right)$, so under the sample complexity assumption $n = \Omega\left(\left(p + \log\left(\frac{1}{\tau}\right)\right)\frac{U^2}{L^2\epsilon^2}\right)$, we have $R_n \leq r^*$, implying in particular that $\eta^* = 1$ and $\widehat{\beta} = \widehat{\beta}_{\eta^*}$ satisfies the stated error bound.

The statement about L-strong convexity follows from the triangle inequality, since for sufficiently large n, we have $\|\widehat{\beta} - \beta^*\|_2 \leq R_n \leq \frac{r^*}{2}$, so the function $\mathcal{L}_\gamma$ is L-strongly convex in a ball of radius $\frac{r^*}{2}$ around $\widehat{\beta}$.

C.7.3 Proof of Theorem 5.3.4

We first note that by taking pairwise differences, we reduce our case to the symmetric noise setting analyzed in Section 5.3.1: Given $2n$ data points, Algorithm 11 creates a data set $\{(x_i', y_i')\}_{i=1}^n$ satisfying the linear model $y_i' = (x_i')^\top\beta^* + z_i'$, where $z_i' = \frac{z_i - z_{n+i}}{\sqrt{2}}$. Note that the new covariates still satisfy $\mathbb{E}\,x_i' = 0$ and $\mathbb{E}\,x_i'(x_i')^\top = I$. Importantly, the errors are now drawn from a symmetric distribution.

Let S_1 be the set returned by the filter algorithm with cardinality $\Omega(n)$, and define

the event

$$\mathcal{E} = \{S_1 \text{ satisfies weak stability with } \epsilon = \Omega(1), L = \Omega(1), \text{ and } U = O(1)\}.$$

We first give the proof of the theorem statement on the event $\mathcal{E}$. Since the noise is symmetric and independent of the covariates (thus also of $\mathcal{E}$), we have $\mathbb{P}(|z_i'| \geq \gamma/2) = O(\epsilon)$, so Theorem 5.3.1 applies and gives the desired result. In the rest of the proof, we will show that $\mathcal{E}$ holds with probability $1 - \exp(-\Omega(n)) \geq 1 - \tau$.

By Proposition C.6.1, if S_1 is (ϵ_1, δ_1)-stable, then it also satisfies weak stability with $\epsilon = \epsilon_1, L = (1-\epsilon_1)\left(1 - \frac{\delta_1^2}{\epsilon_1}\right)$, and $U = 1 + \frac{\delta_1^2}{\epsilon_1}$. Therefore, it suffices to show that S_1 is (ϵ_1, δ_1)-stable such that $\epsilon_1 = \Omega(1)$ and (say) $\frac{\delta_1^2}{\epsilon_1} < 0.5$. By Proposition C.6.3, we know that if $\epsilon' < c_*$ and $n = \Omega\left(\frac{p \log p}{\epsilon'}\right)$, then with probability at least $1 - O(\exp(-\Omega(n\epsilon')))$, the set S_1 is (ϵ_1, δ_1)-stable with $\frac{\delta_1^2}{\epsilon_1} < 0.2$ and $\epsilon_1 = \Omega(\epsilon')$. Therefore, choosing ϵ' to be a small enough constant, say $\frac{c^*}{2}$, we conclude that the event $\mathcal{E}$ holds with probability $1 - O(\exp(-\Omega(n)))$. This requires that $n = \Omega\left(\frac{p \log p}{\epsilon'}\right) = \Omega(p \log p)$, completing the proof.

C.7.4 Proof of Theorem 5.3.6

In the course of this proof, we will need to refer to set functions that take a finite set as the argument and return a value in $\mathbb{R}$. The sets we consider will be of the form $S = \{(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n)\}$, where $u_i \in \mathbb{R}^p$, $v_i \in \mathbb{R}$, and $n \geq 1$. The set functions will be of the following form:

$$F(S) := \sum_{i=1}^{n} f(u_i, v_i),$$

for some $f : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$. For ease of notation, we will use the following convention:

$$F(S) = \sum_{(x,y) \in S} f(x, y).$$

This simplifies notation by avoiding explicit indexing of the elements in the sets being considered. For example, if $S' \subseteq S$, we may express $F(S') = \sum_{(x,y) \in S'} f(x, y)$.

For ease of presentation, we also redefine the algorithm with different notation, as reflected in Algorithm 15.

We state the following technical lemma, which is proved in Appendix C.7.5:

Algorithm 15 Huber Regression - Adversarial Corruption

function HUBER_REGRESSION_WITH_FILTERING($T = \{x'_i, y'_i : i \in [2n]\}, \gamma, \epsilon'_1$)
  for $i \leftarrow 1$ to n **do**
    $(\tilde{x}_i, \tilde{y}_i) \leftarrow \left(\frac{x'_i - x'_{n+i}}{\sqrt{2}}, \frac{y'_i - y'_{n+i}}{\sqrt{2}}\right)$
  end for
  $T_1 \leftarrow \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n$
  $T_2 \leftarrow$ FilteredCovariates(T_1, ϵ'_1)
  $\widehat{\beta} \leftarrow$ HuberRegression(T_2, γ)
  return $\widehat{\beta}$
end function

Lemma C.7.4. *Under the setting of Theorem 5.3.6, with probability at least $1 - 2\tau$, we have the following statements:*

(*i*) *The filtered set of covariates T_2 satisfies weak stability with parameters $\epsilon_1 = \Omega(1), L = \Omega(1)$, and $U = O(1)$.*

(*ii*) *The gradient of the loss function satisfies $\|\nabla \mathcal{L}_\gamma(\beta^*)\|_2 \lesssim \gamma\left(\sqrt{\frac{p \log p}{n}} + \epsilon^{1-1/k} + \sqrt{\frac{\log(1/\tau)}{n}}\right)$.*

(*iii*) *For $r \gtrsim \frac{\epsilon_1 \gamma}{\sqrt{U}}$, $\gamma \gtrsim \frac{\sigma}{\sqrt{\epsilon_1}}$, and $\frac{\log(1/\tau)}{n} \lesssim \epsilon_1$, the function $\mathcal{L}_\gamma$ is L-strongly convex in a ball of radius r around β^*.*

Note that we can then follow the proof of Theorem 5.3.1 exactly, where we replace Lemmas C.7.2 and C.7.3 with statements (ii) and (iii) of Lemma C.7.4 and impose the condition that ϵ is less than a small enough constant.

C.7.5 Proof of Lemma C.7.4

Proof of (i): Recall that T_1 is a set of cardinality n, where we subtract pairs of points in the corrupted data set (and rescale by $\sqrt{2}$). Analogously, we define the set S_1, where we perform pairwise subtraction on the uncorrupted data set S (and rescale by $\sqrt{2}$). It can be shown that T_1 is an (at most) 2ϵ-corrupted version of set S_1, and S_1 is a set of n i.i.d. data points from a linear model, where (i) the covariates are drawn from a centered isotropic distribution with k^{th} moment bounded by $c\sigma_{x,k}$; and (ii) the additive noise is zero-mean, symmetric, independent of the covariates, and of variance σ^2.

By Theorem C.1.3, we know that with probability $1 - \tau$, there exists a set $S_2 \subseteq S_1$ such that $|S_2| \geq (1 - \epsilon'_1)n$ and S_2 is (ϵ_2, δ_2)-stable, where $\epsilon_2 = C\epsilon'_1$ and $\delta_2 \lesssim \sqrt{\frac{p \log p}{n}} + \sigma_{x,k}{\epsilon_1}^{1-1/k} + \sigma_{x,4}\sqrt{\frac{\log(1/\tau)}{n}}$. Here, we take $\epsilon_1 = \frac{p \log p}{n} + 2\epsilon$ and define $\epsilon'_1 = C\left(\epsilon_1 + \frac{\log(1/\tau)}{n}\right)$,

and note that $\epsilon_1, \epsilon_1' = O(1)$ by our assumptions. Recall that T_2 is the output of the filter algorithm on the set T_1 with parameter $\epsilon_1' \geq 2\epsilon$. Since T_1 is an (at most) 2ϵ-corrupted version of S_1, the existence of the stable set S_2, in conjunction with Theorem C.1.1, implies that with probability $1-\tau$: (i) T_2 has cardinality at least $(1-c_2\epsilon_1')n$, and (ii) T_2 is (ϵ_3, δ_3)-stable, where $\epsilon_3 = c_2\epsilon_2$ and $\delta_3 = c_4\delta_2$.

Moreover, by Proposition C.6.2, we know that for ϵ_5 a small enough constant, with probability at least $1 - O(\exp(-\Omega(n\epsilon_5)))$, every $S_3 \subseteq S_1$ with cardinality at least $(1-\epsilon_5)n$ satisfies the inequality $\lambda_{\min}\left(\frac{1}{n}\sum_{(x,y)\in S_3} xx^\top\right) \geq 0.8$. Since the amount of corruption is sufficiently small, we will be able to translate this guarantee to the filtered set T_2.

We now choose $\epsilon_5 \in (0,1)$ to be a small enough constant and choose ϵ_1' sufficiently small (note that the latter is possible for a small enough choice of ϵ and large enough choice of n), so that the following are satisfied simultaneously:

1. Both $\frac{\delta_2^2}{\epsilon_2} = O(1)$ and $\frac{\delta_3^2}{\epsilon_3} = O(1)$: note that
$$\frac{\delta_2^2}{\epsilon_2} \lesssim \frac{\frac{p\log p}{n} + \sigma_{x,k}^2\epsilon_1^{2-2/k} + \sigma_{x,4}^2\frac{\log(1/\tau)}{n}}{\epsilon_1 + \frac{\log(1/\tau)}{n}} \lesssim 1.$$
2. The cardinality of S_2 satisfies $|S_2| \geq (1-\epsilon_1')n \geq \left(1 - \frac{\epsilon_5}{20}\right)n \geq \frac{n}{2}$.
3. The cardinality of T_2 satisfies $|T_2| \geq (1-c_2\epsilon_1')n \geq \left(1 - \frac{\epsilon_5}{20}\right)n \geq \frac{n}{2}$.
4. The inequality $4\epsilon < 4\epsilon_1' \leq \frac{\epsilon_5}{10}$ holds.

We now show that the covariates in T_2 satisfy weak stability with $\epsilon_6 = \frac{\epsilon_5}{3} = \Omega(1)$, $L = \Omega(1)$, and $U = O(1)$. Suppose $T_2' \subseteq T_2$ is such that $|T_2'| \geq (1-\epsilon_6)|T_2|$. Then
$$\frac{1}{|T_2|}\lambda_{\max}\left(\sum_{(x,y)\in T_2'} xx^\top\right) \leq \frac{1}{|T_2|}\lambda_{\max}\left(\sum_{(x,y)\in T_2} xx^\top\right) \leq 1 + \frac{\delta_3^2}{\epsilon_3} = O(1),$$
using the (ϵ_3, δ_3)-stability of T_2, giving the upper bound $U = O(1)$. To obtain the lower bound, note that
$$\begin{aligned}
|T_2' \cap S_1| &\geq |T_2'| - |T_2 \triangle S_1| \\
&\geq |T_2|\left(1 - \frac{\epsilon_5}{3}\right) - 2\epsilon n \\
&\geq n\left(1 - \frac{\epsilon_5}{3}\right)\left(1 - \frac{\epsilon_5}{20}\right) - \frac{\epsilon_5 n}{20} \geq (1-\epsilon_5)n.
\end{aligned}$$

Therefore, $T_2' \cap S_1$ is a subset of S_1 with cardinality at least $(1-\epsilon_5)n$, and we conclude that

$$\frac{1}{|T_2|}\lambda_{\min}\left(\sum_{(x,y)\in T_2'} xx^\top\right) \geq \frac{1}{n}\lambda_{\min}\left(\sum_{(x,y)\in T_2'\cap S_1} xx^\top\right) \geq 0.8.$$

This gives the desired lower bound $L = \Omega(1)$.

Proof of (ii): Using the same strategy as in previous step, we can show that weak stability also holds on S_2 with parameters ϵ_6, $L = \Omega(1)$, and $U = O(1)$. We will use this fact to prove concentration results analogous to Lemmas C.7.2 and C.7.3.

In fact, in the proof of Lemma C.7.2, the only property of the covariates that we leveraged was the fact that they satisfy weak stability with $U = O(1)$. Thus, we can analogously argue that (since the additive noise is independent of the covariates)

$$\left\|\frac{1}{|S_2|}\sum_{(x,y)\in S_2}\nabla\ell_\gamma(y-x^\top\beta^*)\right\|_2 \lesssim \gamma\left(\sqrt{\frac{p}{n}}+\sqrt{\frac{\log(1/\tau)}{n}}\right), \tag{C.15}$$

with probability at least $1-\tau$.

We will now translate this result back to T_2 using the fact that both S_2 and T_2 are stable. Let $\mathcal{L}_\gamma$ denote the Huber loss function with parameter γ applied to the set T_2:

$$\mathcal{L}_\gamma(\beta) = \frac{1}{|T_2|}\sum_{(x,y)\in T_2}\ell_\gamma(y-x^\top\beta).$$

Using the triangle inequality together with the bound (C.15) and the notation $z = y - x^\top\beta^*$, we then obtain

$$\begin{aligned}
\|\nabla\mathcal{L}_\gamma(\beta^*)\|_2 &= \left\|\frac{1}{|T_2|}\sum_{(x,y)\in T_2} x\psi_\gamma(z)\right\|_2 \\
&\leq \left\|\frac{1}{|T_2|}\sum_{(x,y)\in S_2} x\psi_\gamma(z)\right\|_2 + \left\|\frac{1}{|T_2|}\sum_{(x,y)\in S_2\setminus T_2} x\psi_\gamma(z)\right\|_2 \\
&\qquad + \left\|\frac{1}{|T_2|}\sum_{(x,y)\in T_2\setminus S_2} x\psi_\gamma(z)\right\|_2 \\
&\lesssim \left\|\frac{1}{|S_2|}\sum_{(x,y)\in S_2} x\psi_\gamma(z)\right\|_2 + \left\|\frac{1}{|S_2|}\sum_{(x,y)\in S_2\setminus T_2} x\psi_\gamma(z)\right\|_2
\end{aligned}$$

$$+\left\|\frac{1}{|T_2|}\sum_{(x,y)\in T_2\setminus S_2} x\psi_\gamma(z)\right\|_2$$
$$\lesssim \gamma\left(\sqrt{\frac{p}{n}}+\sqrt{\frac{\log(1/\tau)}{n}}+\delta_2+\delta_3\right),$$

with probability at least $1-\tau$, where the last step uses Proposition C.6.6 and the stability of S_2 and T_2. Using the bounds on δ_2 and δ_3 completes the proof.

Proof of (iii): We have shown that with probability at least $1-2\tau$, the sets S_2 and T_2 both satisfy weak stability with ϵ_6, $L=\Omega(1)$, and $U=O(1)$; in addition, statements (1)–(4) hold in the proof of part (i) above. We denote this high-probability event by $\mathcal{E}$, and show that under the additional assumptions, the desired strong convexity statement holds on the event $\mathcal{E}$.

By the same argument used in the proof of Lemma C.7.3, we know that on event $\mathcal{E}$, if r,γ, and τ satisfy the inequality

$$\frac{r\sqrt{U}}{\gamma}+\frac{\sigma^2}{\gamma^2}+\frac{\log(1/\tau)}{n}\lesssim\epsilon_6,$$

then

$$\sup_{\beta:\|\beta-\beta^*\|_2\le r}\frac{1}{|S_2|}\sum_{(x,y)\in S_2}\mathbb{I}\left(|y-x^\top\beta|\ge\gamma\right)\le\frac{\epsilon_6}{10}. \tag{C.16}$$

Crucially, we use the fact that conditioned on the event $\mathcal{E}$ (which is entirely defined in terms of the covariates), the noise random variables $\{z_i=y_i-x_i^\top\beta^*:(x_i,y_i)\in S_2\}$ remain i.i.d.

Now let $W:=\sup_{\beta:\|\beta-\beta^*\|_2\le r}\frac{1}{|T_2|}\sum_{(x,y)\in T_2}\sum_{i=1}^n\mathbb{I}\left(|y-x^\top\beta|\ge\gamma\right)$. Note that

$$W\le\frac{|T_2\setminus S_2|}{|T_2|}+\sup_{\beta:\|\beta-\beta^*\|_2\le r}\frac{1}{|T_2|}\sum_{(x,y)\in S_2}\mathbb{I}\left(|y-x^\top\beta|\ge\gamma\right). \tag{C.17}$$

On the event $\mathcal{E}$, we can bound the first term by

$$\frac{|T_2\setminus S_2|}{|T_2|}\le\frac{|T_1\setminus S_2|}{n/2}\le\frac{2}{n}\left(|T_1\setminus S_1|+|S_1\setminus S_2|\right)\le\frac{2}{n}\left(2\epsilon n+\frac{\epsilon_6 n}{20}\right)\le\frac{2\epsilon_6}{5},$$

where the third inequality uses the fact that $|S_2|\ge(1-\epsilon_5/20)n$, and the last inequality uses the bound $4\epsilon\le\frac{\epsilon_5}{10}=\frac{3\epsilon_6}{10}$, which were established in the proof of part (i). The

second term of inequality (C.17) can be bounded by

$$\sup_{\beta:\|\beta-\beta^*\|_2\le r}\frac{1}{|T_2|}\sum_{(x,y)\in S_2}\mathbb{I}\left(|y-x^\top\beta|\ge\gamma\right)=\frac{|S_2|}{|T_2|}\cdot\sup_{\beta:\|\beta-\beta^*\|_2\le r}\frac{1}{|S_2|}\sum_{(x,y)\in S_2}\mathbb{I}\left(|y-x^\top\beta|\ge\gamma\right)$$
$$\le\frac{n}{n/2}\cdot\frac{\epsilon_6}{10}=\frac{\epsilon_6}{5},$$

using inequality (C.16). Thus,

$$W\le\frac{2\epsilon_6}{5}+\frac{\epsilon_6}{5}<\epsilon_6.$$

Now define the matrix

$$H_n(\beta):=\frac{1}{|T_2|}\sum_{(x,y)\in T_2}xx^\top\mathbb{I}\left(|y-x^\top\beta|<\gamma\right).$$

It follows that the strong convexity parameter of $\mathcal{L}_\gamma(\beta)$ is at least $\lambda_{\min}(H_n)$. Using the fact that T_2 satisfies weak stability and $W\le\epsilon_6$, we conclude that on the event $\mathcal{E}$, we have $\lambda_{\min}(H_n(\beta))\ge L$ for any β such that $\|\beta-\beta^*\|_2\le r$, as wanted.

C.7.6 Proof of Theorem 5.3.11

We will show that conditions analogous to the ones stated in Lemma C.7.4 hold in this setting. As the proof is very similar to the proof in Section C.7.5, we only highlight several arguments which need to be adapted. We use the same notation defined in the previous section.

Condition (i): Since the distribution of the covariates has a bounded covariance, Theorem C.10.1 implies that, with probability at least $1-\tau$, the set S_2 is (ϵ_2,δ_2)-stable, where $\delta_2\lesssim\sqrt{\frac{p\log p}{n}}+\sqrt{\epsilon}+\sqrt{\frac{\log(1/\tau)}{n}}$. Recall that we needed $\frac{\delta_2^2}{\epsilon_2}=O(1)$. This is still satisfied, since $n\gtrsim p\log p$ and $\epsilon+\frac{\log(1/\tau)}{n}<c$, for a sufficiently small positive constant c.

It remains to establish (ϵ,L,U)-weak stability of T_2 with $\epsilon=\Omega(1)$, $L=\Omega(1)$, and $U=O(1)$. Similar to the proof of Lemma C.7.4, the lower bounds on ϵ and L follow from the properties of S_2 which hold by the small ball property of the covariates, as shown in Proposition C.6.2.

Condition (ii): As shown in the proof of Lemma C.7.4, the norm of the gradient is bounded as $\|\nabla\mathcal{L}_\gamma(\beta^*)\|_2\lesssim\gamma\left(\sqrt{\frac{p}{n}}+\sqrt{\frac{\log(1/\tau)}{n}}\right)+\delta_2+\delta_3$. Since $\delta_3=O(\delta_2)$, the bound on

δ_2 established in the previous paragraph suffices.

Condition (iii): This is exactly same as before, because we only used weak stability of the sets S_2 and T_2 to show this result.

C.8 Least Trimmed Squares

In this appendix, we provide additional proof details for the results in Section 5.4.

C.8.1 Proof of Theorem 5.4.2

We begin by stating the following deterministic result, which is implicit in Bhatia et al. [BJKK17]. For completeness, we provide a proof in Appendix C.8.2. Recall the definitions of the SSC and SSS properties from Definition 5.2.7.

Lemma C.8.1. (*Adapted from Lemma 5 of Bhatia et al. [BJKK17]*) *Suppose $y = X\beta^* + z$, where the SSC and SSS parameters of the x_i's, denoted by $\{\lambda_k\}$ and $\{\Lambda_k\}$, respectively, satisfy $\frac{\Lambda_{2m}}{\lambda_n} < \frac{1}{4}$ and $\Lambda_n = O(\lambda_n)$. Suppose $z = w + b^*$, for some vector $w \in \mathbb{R}^n$ and an m-sparse vector $b^* \in \mathbb{R}^n$, and let G and H be numbers such that $G \geq \sup_{S':|S'|\leq 2m} \sqrt{\sum_{i\in S'} w_i^2}$ and $H \geq \|\sum_{i=1}^n x_i w_i\|_2$. Then Algorithm 12, after $J \gtrsim \log_2\left(\frac{\|b^*\|_2}{2G+2H/\sqrt{\lambda_n}}\right)$ iterations, outputs an estimator $\widehat{\beta}$ such that $\|\widehat{\beta} - \beta^*\|_2 \lesssim \frac{G\sqrt{\Lambda_n}+H}{\lambda_n}$.*

Remark C.8.2. *The proof of Lemma C.8.1 actually implies that for any error level $e \gtrsim \frac{G\sqrt{\Lambda_n}+H}{\lambda_n}$, Algorithm 12 is guaranteed to output an estimator satisfying the error bound $\|\widehat{\beta} - \beta^*\|_2 \leq e$ after $J \gtrsim \log_2\left(\frac{\|b^*\|_2}{e}\right)$ iterations. This form of the result is helpful in settings such as Theorem 5.4.2 below, where we can obtain data-driven upper bounds on G and H, and consequently also on the term $\frac{G\sqrt{\Lambda_n}+H}{\lambda_n}$, which hold with high probability. Together with a data-driven upper bound on $\|b^*\|_2$, this provides a calculable lower bound on the number of iterations required for Algorithm 12 to succeed in outputting an estimator with small error.*

Note that the statement of Lemma C.8.1 is deterministic: In Bhatia et al. [BJKK17], it was shown that when the covariates are i.i.d. Gaussian, the SSC and SSS conditions hold with high probability. Our proof of Theorem 5.4.2 essentially proceeds by showing that these conditions hold with high probability for possibly heavy-tailed, adversarially contaminated covariates after applying filtering.

Turning to the proof of the theorem, we will use the notation $z_i := y_i - x_i^\top \beta^*$ and $z_i' := y_i' - (x_i')^\top \beta^*$. Let $m = C_1\left(p\log p + \epsilon n + \log\left(\frac{1}{\tau}\right)\right)$, for a large enough constant $C_1 > 6$

to be chosen later. We will now apply Proposition C.6.3 with $\epsilon_1 = \frac{C_2 m}{n}$, for a constant $C_2 \geq 1$ to be decided later. In order for Proposition C.6.3 to be applicable, we need $\epsilon_1 < c^*$ and $n = \Omega\left(\frac{p \log p}{\epsilon_1}\right)$: For any C_2, the latter condition can be satisfied by choosing C_1 sufficiently large, and then the former condition can be satisfied by restricting ϵ, $\frac{\log(1/\tau)}{n}$, and $\frac{p \log p}{n}$ to be less than sufficiently small constants. Let $T_1 \subseteq T$ be the set of data points corresponding to covariates which survive the filter algorithm, and let $n_1 := |T_1|$. Proposition C.6.3 guarantees that with probability at least $1 - 2\tau$, we have

- $|T_1| \geq (1 - c_1\epsilon_1)n \geq \frac{n}{2}$,
- the covariates of the points in T_1 are (ϵ_2, δ_2)-stable, where

$$\delta_2 = O\left(\sqrt{\frac{p \log p}{n}} + \sigma_{x,4}\epsilon_1^{3/4} + \sigma_4\sqrt{\frac{\log(1/\tau)}{n}}\right)$$

  and $\epsilon_2 = \Theta(\epsilon_1)$, and
- $\frac{\delta_2^2}{\epsilon_2} < 0.05$.

We will now choose C_2 sufficiently large such that $\epsilon_2 n = \Theta(\epsilon_1 n) = \Theta(C_2 m) > 4m$. From here on, we will also assume that ϵ, $\frac{\log(1/\tau)}{n}$, and $\frac{p \log p}{n}$ are bounded such that $4m \leq n$.

We now show that the SSC and SSS parameters of the covariates in T_1 are well-behaved, so that Lemma C.8.1 applies. We will apply the lemma to the model

$$y_i' = (x_i')^\top \beta^* + w_i + b_i^*, \qquad 1 \leq i \leq n_1, \tag{C.18}$$

where for a set $T_2 \subseteq T_1$ to be defined later, we define the vector $w \in \mathbb{R}^{n_1}$ according to

$$w_i := \begin{cases} z_i, & \text{if } (x_i, y_i) \in T_2, \\ 0, & \text{otherwise,} \end{cases}$$

and then simply define $b^* := y_i' - (x_i')^\top \beta^* - w$. Let the SSC and SSS parameters of T_1 be denoted by $\{\lambda_k\}$ and $\{\Lambda_k\}$, respectively. Note that

$$\Lambda_{2m} \leq \Lambda_{\epsilon_2 n/2} \leq \Lambda_{\epsilon_2 n_1} \leq \frac{3n_1\delta_2^2}{\epsilon_2} \leq 0.15 n_1,$$

where we have used Proposition C.6.4 in the third inequality. By the (ϵ_2, δ_2)-stability of T_1, we have $\lambda_{n_1} \geq n_1 \left(1 - \frac{\delta_2^2}{\epsilon_2}\right) \geq 0.9 n_1$. Therefore, $\frac{\Lambda_{2m}}{\lambda_{n_1}} \leq \frac{1}{4}$. Since

$$\Lambda_{n_1} \leq n_1 \left(1 + \frac{\delta_2^2}{\epsilon_2}\right) \leq 1.05 n_1,$$

we also have $\Lambda_{n_1} = O(\lambda_{n_1})$. Thus, the eigenvalue conditions of Lemma C.8.1 are indeed satisfied.

We now turn to the definition of T_2 and show that with this definition, b^* is m-sparse. Let $S_2 \subseteq S$ be the set of $n - \frac{m}{4}$ uncontaminated data points with the smallest values of $|z_i|$. Let F be the cumulative distribution function of $|z_i|$ and let F^{-1} be its generalized inverse, i.e., $F^{-1}(p) = \inf_t \mathbb{P}(|z| \leq t) \geq p$. Note that by a Chernoff bound, we have

$$\left| \left\{ i \in [n] : |z_i| > F^{-1}\left(1 - \frac{m}{8n}\right) \right\} \right| \leq \frac{m}{4}, \tag{C.19}$$

with probability at least $1 - \exp(-\Omega(m))$. Let $S_2' := S_2 \cap T$ denote the corresponding set of data points that are preserved after corruption.

Next, let $q_i := x_i z_i$, for $1 \leq i \leq n$, and note that the q_i's are i.i.d. random variables with mean zero and covariance $\sigma^2 I$. Applying Theorem C.1.3 with $\epsilon_3 = \frac{m}{3n}$ on the set $S' := \{q_1, \ldots, q_n\}$, we see that, with probability except $O(\exp(-\Omega(m)))$, there exists a set $S_3 \subseteq S'$ such that (i) $|S_3| \geq (1 - \epsilon_3)n$, and (ii) S_3 is $(C_4 \epsilon_3, \delta_3)$-stable with respect to σ^2, where $C_4 = c_1 C_2 + 1$ and $\delta_3 = O\left(\sqrt{\frac{p \log p}{n}} + \sigma\sqrt{\frac{m}{n}}\right)$. Let $S_3' := \{(x_i, y_i) : x_i z_i \in S_3\} \cap T$ denote the corresponding set of (x, y) pairs that are also preserved after corruption.

Finally, we define the set

$$T_2 := T_1 \cap S_2' \cap S_3'.$$

Note that

$$|T_1 \setminus T_2| \leq (|S \setminus S_2| + |T \setminus S|) + (|S' \setminus S_3| + |T \setminus S|) \leq 2\epsilon n + \frac{m}{4} + \frac{m}{3} \leq m,$$

where we use the fact that $m \geq 6\epsilon n$ (since $C_1 > 3$). Thus, the vector $b^* \in \mathbb{R}^{n_1}$ is indeed m-sparse, and Lemma C.8.1 implies an error bound of order $\frac{G}{\sqrt{n_1}} + \frac{H}{n_1} = O\left(\frac{G}{\sqrt{n}} + \frac{H}{n}\right)$. It remains to control the parameters G and H.

Recall that with high probability, inequality (C.19) holds, in which case the nonzero

entries of w_i have magnitude at most $F^{-1}\left(1-\frac{m}{8n}\right)$. Thus, we have

$$\sup_{S':|S'|\le 2m}\sqrt{\sum_{i\in S'} w_i^2} \le \sqrt{2m}F^{-1}\left(1-\frac{m}{8n}\right) \lesssim \sqrt{m}\left(\frac{m}{n}\right)^{-1/k'},$$

where the second inequality follows from the $(k')^{\text{th}}$ moment condition on z_i. Thus, we may take $G = O\left(\sqrt{m}\left(\frac{m}{n}\right)^{-1/k'}\right)$.

Turning to H, note that with high probability, we have

$$\frac{|T_2|}{|S_3|} \ge \frac{|T_1|-m}{n} \ge 1 - c_1\epsilon_1 - \frac{m}{n} = 1-(c_1C_2+1)\frac{m}{n} = 1-C_4\epsilon_3.$$

Hence, the $(C_4\epsilon_3, \delta_3)$-stability of S_3 implies that

$$\left\|\sum_{i=1}^{n_1} x_i' w_i\right\|_2 = \left\|\sum_{(x,y)\in T_2} x_i z_i\right\|_2 \le |T_2|\sigma\delta_3 \le n\sigma\delta_3,$$

where we employ the notation used in the proof of Theorem 5.3.6 in the second expression. Therefore, $H \le n\sigma\delta_3$.

Altogether, we arrive at the error bound

$$\|\widehat{\beta}-\beta^*\|_2 \le \frac{G}{\sqrt{n}} + \frac{H}{n} \lesssim \sigma\left(\delta_3 + \sigma_{z,k'}\left(\frac{m}{n}\right)^{\frac{1}{2}-\frac{1}{k'}}\right) \lesssim \sigma\sigma_{z,k'}\left(\frac{p\log p}{n} + \epsilon + \frac{\log(1/\tau)}{n}\right)^{\frac{1}{2}-\frac{1}{k'}},$$

where we use the value of m and the fact that $\delta_3 \lesssim \sigma_{z,k'}\left(\frac{m}{n}\right)^{1/2-1/k'}$. Moreover, the probability of error is at most $O(\exp(-\Omega(m)))$. Lastly, we choose C_1 large enough so that the error probability is at most $O(\tau)$.

Finally, we bound the number of iterations of the alternating minimization algorithm required to guarantee the desired accuracy bound. In light of Remark C.8.2, it suffices to obtain a high-probability upper bound on $\|b^*\|_2$ that can be computed from the data. Recall the notation $S = (X, y)$ and $T = (X', y')$ for the i.i.d. and corrupted data sets, respectively, and recall that $T_1 \subseteq T$ denotes the filtered data set. Abusing notation slightly, we write the model (C.18) in matrix/vector form as $y'_{T_1} = X'_{T_1}\beta^* + w_{T_1} + b^*_{T_1}$. We claim that

$$\|b^*_{T_1}\|_2 = O\left(\|y'\|_2(1+\|X'\|_2)\right), \tag{C.20}$$

with probability at least $1 - O(\exp(-\Omega(n)))$.

Recall that by construction, either $b_i^* = 0$ or $w_i = 0$ for each i in the model (C.18).

Thus, by the triangle inequality, we have

$$\|b^*_{T_1}\|_2 \le \|y'_{T_1}\|_2 + \|X'_{T_1}\beta^*\|_2 \le \|y'\|_2 + \|X'\beta^*\|_2 \le \|y'\|_2 + \|X'\|_2\|\beta^*\|_2.$$

We now use concentration properties of the i.i.d. points in S to obtain a data-driven upper bound on $\|\beta^*\|_2$. Note that $\mathbb{E}(y_i^2) = \|\beta^*\|_2^2 + \sigma^2$. Furthermore, by Lemma C.4.6 and the convexity of the absolute value function, we have

$$\mathbb{E}\,|y_i| = \mathbb{E}\,|x_i^\top\beta^* + z_i| \ge \max\{\mathbb{E}\,|x_i^\top\beta^*|, \mathbb{E}\,|z_i|\}.$$

Furthermore, we can lower-bound both $\mathbb{E}\,|x_i^\top\beta^*|$ and $\mathbb{E}\,|z_i|$ using Proposition C.9.1 and Assumption 5.2.1:

$$\mathbb{E}\,|x_i^\top\beta^*| \ge \frac{\|\beta^*\|_2}{\sigma_{x,4}^2},$$
$$\mathbb{E}\,|z_i| \ge \frac{\sigma}{\sigma_{z,4}^2},$$

using the assumption that $(\mathbb{E}\,|z_i|^4)^{1/4} \le \sigma_{z,4}\sigma$ by $(4,2)$-hypercontractivity.

By the Paley-Zygmund inequality (e.g., see Exercise 2.4 of Boucheron et al. [BLM13]), we have

$$\begin{aligned}
\mathbb{P}\left(|y_i| \ge \frac{\mathbb{E}\,|y_i|}{2}\right) &\ge \frac{(\mathbb{E}\,|y_i|)^2}{4\,\mathbb{E}\,y_i^2} \\
&\ge \frac{\max\left\{\frac{\|\beta^*\|_2^2}{\sigma_{x,4}^4}, \frac{\sigma^2}{\sigma_{z,4}^4}\right\}}{4(\|\beta^*\|_2^2 + \sigma^2)} \\
&\ge \frac{1}{\max\{\sigma_{x,4}^4, \sigma_{z,4}^4\}} \cdot \frac{\frac{1}{2}\left(\|\beta^*\|_2^2 + \sigma^2\right)}{4(\|\beta^*\|_2^2 + \sigma^2)} \\
&= \frac{1}{8\max(\sigma_{x,4}^4, \sigma_{z,4}^4)}.
\end{aligned}$$

Thus,

$$\mathbb{P}\left(|y_i| \ge \frac{\|\beta^*\|_2}{2\sigma_{x,4}^2}\right) \ge \frac{1}{8\max\{\sigma_{x,4}^4, \sigma_{z,4}^4\}}.$$

Let $\gamma = 16\max\{\sigma_{x,4}^4, \sigma_{z,4}^4\}$, which is assumed to be $O(1)$. Let W be the $\lceil(1 - 1/\gamma)\,n\rceil^{\text{th}}$ largest $|y_i|$. Then by a Chernoff bound, we have

$$\mathbb{P}\left(W < \frac{\|\beta^*\|_2}{2\sigma_{x,4}^2}\right) \le \exp\left(-\Omega\left(\frac{n}{\alpha}\right)\right).$$

Finally, for $\epsilon < \frac{1}{2\gamma}$, we have $\max_i |y'_i| \geq W$. Therefore, with high probability,

$$\|\beta^*\|_2 \leq 2\sigma_{x,4}^2 \max_i |y'_i| = O(\|y'\|_2).$$

This completes the proof.

C.8.2 Proof of Lemma C.8.1

In this appendix, we reproduce the proof of the convergence guarantee for alternating minimization from Bhatia et al. [BJKK17].

We begin by introducing some additional notation: For a vector $a \in \mathbb{R}^n$ and a set $S \subseteq [n]$, we will use a_S to denote the vector $q \in \mathbb{R}^n$ such that (i) for $i \in S$, $q_i = v_i$; and (ii) for $i \notin S$, $q_i = 0$. Similarly, for a matrix $A \in \mathbb{R}^{n \times p}$ and a set $S \subseteq [n]$, we will use A_S to denote the matrix $Q \in \mathbb{R}^{n \times p}$ such that (i) for $i \in S$, the i^{th} row of Q is the same as the i^{th} row of A; and (ii) for $i \notin S$, all entries in the i^{th} row of Q are 0.

Lemma C.8.3. *Suppose $a \in \mathbb{R}^n$. Let $b = \mathrm{HT}_r(a)$, let $S_1 = supp(b)$, and let $S \subseteq [n]$ be such that $S_1 \subseteq S$. Then for any r-sparse vector c, we have $\|b - a_S\|_2 \leq \|c - a_S\|_2$.*

Proof. Without loss of generality, let a be such that $|a_1| \geq |a_2| \geq \cdots \geq |a_n|$. Then $S_1 = [r]$. Note that for any vector c, we have

$$\|c - a_S\|_2^2 \geq \|c_S - a_S\|_2^2 = \sum_{i \in S} (c_i - a_i)^2.$$

It is not hard to see that the right-hand expression is minimized over r-sparse vectors when $c_i = a_i$ for $i \in S_1$ and $c_i = 0$ for $i \in S \setminus S_1$. This yields the expression $\|b - a_S\|_2^2$, completing the proof. □

Using the notation from Bhatia et al. [BJKK17], let $X \in \mathbb{R}^{d \times n}$ denote the matrix of covariates, let $Y \in \mathbb{R}^n$ denote the vector of responses, and let $Z := Y - X^\top \beta^*$. (Note that the matrix X is now defined to be the transpose of the design matrix that we denote by X elsewhere in the paper.) Recall that the model is $Y = X^\top \beta^* + w + b^*$, where the idea is that w has small entries and is nearly orthogonal to X, whereas b^* is m-sparse.

Recall that b^j was defined iteratively in the algorithm, and further define

$$\lambda^j := (XX^\top)^{-1} X(b^j - b^*),$$
$$g := (I - P_X)w.$$

Note that the update step can be written as follows:

$$b^{j+1} = \mathrm{HT}_m\left(P_X b^j + (I - P_X)(X^\top \beta^* + w + b^*)\right) = \mathrm{HT}_m(b^* + X^\top \lambda^j + g),$$

using the fact that $X^\top = P_X X^\top$. Denote $I_j := \mathrm{supp}(b^j) \cup \mathrm{supp}(b^*)$. Applying Lemma C.8.3 with $a = b^* + X^\top \lambda^j + g$ and $S = I_{j+1}$, we have

$$\begin{aligned}\|b^{j+1} - (b^* + X^\top \lambda^j + g)_{I_{j+1}}\|_2 &\leq \|b^* - (b^* + X^\top \lambda^j + g)_{I_{j+1}}\|_2 \\ &= \|b^* - b^* - X^\top_{I_{j+1}} \lambda^j - g_{I_{j+1}}\|_2 = \|X^\top_{I_{j+1}} \lambda^j + g_{I_{j+1}}\|_2,\end{aligned}$$

where we use the fact that $\mathrm{supp}(b^*) \subseteq I_{j+1}$. By the triangle inequality, we then have

$$\begin{aligned}\|b^{j+1} - b^*\|_2 &\leq \|b^{j+1} - b^* - X^\top_{I_{j+1}} \lambda^j - g_{I_{j+1}}\|_2 + \|X^\top_{I_{j+1}} \lambda^j + g_{I_{j+1}}\|_2 \\ &\leq 2\|X^\top_{I_{j+1}} \lambda^j + g_{I_{j+1}}\|_2 \leq 2\|X^\top_{I_{j+1}} \lambda^j\|_2 + 2\|g_{I_{j+1}}\|_2.\end{aligned}$$

We bound each of the latter two terms separately. For the first term, we use the definition of λ^j and the eigenvalue bounds on the covariates to write the following:

$$\|X^\top_{I_{j+1}} \lambda^j\|_2 = \|X^\top_{I_{j+1}} (XX^\top)^{-1} X_{I_{j+1}} (b^j - b^*)\|_2 \leq \frac{\Lambda_{2m}}{\lambda_n} \|b^j - b^*\|_2.$$

We now focus on the second term. By the triangle inequality, we have

$$\begin{aligned}\|g_{I_{j+1}}\|_2 &= \|W_{I_{j+1}} - X^\top_{I_{j+1}} (XX^\top)^{-1} XW\|_2 \\ &\leq \|W_{I_{j+1}}\|_2 + \|X^\top_{I_{j+1}} (XX^\top)^{-1} XW\|_2 \\ &\leq G + \frac{H}{\sqrt{\lambda_n}},\end{aligned}$$

using the fact that $W_{I_{j+1}}$ is at most $2m$-sparse and the bound

$$\|X^\top_{I_{j+1}} (XX^\top)^{-1} XW\|_2 \leq \frac{\sqrt{\Lambda_{2m}} H}{\lambda_n} \leq \frac{H}{\sqrt{\lambda_n}}.$$

Combining the inequalities yields the bound

$$\|b^{j+1} - b^*\|_2 \leq \frac{2\Lambda_{2m}}{\lambda_n} \|b^j - b^*\|_2 + e_0 \leq \frac{1}{2} \|b^j - b^*\|_2 + e_0, \tag{C.21}$$

where $e_0 := 2G + 2\frac{H}{\sqrt{\lambda_n}}$ and we have used the assumption that $\frac{2\Lambda_{2m}}{\lambda_n} \leq \frac{1}{2}$. Iterating the

bound, we see that $\|b^j - b^*\| \le 3e_0$ whenever $j \ge \log_2\left(\frac{\|b^0 - b^*\|_2}{e_0}\right)$.

To bound the final error between β^j and β^*, we note that $\beta^j - \beta^* = (XX^\top)^{-1}X(W + b^* - b^j)$. Using the definitions of G and H, we have

$$\begin{aligned}\|\beta^j - \beta^*\|_2 = \|(XX^\top)^{-1}X(W + b^* - b^j)\|_2 &\le \frac{\|X(W + (b^* - b^j))\|_2}{\lambda_n} \\ &\le \frac{\|XW\|_2 + \|X(b^* - b^j)\|_2}{\lambda_n} \lesssim \left(\frac{H + \sqrt{\Lambda_n}\left(G + \frac{H}{\sqrt{\lambda_n}}\right)}{\lambda_n}\right) \\ &\lesssim \frac{H + G\sqrt{\Lambda_n}}{\lambda_n},\end{aligned}$$

completing the proof.

C.9 Least Absolute Deviation

In this appendix, we provide additional proof details for the results in Section 5.5.

C.9.1 Auxiliary Results

Proposition C.9.1. *Suppose Z satisfies $\mathbb{E}\, Z^2 = 1$ and $\mathbb{E}\, Z^4 < \infty$. Then $\mathbb{E}\,|Z| > 1/\sqrt{\mathbb{E}\,|Z|^4}$.*

Proof. We apply Hölder's inequality, which states that

$$\mathbb{E}\,|XY| \le (\mathbb{E}\,|X|^p)^{1/p}(\mathbb{E}\,|Y|^q)^{1/q},$$

for $p \in (1, \infty)$ and $q = \frac{p}{p-1}$. Taking $X = Z^{4/3}$, $Y = Z^{2/3}$, and $p = 3$, we have

$$1 = \mathbb{E}\, Z^2 \le (\mathbb{E}(|Z|^{4/3})^3)^{1/3}(\mathbb{E}(|Z|^{2/3})^{3/2})^{2/3} = (\mathbb{E}\,|Z|^4)^{1/3}(\mathbb{E}\,|Z|)^{2/3}.$$

□

Lemma C.9.2. *Let $X_1, \ldots, X_n$ be i.i.d. nonnegative random variables and let $\epsilon \in (0, 1)$. Then with probability $1 - 2\exp(-cn\epsilon)$, the trimmed sum satisfies*

$$\sum_{i=1}^{(1-\epsilon)n} X_{(i)} = O\left(\frac{n\,\mathbb{E}\, X_i}{\epsilon}\right),$$

where $\{X_{(i)}\}_{i=1}^n$ are order statistics.

Proof. Let F be the cdf of the X_i's, and let F^{-1} be its inverse, so $F^{-1}(1-\epsilon) = \inf\{t : \mathbb{P}(X_i > t) \le \epsilon\}$ for $\epsilon \in [0,1]$. Let $a := F^{-1}\left(1 - \frac{\epsilon}{3}\right)$ and define $Z_i = \min(X_i, a)$. Note that $\sum_{i=1}^n Z_i \le an$.

Now let $Y_i = \mathbb{I}\{X_i > a\}$ and define the event

$$\mathcal{E} := \left\{\sum_{i=1}^n Y_i < \epsilon n\right\}.$$

We have

$$\mathbb{E}\, Y_i = \mathbb{P}(X_i > a) = \mathbb{P}\left(X_i > F^{-1}\left(1 - \frac{\epsilon}{3}\right)\right) \le \frac{\epsilon}{3}.$$

Applying a Chernoff bound, we therefore have

$$\sum_{i=1}^n Y_i \le \frac{2\epsilon n}{3},$$

with probability at least $1 - \exp(-cn\epsilon)$, implying that $\mathbb{P}(\mathcal{E}) \ge 1 - \exp(-c_2 n\epsilon)$.

Finally, note that on the event $\mathcal{E}$, we have

$$\sum_{i=1}^{(1-\epsilon)n} X_{(i)} \le \sum_{i=1}^n Z_i \le an.$$

Applying Markov's inequality, we have $\mathbb{P}\left(X_i \ge \frac{4\,\mathbb{E}\, X_i}{\epsilon}\right) \le \frac{\epsilon}{4} < \frac{\epsilon}{3}$. Therefore, $a \le \frac{4\,\mathbb{E}\, X_i}{\epsilon}$, completing the proof. □

Lemma C.9.3. *Suppose the covariates $x_1, \dots, x_n$ are sampled i.i.d. from a distribution satisfying Assumption 5.2.1. With probability $1 - 2\exp(-cn\epsilon)$, we have that for any unit vector v and any $S \subseteq [n]$ with $|S| \ge (1-\epsilon)n$, the following holds:*

$$\frac{1}{n}\sum_{i \in S} |x_i^\top v| \ge \frac{1}{\sigma_{x,4}^2} - O\left(\sqrt{\epsilon} + \sqrt{\frac{p}{n}}\right).$$

Proof. Let Q be the threshold $C\left(\sqrt{\frac{1}{\epsilon}} + \frac{1}{\epsilon}\sqrt{\frac{p}{n}}\right)$ from Lemma C.4.3. Let $\mathcal{E}$ denote the event from Lemma C.4.3, stating that for any unit vector v, we have $\left|\{i : |x_i^\top v| \ge Q\}\right| \le \epsilon n$. By the lemma, we know that $\mathbb{P}(\mathcal{E}) \ge 1 - \exp(-cn\epsilon)$.

We will now assume that the event $\mathcal{E}$ holds and incur an additional failure probability

of $\exp(-cn\epsilon)$ by a union bound. Define the function $f : \mathbb{R}_+ \to \mathbb{R}_+$, as follows:

$$f(x) = \begin{cases} x, & \text{if } x \in [0, Q], \\ Q, & \text{otherwise,} \end{cases},$$

and let $g(x) = -f(x)$. For any $v \in \mathcal{S}^{p-1}$, on the event $\mathcal{E}$, we have the following bound:

$$\begin{aligned} \min_{S:|S|\geq(1-\epsilon)n} \sum_{i\in S} |x_i^\top v| &\geq \sum_{i=1}^n f(|x_i^\top v|) - \epsilon Q n \\ &= -\left(\sum_{i=1}^n g(|x_i^\top v|) - \mathbb{E}\, g(|x_i^\top v|)\right) + n\,\mathbb{E}\, f(|x_i^\top v|) - \epsilon Q n. \end{aligned}$$

Taking an infimum over v, we then have

$$\begin{aligned} &\inf_{v\in\mathcal{S}^{p-1}} \min_{S:|S|\geq(1-\epsilon)n} \sum_{i\in S} |x_i^\top v| \\ &\quad\geq -\epsilon Q n - \sup_{v\in\mathcal{S}^{p-1}} \left(\sum_{i=1}^n g(|x_i^\top v|) - \mathbb{E}\, g(|x_i^\top v|)\right) + n\left(\inf_{v\in\mathcal{S}^{p-1}} \mathbb{E}\, f(|x_i^\top v|)\right). \end{aligned} \tag{C.22}$$

Now define the random variable

$$N := \sup_{v\in\mathcal{S}^{p-1}} \sum_{i=1}^n g(|x_i^\top v|) - \mathbb{E}\, g(|x_i^\top v|).$$

We first bound the expectation of N using symmetrization and contraction of Rademacher averages [LT91; BLM13]:

$$\begin{aligned} \mathbb{E}\, N &\leq 2\,\mathbb{E} \sup_{v\in\mathcal{S}^{p-1}} \left|\sum_{i=1}^n \xi_i g(|x_i^\top v|)\right| \leq 4\,\mathbb{E} \sup_{v\in\mathcal{S}^{p-1}} \left|\sum_{i=1}^n \xi_i x_i^\top v\right| \\ &\leq 4\,\mathbb{E}\left(\left\|\sum_{i=1}^n \xi_i x_i\right\|_2 \sup_{v\in\mathcal{S}^{p-1}} \|v\|_2\right) \leq 4\sqrt{\mathbb{E}\left(\left\|\sum_{i=1}^n \xi_i x_i\right\|_2^2\right)} \\ &= 4\sqrt{\mathbb{E}\left(\sum_{i=1}^n x_i^\top x_i\right)} = 4\sqrt{\sum_{i=1}^n \mathbb{E}\left(\operatorname{tr}(x_i^\top x_i)\right)} = 4\sqrt{\sum_{i=1}^n \mathbb{E}\left(\operatorname{tr}(x_i x_i^\top)\right)} \\ &= 4\sqrt{\sum_{i=1}^n \operatorname{tr}\left(\mathbb{E}\left(x_i x_i^\top\right)\right)} \\ &= 4\sqrt{pn}, \end{aligned}$$

where the ξ_i's are i.i.d. Rademacher random variables. We now bound the following

term (which is usually called the *wimpy variance* [BLM13]):

$$\sigma^2 := \sup_v n \, \mathbf{Var}(g(|x_i^\top v|)) \le \sup_v n \, \mathbb{E} \, |x_i^\top v|^2 = n \, .$$

Using Talagrand's inequality for bounded empirical processes (cf. Lemma C.4.2), we therefore have

$$N = O(\sqrt{pn} + \sqrt{n}\sqrt{n\epsilon} + Qn\epsilon) = O(\sqrt{pn} + n\sqrt{\epsilon} + n\sqrt{\epsilon} + \sqrt{pn}) = O(\sqrt{pn} + n\sqrt{\epsilon}), \quad \text{(C.23)}$$

with probability at least $1 - \exp(-c'n\epsilon)$.

Finally, note that for any $v \in \mathcal{S}^{p-1}$, the Cauchy-Schwarz inequality gives

$$\begin{aligned}
\mathbb{E} \left| f(|x_i^\top v|) - |x_i^\top v| \right| &\le \mathbb{E} \left(|x_i^\top v| \, \mathbb{I}\{|x_i^\top v| > Q\} \right) \\
&\le \sqrt{\mathbb{E}(x_i^\top v)^2} \sqrt{\mathbb{P}(|x_i^\top v| \ge Q)} \le \sqrt{\frac{\mathbb{E}(x_i^\top v)^2}{Q^2}} \\
&= O(\sqrt{\epsilon}),
\end{aligned}$$

where the last two steps use Markov's inequality and the fact that $Q = \Omega(1/\sqrt{\epsilon})$. Thus,

$$\mathbb{E} \, f(|x_i^\top v|) \ge \mathbb{E} \, |x_i^\top v| - O(\sqrt{\epsilon}) \ge \frac{1}{\sigma_{x,4}^2} - O(\sqrt{\epsilon}), \quad \text{(C.24)}$$

where the second inequality follows from Proposition C.9.1.

Combining inequalities (C.22), (C.23), and (C.24), we obtain the bound

$$\begin{aligned}
\frac{1}{n} \inf_{v \in \mathcal{S}^{p-1}} \min_{S : |S| \ge (1-\epsilon)n} \sum_{i \in S} |x_i^\top v| &\ge \inf_v \mathbb{E} \, f(|x_i^\top v|) - \epsilon Q - \frac{N}{n} \\
&\ge \frac{1}{\sigma_{x,4}^2} - O\left(\sqrt{\frac{p}{n}} + \sqrt{\epsilon} \right).
\end{aligned}$$

This completes the proof. □

C.9.2 Proof of Theorem 5.5.1

Our main result relies on the following lemma from Karmalkar and Price [KP19], who showed that if the covariates satisfy (ϵ, m, M, ℓ_1)-stability, then the LAD estimator is robust to corruption in responses. We provide a proof for completeness:

Lemma C.9.4. (*Karmalkar and Price* [KP19]) *Suppose the covariates satisfy* (m, M, ϵ, ℓ_1)*-stability such that* $M > m$*. Then* $\|\widehat{\beta}_{LAD} - \beta^*\|_2 = O\left(\frac{\sum_{i=1}^{(1-\epsilon)n} |z|_{(i)}}{n(M-m)}\right)$.

Proof. We denote $\widehat{\beta} = \widehat{\beta}_{\text{LAD}}$ for brevity. Let S be the set of $(1-\epsilon)n$ indices with the smallest magnitudes of additive errors. We have the following:

$$\begin{aligned}
0 &\geq \sum_{i\in S} |y_i - x_i^\top \widehat{\beta}| - \sum_{i\in S} |y_i - x_i^\top \beta^*| + \sum_{i\in S^c} |y_i - x_i^\top \widehat{\beta}| - \sum_{i\in S^c} |y_i - x_i^\top \beta^*| \\
&\geq \sum_{i\in S} |x_i^\top(\widehat{\beta} - \beta^*)| - 2\sum_{i\in S} |y_i - x_i^\top \beta^*| - \sum_{i\in S^c} |x_i^\top(\widehat{\beta} - \beta^*)|, \\
&\geq nM\|\widehat{\beta} - \beta^*\|_2 - 2\sum_{i\in S} |z_i| - nm\|\widehat{\beta} - \beta^*\|_2,
\end{aligned}$$

where the first inequality follows by the optimality of $\widehat{\beta}$, the second inequality uses the triangle inequality, and the third inequality uses the property of (ϵ, m, M, ℓ_1)-stability. Rearranging the inequality and using the fact that $\sum_{i\in S} |z_i| \leq \sum_{i=1}^{(1-\epsilon)n} |z|_{(i)}$, we obtain the desired result. □

The following lemma shows that the filtered covariates satisfy (m, M, ϵ, ℓ_1)-stability:

Lemma C.9.5. *Let* S *be the data set described in Theorem 5.5.1. For an* $\epsilon_1 < c_*$*, let* T *be an* ϵ_1*-corrupted version of set* S*. Let* T_1 *be the output of the filter algorithm on input* T *and* ϵ'*, where* $\epsilon' = \Theta(1)$*. Then with probability at least* $1 - O(\exp(-\Omega(n)))$*, the set* T_1 *satisfies* $(\epsilon_2, m, M, \ell_1)$*-stability with* $\epsilon_2 = \Theta(1)$*,* $m = \Theta(1)$*,* $M = \Theta(1)$*, and* $M \geq 2m$*, and these parameters do not depend on* ϵ_1*. Moreover,* $|T_1| \geq \frac{n}{2}$.

Proof. We provide a sketch of the proof here; more details may be found in Appendix C.9.3. We show that the lower bound (on M) in Definition 5.2.8 is satisfied due to the small-ball property [Men15], and that the filtering algorithm removes the "outliers" in the data set, leading to the upper bound (on m). The proof of the lower bound is given in Lemma C.9.3, which follows similar calculations from previous work [KM15; DKP20]. These arguments show that if $n = \Omega(p \log p)$, the ℓ_1-stability lower bound holds with $M \geq \frac{1}{2\sigma_4^2}$. For the upper bound, we use the fact that the filtered set T_1 is (ϵ, δ)-stable. Then Proposition C.6.5 implies that for $T' \subseteq T_1$ with $|T'| \leq \epsilon|T_1|$, and any unit vector v, we have $\frac{1}{|T_1|}\sum_{i\in T'} |x_i^\top v| \leq 2\delta$, so the stability upper bound holds with $m \leq 2\delta$. We choose the parameter values such that $M \geq \frac{1}{2\sigma_4^2} \geq 4\delta \geq 2m = \Omega(1)$. □

Lemma C.9.5 states that, with probability at least $1 - O(\exp(-\Omega(n))$, the set T_1 obtained by running the filtering algorithm on T satisfies $(\epsilon_2, m, M, \ell_1)$-stability, where

$2m \le M = \Theta(m)$ and $\epsilon_2 = \Theta(1)$. We assume that ϵ is small enough such that $\epsilon_2 > 4\epsilon$. Applying Lemma C.9.4, we claim that the ℓ_2-estimation error is bounded by a constant times $\sum_{i=1}^{n-\epsilon_2 n_1} |y' - X'\beta^*|_{(i)}$, where we denote the corrupted data set by $T = \{(x_i', y_i')\}_{i=1}^n$ and $n_1 = |T_1| = (1-\epsilon')n$. Indeed, the bound in Lemma C.9.4 involves a sum of the $(1-\epsilon_2)n_1$ smallest residuals in the filtered data set. Each of these terms appears in the set of residuals $\{|y_i' - x_i'^\top \beta^*|\}_{i=1}^n$ for T, so the aforementioned sum is certainly upper-bounded by the sum of all but the $\epsilon_2 n_1$ largest residuals for T. Furthermore, we have

$$\sum_{i=1}^{n-\epsilon_2 n_1} |y' - X'\beta^*|_{(i)} \le \sum_{i=1}^{n-\epsilon_2 n/2} |y' - X'\beta^*|_{(i)} \le \sum_{i=1}^{n-\epsilon_2 n/2+\epsilon n} |y - X\beta^*|_{(i)} \le \sum_{i=1}^{n-\epsilon_2 n/4} |y - X\beta^*|_{(i)},$$

where the first inequality uses the fact that $n_1 \ge \frac{n}{2}$, the second inequality uses the fact that T differs from S in at most ϵn points, and the last inequality uses the fact that $\epsilon \le \frac{\epsilon_2}{4}$. Applying Lemma C.9.2, we see that the final quantity is at most $O\left(\frac{n\kappa}{\epsilon_2}\right)$, with probability at least $1 - O(\exp(-\Omega(n\epsilon_2)))$. Since $\epsilon_2 = \Omega(1)$, this completes the proof.

C.9.3 Proof of Lemma C.9.5

We follow the proof strategy from Koltchinskii and Mendelson [KM15] and Diakonikolas et al. [DKP20].

Let $T_1 = \{(x_i', y_i')\}_{i=1}^{n_1}$ be the output of the filter algorithm with inputs T and ϵ', where $\epsilon_1 < \epsilon'$. By Proposition C.6.3, with probability at least $1 - 2\exp(-n\epsilon')$, the set T_1 is (ϵ_2, δ_2)-stable, where $\epsilon_2 = \Theta(\epsilon')$ and $\delta_2 = O\left(\sqrt{\frac{p \log p}{n}} + \sqrt{\epsilon'}\right)$, and T_1 has cardinality $n_1 \ge (1 - c_1\epsilon')n$. Furthermore, we choose ϵ_1 and ϵ' sufficiently small to guarantee that $n_1 \ge \frac{n}{2}$. Therefore, for any $T' \subseteq T_1$ such that $|T'| \le \epsilon_2 |T_1|$, Proposition C.6.5 states that for all unit vectors v,

$$\frac{1}{n_1} \sum_{x_i' \in T'} |v^\top x_i'| \le 2\delta_2. \tag{C.25}$$

Let $T_2 \subseteq T_1$ be a set such that $|T_2| \ge (1-\epsilon_2)|T_1|$. Since $|T_1| = n_1 \ge \frac{n}{2}$, we have

$$\begin{aligned} |T_2 \cap S| &= |S| - |S \setminus T| - |T \setminus T_1| - |T_1 \setminus T_2| \ge n - \epsilon_1 n - c_1\epsilon' n - \epsilon_2 n_1 \ge (1 - \epsilon_1 - c_1\epsilon' - \epsilon_2)n \\ &\ge (1 - c_3\epsilon')n, \end{aligned} \tag{C.26}$$

where c_3 is a constant, using the facts that $\epsilon_1 < \epsilon'$ and $\epsilon_2 = \Theta(\epsilon')$.

Now suppose $n = \Omega(p\sigma_{x,4}^4)$ and $\epsilon_0 = O\left(\frac{1}{\sigma_{x,4}^4}\right)$. By Lemma C.9.3, we know that, with

probability at least $1 - \exp(-\Omega(n\epsilon_0))$, we have

$$\frac{1}{n}\sum_{i\in S'} |x_i^\top v| \geq \frac{1}{2\sigma_{x,4}^2}, \tag{C.27}$$

for any $S' \subseteq [n]$ such that $|S'| \geq (1-\epsilon_0)n$ and any $v \in \mathcal{S}^{p-1}$. Hence, if $c_3\epsilon' \leq \epsilon_0$, inequalities (C.26) and (C.27) together imply that

$$\frac{1}{|T_1|}\sum_{x_i'\in T_2} |v^\top x_i'| \geq \frac{1}{n}\sum_{x_i\in T_2\cap S} |v^\top x_i| \geq \frac{1}{2\sigma_{x,4}^2}. \tag{C.28}$$

From inequalities (C.25) and (C.28), we conclude that T_1 satisfies $\left(\epsilon_2, m = 2\delta_2, M = \frac{1}{2\sigma_{x,4}^2}, \ell_1\right)$-stability with the desired probability. Note that if we choose $n = \Omega(p\log p)$ large enough and ϵ' to be a sufficiently small constant, we can guarantee that $c_2\epsilon_2 \leq \epsilon_0$ and δ_2 is sufficiently small, so $2m \leq M$.

C.10 Postprocessing

In this appendix, we provide additional proof details for the results in Section 5.6. We will use the following result from Diakonikolas et al. [DKP20], which gives a result corresponding to Theorem C.1.3 when the distribution only has a finite variance:

Theorem C.10.1. (*Diakonikolas et al.* [*DKP20*]) *Let S be a set of n i.i.d. points from a distribution in $\mathbb{R}^p$ with mean μ and covariance $\Sigma \preceq \sigma^2 I$ for some $\sigma \geq 0$. Let ϵ and τ be such that $\epsilon' = C\left(\epsilon + \frac{\log(1/\tau)}{n}\right) = O(1)$, for a large enough constant C. Then with probability at least $1-\tau$, there exists a subset $S' \subseteq S$ such that $|S'| \geq (1-\epsilon')|S|$ and S' is $(C_1\epsilon', \delta)$-stable with respect to μ and σ^2, where $C_1 > 2$ is any large constant and $\delta = O\left(\sqrt{\frac{p\log p}{n}} + \sqrt{\epsilon} + \sqrt{\frac{\log(1/\tau)}{n}}\right)$, with prefactor depending on C_1.*

C.10.1 Proof of Theorem 5.6.1

We will use the following result from Diakonikolas et al. [DKP20], which shows that applying iterative filtering to $\{z_1, \ldots, z_k\}$ returns a sub-Gaussian estimate of the mean of the original sample:

Theorem C.10.2. (*Diakonikolas et al.* [*DKP20*]) *Let S be a set of n i.i.d. samples from a distribution with mean μ and covariance Σ. Let T be an ϵ-corrupted version of S. For a probability τ, let $\epsilon' = \Theta\left(\epsilon + \frac{\log(1/\tau)}{n}\right)$, where ϵ' is less than a small constant. Let $k = \lceil \epsilon' n \rceil$. Let*

$T_k := \{z_1, \ldots, z_k\}$ be the set obtained by median-of-means preprocessing on the set T. Then running the filtering algorithm in Theorem C.1.1 with inputs T_k and $\epsilon' = \Theta(1)$ returns a set T' such that, with probability at least $1 - \exp(-\Omega(k))$,

$$\|\widehat{\mu}_{T'} - \mu\|_2 = O\left(\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \sqrt{\|\Sigma\|_2 \epsilon} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\tau)}{n}}\right),$$

where $\widehat{\mu}_{T'}$ is the empirical mean of the set T'.

Turning to the proof of Theorem 5.6.1, we will condition on the value of the initial estimator $\widehat{\beta}_1$. Let $S_1 := \left\{\widehat{\beta}_1 + (y_i - x_i^\top \widehat{\beta}_1)x_i : (x_i, y_i) \in S\right\}$. Since $\widehat{\beta}_1$ is independent of S by assumption, the set S_1 consists of i.i.d. samples when we condition on $\widehat{\beta}_1$. It is easy to see that T_1 is an ϵ-corrupted version of S_1 and $\mathbb{E}\left[\widehat{\beta}_1 + (y_i - x_i^\top \widehat{\beta}_1)x_i\right] = \beta^*$. Thus, the desired result follows from Theorem C.10.2 if we can show that the set S_1 satisfies the stated conditions. For simplicity, set

$$w_i := \widehat{\beta}_1 + (y_i - x_i^\top \widehat{\beta}_1)x_i = \widehat{\beta}_1 + x_i^\top x_i(\beta^* - \widehat{\beta}_1) + x_i z_i.$$

We will work conditionally on $\widehat{\beta}_1$ in the remainder of the proof. Since $\widehat{\beta}_1$ is independent of S, the w_i's are then conditionally i.i.d. Set $\Delta := \widehat{\beta}_1 - \beta^*$, so $\|\Delta\|_2 \le \sigma$ by assumption, and observe that $w_i - \beta^* = \Delta - x_i^\top x_i \Delta + x_i z_i$. Therefore, for any unit vector v, we have

$$\begin{aligned}
v^\top \Sigma_{w_i} v = \mathbb{E}(v^\top (w_i - \beta^*))^2 &= \mathbb{E}(v^\top \Delta - (v^\top x_i)(\Delta^\top x_i) + v^\top x_i z_i)^2 \\
&\lesssim (v^\top \Delta)^2 + \mathbb{E}\left((v^\top x_i)^2 (\Delta^\top x_i)^2\right) + \mathbb{E}\left((v^\top x_i)^2 z_i^2\right) \\
&\lesssim \|\Delta\|_2^2 + \sqrt{\mathbb{E}(v^\top x_i)^4}\sqrt{\mathbb{E}(\Delta^\top x_i)^4} + \sigma^2 \\
&\lesssim \|\Delta\|_2^2 + \sigma_{x,4}^4 \|\Delta\|_2^2 + \sigma^2 \\
&\lesssim \sigma^2. \qquad \text{(C.29)}
\end{aligned}$$

Therefore, $\mathrm{tr}(\Sigma_w) \lesssim \sigma^2 p$ and $\|\Sigma_w\|_2 \lesssim \sigma^2$. This completes the proof. (Observe that if $\|\widehat{\beta}_1 - \beta^*\|_2$ were much larger than σ, this argument yields an error bound which depends on $\sqrt{\sigma^2 + \|\widehat{\beta} - \beta^*\|_2^2}$.)

C.10.2 Proof of Theorem 5.6.2

Our approach differs from the proof of Theorem 5.6.1 in that the vectors in the set

$$S_1 = \left\{\widehat{\beta}_1 + (y_i - x_i^\top \widehat{\beta}_1)x_i : (x_i, y_i) \in S\right\}$$

may no longer be i.i.d. when we condition on the initial estimator $\widehat{\beta}_1$. Thus, we cannot directly apply Theorem C.10.2 to obtain an error bound. On the other hand, recall from Remark C.1.2 that if we can show the existence of a sufficiently large stable subset of the set S_1, Theorem C.1.1 implies a corresponding error bound.

For any fixed $v \in \mathbb{R}^p$, define the random variables

$$W_i^v := v + (y_i - x_i^\top v)x_i = v + x_i x_i^\top(\beta^* - v) + z_i x_i, \qquad \forall 1 \le i \le n,$$

and define the multiset $S_v := \{W_1^v, \ldots, W_n^v\}$. Note that each set S_v consists of n i.i.d. data points, so that stability properties can be obtained easily; the additional challenge is that we need to show the existence of a stable subset for all $v \in \mathbb{R}^p$ simultaneously, so that we can apply the result when $v = \widehat{\beta}_1$. To this end, we will use a covering argument. Let $r = \Theta(\sigma)$ be such that $\|\widehat{\beta}_1 - \beta^*\|_2 \le r$, and define the set $\mathcal{T} := \{v : \|\beta^* - v\|_2 \le r\}$. We now define $\mathcal{C}_\eta \subseteq \mathcal{T}$ to be an η-cover of $\mathcal{T}$, i.e., for every $v \in \mathcal{T}$, there exists $v' \in \mathcal{C}_\eta$ such that $\|v - v'\|_2 \le \eta$. Note that for $\eta \le r$, we can choose $\mathcal{C}_\eta$ such that $\log(|\mathcal{C}_\eta|) \le p \log\left(\frac{3r}{\eta}\right)$ (cf. Corollary 4.2.13 of Vershynin [Ver18]).

For any $v \in \mathcal{C}_\eta$, we have $\mathbb{E}\, W_i^v = \beta^*$. Let $\Delta = v - \beta^*$. As in inequality (C.29) in the proof of Theorem 5.6.1, we can argue that $\|\mathrm{Cov}(W_v)\|_2 \le C_0\sigma^2$. Applying Theorem C.10.1 with parameters $\tau' = \tau\exp(-C_1 p\log(pn))$ and $\epsilon' = \Theta\left(\epsilon + \frac{\log(1/\tau')}{n}\right) = \Theta\left(\epsilon + \frac{\log(1/\tau)}{n} + \frac{p\log(pn)}{n}\right)$, for a large constant $C_1 > 0$ to be defined later, we see that with probability at least $1 - \tau'$, there exists a set $S_v' \subseteq S_v$ such that $|S_v'| \ge (1-\epsilon')n$ and S_v' is $(C\epsilon', \delta)$-stable with respect to β^* and $\sigma_*^2 := C_0\sigma^2$, where $\delta := \Theta\left(\sqrt{\frac{p\log(pn)}{n}} + \sqrt{\epsilon} + \sqrt{\frac{\log(1/\tau)}{n}}\right)$.

Suppose a stable set exists for every element of C_η (we will bound the error probability later). Now consider an arbitrary $v' \in \mathbb{R}^p$, and let $v \in C_\eta$ be such that $\|v' - v\|_2 \le \eta$. We know that there exists a set $S_v' \subseteq S_v$ which is $(C\epsilon', \delta)$-stable with respect to β^* and σ_*^2; we will show how to obtain a stable set $S_{v'}' \subseteq S_{v'}$ using S_v'. Note that S_v' corresponds to a set of indices which we define as $T_v \subseteq [n]$, so $S_v' = \{W_i^v\}_{i \in T_v}$.

Define the set

$$S_2 := \left\{(x_i, y_i) : \|x_i\|_2 \le \sqrt{\frac{p}{\epsilon'}} \text{ and } |y_i - x_i^\top \beta^*| \le \frac{\sigma}{\sqrt{\epsilon'}}\right\}.$$

By a Chernoff bound, we can argue that with probability at least $1-\exp(-cn\epsilon') = 1-O(\tau)$, we have $|S_2| \ge (1-4\epsilon')n$. Indeed, define the indicator variables $E_i = 1\{(x_i, y_i) \in S_2\}$. Then

$$\begin{aligned}\mathbb{E}(E_i) &= \mathbb{P}\left(\|x_i\|_2 \le \sqrt{\frac{p}{\epsilon'}} \text{ and } |z_i| \le \frac{\sigma}{\sqrt{\epsilon'}}\right) \ge 1 - \mathbb{P}\left(\|x_i\|_2^2 \ge \frac{p}{\epsilon'}\right) - \mathbb{P}\left(z_i^2 \ge \frac{\sigma^2}{\epsilon'}\right) \\ &\ge 1 - \frac{\mathbb{E}(\|x_i\|_2^2)}{d/\epsilon'} - \frac{\mathbb{E}(z_i^2)}{\sigma^2/\epsilon'} = 1 - 2\epsilon',\end{aligned}$$

using Markov's inequality. Applying the multiplicative Chernoff bound in Lemma C.4.1 to the random variables $(1 - E_i)$, we then obtain

$$\mathbb{P}\left(|S_2| \ge (1-4\epsilon')n\right) \ge \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (1 - E_i) \le 4\epsilon'\right) \ge 1 - \exp(-cn\epsilon'),$$

as claimed. We also define the set of indices $T_0 \subseteq [n]$ such that $S_2 = \{(x_i, y_i)\}_{i \in T_0}$.

Now let $T_{v'} := T_v \cap T_0$ and consider the set $S'_{v'} := \{W_i^{v'}\}_{i \in T_{v'}}$, which we will show is stable with high probability. Note that $|T'_{v'}| \ge (1-5\epsilon')n$. We have the following lemma, proved in Appendix C.10.3:

Lemma C.10.3. *Suppose S'_v is $(C\epsilon', \delta)$-stable with respect to β^* and σ_*^2 such that $|S'_v| \ge (1-\epsilon')n$, and suppose $|S_2| \ge (1-4\epsilon')n$. Suppose $\|v - v'\|_2 \le \eta$ and $\eta = \frac{r\sqrt{\epsilon'}}{f(d/\epsilon')}$, where f is an appropriately defined second-degree polynomial. Then $S'_{v'}$ is $(C\epsilon'/2, \delta')$-stable with respect to β^* and σ_*^2, where $\delta' = \Theta\left(\sqrt{\frac{p\log(pn)}{n}} + \sqrt{\epsilon} + \sqrt{\frac{\log(1/\tau)}{n}}\right)$.*

Finally, we use a union bound to control the failure probability. Combining the error probability for the Chernoff bound for S_2 with the error probabilities for the elements of C_η, we see that the overall probability of error is bounded by

$$\begin{aligned}\exp(-cn\epsilon') + \tau'|\mathcal{C}_\eta| &\le \exp(-cn\epsilon') + \tau' \exp\left(p\log\left(\frac{3r}{\eta}\right)\right) \\ &= \exp(-cn\epsilon') + \tau\exp\left(-C_1 p\log(pn) + p\log\left(\frac{3f(d/\epsilon')}{\sqrt{\epsilon'}}\right)\right) \\ &\le \exp(-cn\epsilon') + \tau\exp\left(-C_1 p\log(pn) + c_1 p\log n + \frac{p}{2}\log\left(\frac{1}{\epsilon'}\right)\right)\end{aligned}$$

$$\leq \exp(-cn\epsilon') + \tau \exp\left(-C_1 p \log(pn) + c_1 p \log n + c_2 p \log n\right),$$

using the choice of η in Lemma C.10.3 and the fact that $\epsilon' = \Omega\left(\frac{p}{n}\right)$ in the last two inequalities. The final expression can be made smaller than 2τ for a sufficiently large choice of C_1, completing the proof.

C.10.3 Proof of Lemma C.10.3

Consider any set $T' \subseteq T_{v'}$ such that $|T'| \geq \left(1 - \frac{C\epsilon'}{2}\right)|T'_{v'}|$, and define $\Delta := \beta^* - v$ and $\Delta' := \beta^* - v'$, so $\Delta' - \Delta = v - v'$. Using the triangle inequality, we write

$$\begin{aligned}
\left\| \frac{1}{|T'|} \sum_{i \in T'} W_i^{v'} - \beta^* \right\|_2 &= \left\| \frac{1}{|T'|} \sum_{i \in T'} v' + x_i x_i^\top (\beta^* - v') + x_i z_i - \beta^* \right\|_2 \\
&= \left\| \frac{1}{|T'|} \sum_{i \in T'} x_i x_i^\top \Delta' + x_i z_i - \Delta' \right\|_2 \\
&\leq \left\| \frac{1}{|T'|} \sum_{i \in T'} x_i x_i^\top \Delta + x_i z_i - \Delta \right\|_2 + \left\| \frac{1}{|T'|} \sum_{i \in T'} x_i x_i^\top (\Delta' - \Delta) \right\|_2 + \|\Delta' - \Delta\|_2 \\
&\leq \left\| \frac{1}{|T'|} \sum_{i \in T'} x_i x_i^\top \Delta + x_i z_i - \Delta \right\|_2 + \frac{d\eta}{\epsilon'} + \eta, \qquad \text{(C.30)}
\end{aligned}$$

where we have used the facts that $\|x_i\|_2 \leq \sqrt{\frac{p}{\epsilon'}}$ for $i \in T_0$ and $\|\Delta' - \Delta\|_2 \leq \eta$ in the last line. Furthermore, note that the first term on the right-hand side of inequality (C.30), which can be written as $\left\| \frac{1}{|T'|} \sum_{i \in T'} W_i^v - \beta^* \right\|_2$, can be upper-bounded by $\sigma_* \delta$ using the stability of the set T_v, since $T' \subseteq T_v$ and

$$|T'| \geq \left(1 - \frac{C\epsilon'}{2}\right)|T'_{v'}| \geq \left(1 - \frac{C\epsilon'}{2}\right)(1 - 5\epsilon')n \geq (1 - C\epsilon)|T_v|,$$

if $C \geq 10$. Thus, we conclude that

$$\left\| \frac{1}{|T'|} \sum_{i \in T'} W_i^{v'} - \beta^* \right\|_2 \leq 2\sigma_* \delta,$$

by choosing $\eta \leq \frac{\sigma_* \delta}{1 + d/\epsilon'}$. Note that since $r = \Theta(\sigma_*)$ and $\delta = \Omega(\sqrt{\epsilon'})$, this may be accomplished with the choice

$$\eta = O\left(\frac{r\sqrt{\epsilon'}}{1 + d/\epsilon'}\right). \qquad \text{(C.31)}$$

We also need to establish a spectral norm bound on the second moment matrix. Denoting

$$\begin{aligned} a_i &:= x_i x_i^\top \Delta + x_i z_i - \Delta, \\ b_i &:= x_i x_i^\top (\Delta' - \Delta), \\ c &:= \Delta - \Delta', \end{aligned}$$

we see that

$$\begin{aligned} &\left\| \frac{1}{|T'|} \sum_{i \in T'} \left(W_i^{v'} - \beta^* \right) \left(W_i^{v'} - \beta^* \right)^\top - \sigma_*^2 I \right\|_2 \\ &= \left\| \frac{1}{|T'|} \sum_{i \in T'} (x_i x_i^\top \Delta' + x_i z_i - \Delta')(x_i x_i^\top \Delta' + x_i z_i - \Delta') - \sigma_*^2 I \right\|_2 \\ &= \left\| \frac{1}{|T'|} \sum_{i \in T'} (a_i + b_i + c)(a_i + b_i + c)^\top - \sigma_*^2 I \right\|_2 \\ &\leq \left\| \frac{1}{|T'|} \sum_{i \in T'} a_i a_i^\top - \sigma_*^2 I \right\|_2 + \left\| \frac{1}{|T'|} \sum_{i \in T'} b_i b_i^\top \right\|_2 + \left\| \frac{1}{|T'|} \sum_{i \in T'} c c^\top \right\|_2 \\ &\qquad + 2 \left\| \frac{1}{|T'|} \sum_{i \in T'} a_i b_i^\top \right\|_2 + 2 \left\| \frac{1}{|T'|} \sum_{i \in T'} a_i c^\top \right\|_2 + \left\| \frac{1}{|T'|} \sum_{i \in T'} b_i c^\top \right\|_2 . \end{aligned} \tag{C.32}$$

By the stability of T_v, we have

$$\left\| \frac{1}{|T'|} \sum_{i \in T'} a_i a_i^\top - \sigma_*^2 I \right\|_2 \leq \frac{\sigma_*^2 \delta^2}{C \epsilon'}.$$

Further note that

$$\begin{aligned} \|a_i\|_2 &\leq \frac{d\eta}{\epsilon'} + \sqrt{\frac{p}{\epsilon'}} \cdot \frac{\sigma}{\sqrt{\epsilon'}} + \eta, \\ \|b_i\|_2 &\leq \frac{d\eta}{\epsilon'}, \\ \|c\|_2 &\leq \eta. \end{aligned}$$

Thus, the right-hand expression in inequality (C.32) may be upper-bounded by

$$\frac{\sigma_*^2 \delta^2}{C \epsilon'} + \frac{p^2 \eta^2}{(\epsilon')^2} + \eta^2 + 2 \left(\frac{d\eta}{\epsilon'} + \eta \right) \left(\frac{d\eta}{\epsilon'} + \sqrt{\frac{p}{\epsilon'}} \cdot \frac{\sigma}{\sqrt{\epsilon'}} + \eta \right) + \frac{2d\eta^2}{\epsilon'}$$

$$\leq \frac{\sigma_*^2\delta^2}{C\epsilon'} + \eta\left(\frac{p^2 r}{(\epsilon')^2} + r + 2\left(\frac{p}{\epsilon'}+1\right)\left(\frac{dr}{\epsilon'} + \frac{\sigma\sqrt{p}}{\epsilon'} + r\right) + \frac{2dr}{\epsilon'}\right)$$
$$\leq \frac{\sigma_*^2\delta^2}{C\epsilon'/2},$$

by choosing

$$\eta = O\left(\frac{r}{p^2/(\epsilon')^2 + 1 + 2(d/\epsilon'+1)(2d/\epsilon'+1) + 2d/\epsilon'}\right), \tag{C.33}$$

using the facts that $r = \Theta(\sigma_*)$ and $\delta = \Omega(\sqrt{\epsilon'})$.

Therefore, we see that defining f appropriately and taking $\eta = \frac{r\sqrt{\epsilon'}}{f(d/\epsilon')}$ satisfies conditions (C.31) and (C.33) simultaneously, completing the proof.

C.11 Additional Simulations

We include additional experiment details in this section. Figure C.1 shows how the choice of the tuning parameter γ in the Huber loss affects the resulting error. We note that Huber regression with filtering is quite robust to the choice of γ.

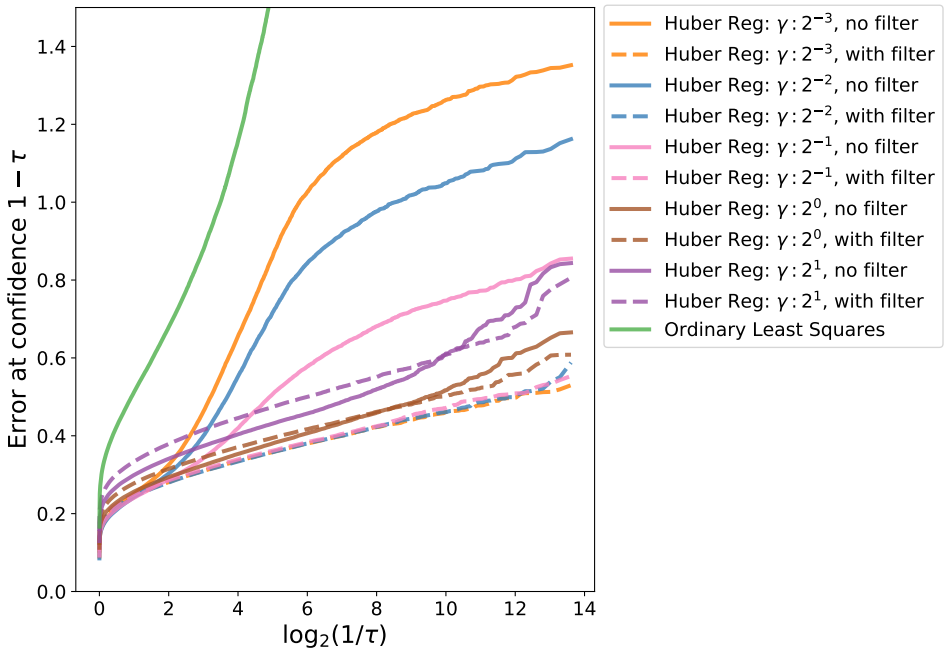

Figure C.1: Plot showing the effect of covariate filtering on Huber regression ($n = 200, p = 40$) for different values of γ. The error is measured in terms of ℓ_2-error, i.e., $\|\widehat{\beta} - \beta^*\|_2$. Solid lines corresponds to "vanilla" version of the estimators (no filtering step), and dashed lines correspond to filtered versions, where the filtering step removes 10 points out of 200 points. We note that the performance of Huber regression with filtering is not greatly affected by the choice of γ.

Figure C.2 shows how the choice of the thresholding parameter m in LTS affects the resulting error. We note that the LTS with filtering is also quite robust to the choice of m.

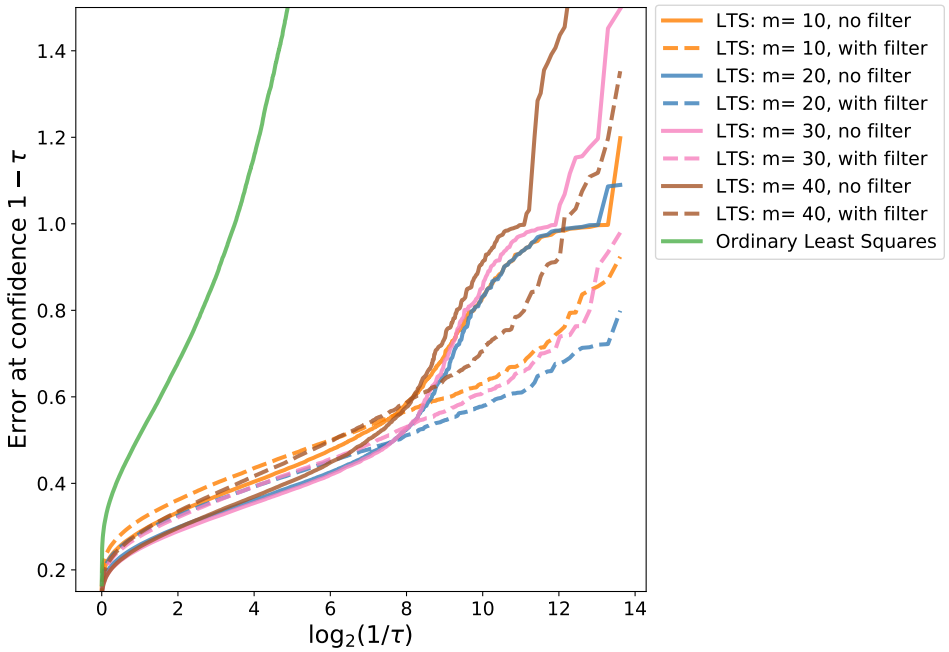

Figure C.2: Plot showing the effect of covariate filtering on LTS regression $(n = 200, p = 40)$ for different values of m. The error is measured in terms of ℓ_2-error, i.e., $\|\widehat{\beta} - \beta^*\|_2$. Solid lines corresponds to "vanilla" version of the estimators (no filtering step), and dashed lines correspond to filtered versions, where the filtering step removes 10 points out of 200 points. We note that the performance of LTS with filtering is not greatly affected by the choice of m.

Appendix

D.1 Additional Technical Facts

Our bounds in Lemma 6.3.7 required the fact below. Here we provide its proof for completeness.

Fact D.1.1. *For any one-dimensional distribution P that matches the first m moments with $\mathcal{N}(0,1)$ and has $\chi^2(P,\mathcal{N}(0,1)) < \infty$ the following identity is true*

$$\chi^2(P,\mathcal{N}(0,1)) = \sum_{i=m+1}^{\infty} \left(\mathop{\mathbb{E}}_{X\sim P}[h_i(X)]\right)^2 .$$

Proof. Let ϕ denote the pdf of the standard one-dimensional Gaussian. For this proof, we use a slightly different definition of the space $L^2(\mathbb{R},\mathcal{N}(0,1))$. We define it as the space of functions for which $\int_{\mathbb{R}} f^2(x)/\phi(x)\mathrm{d}x < \infty$ with the inner product $\langle f,g\rangle := \int_{\mathbb{R}} f(x)g(x)/\phi(x)\mathrm{d}x$ (note the similarity with the definition of χ^2-divergence). The *Hermite functions* (or often called *Hermite-Gauss functions*) $h_i(x)\phi(x)$ for $i = 0,1,\ldots$ form a complete orthonormal basis of the space $L^2(\mathbb{R},\mathcal{N}(0,1))$ with respect to that inner product. It is easy to check that this statement is equivalent to the statement that Hermite polynomials $\{h_i\}_{\mathbb{N}}$ form a complete orthonormal basis of the space of all functions $f:\mathbb{R}\to\mathbb{R}$ for which $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[f^2(x)] < \infty$ (i.e., our old definition of $L^2(\mathbb{R},\mathcal{N}(0,1))$). Since $\chi^2(P,\mathcal{N}(0,1)) < \infty$ we have $P \in L^2(\mathbb{R},\mathcal{N}(0,1))$ and thus we can write $P(x) = \sum_{i=0}^{\infty} a_i h_i(x)\phi(x)$, where $a_i = \mathbb{E}_{X\sim P}[h_i(X)]$. Using the fact that P agrees with the first m moments of $\mathcal{N}(0,1)$ and the property of Hermite polynomials $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[h_i(X)] = \mathbb{I}(i=0)$ we get that $a_0 = \mathbb{E}_{X\sim\mathcal{N}(0,1)}[h_0(X)] = 1$ and $a_i = \mathbb{E}_{X\sim\mathcal{N}(0,1)}[h_i(X)] = 0$ for $0 < i \le m$. Thus

$$P(x) = \phi(x) + \sum_{i=m+1}^{\infty} a_i h_i(x)\phi(x) .$$

The χ^2-divergence can then be written as

$$\chi^2(P,\mathcal{N}(0,1)) = \int_{\mathbb{R}} \frac{(P(x)-\phi(x))^2}{\phi(x)}\mathrm{d}x = \int_{\mathbb{R}} \frac{1}{\phi(x)}\left(\sum_{i=m+1}^{\infty} a_i h_i(x)\phi(x)\right)^2 \mathrm{d}x = \sum_{i=m+1}^{\infty} a_i^2 ,$$

where the last part uses orthonormality of the functions $h_i(x)\phi(x)$. □

We now turn to Claim 6.3.9 which is restated below.

Claim D.1.2. *If $P = \sum_{i=1}^{k} \lambda_i N(\mu_i, \sigma_i^2)$ with $\mu_i \in \mathbb{R}$, $\sigma_i < \sqrt{2}$ and $\lambda_i \geq 0$ such that $\sum_{i=1}^{k} \lambda_i = 1$, we have that $\chi^2(P, \mathcal{N}(0,1)) < \infty$.*

For that we need the following two facts about χ^2-distance between Gaussians. Their proofs can be done by direct calculations.

Fact D.1.3. *Let $k \in \mathbb{Z}_+$, distributions P_i and $\lambda_i \geq 0$, for $i \in [k]$ such that $\sum_{i=1}^{k} \lambda_i = 1$. We have that $\chi^2\left(\sum_{i=1}^{k} \lambda_i P_i, D\right) = \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_i \lambda_i \chi_D(P_i, P_j)$.*

Proof.

$$\begin{aligned}
\chi^2\left(\sum_{i=1}^{k} \lambda_i P_i, D\right) + 1 &= \int_{\mathbb{R}} \left(\sum_{i=1}^{k} \lambda_i P_i(x)\right)^2 / D(x) \mathrm{dx} = \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_i \lambda_j \int_{\mathbb{R}} P_i(x) P_j(x) / D(x) \mathrm{dx} \\
&= \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_i \lambda_j \left(\chi_D(P_i, P_j) + 1\right) = \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_i \lambda_j \chi_D(P_i, P_j) + \left(\sum_{i=1}^{k} \lambda_i\right)^2 \\
&= \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_i \lambda_j \chi_D(P_i, P_j) + 1 \,.
\end{aligned}$$

□

Fact D.1.4.

$$\chi_{\mathcal{N}(0,1)}\left(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)\right) = \frac{\exp\left(-\frac{\mu_1^2(\sigma_2^2 - 1) + 2\mu_1\mu_2 + \mu_2^2(\sigma_1^2 - 1)}{2\sigma_1^2(\sigma_2^2 - 1) - 2\sigma_2^2}\right)}{\sqrt{\sigma_1^2 + \sigma_2^2 - \sigma_1^2\sigma_2^2}} - 1 \,.$$

The proof of Claim D.1.2 then consists of applying Fact D.1.3 and using Fact D.1.4 for each one of the generated terms.

E APPENDIX TO CHAPTER 7

E.1 Proofs of Preliminaries

In this appendix, we provide proofs for the preliminary lemma concerning properties of radially symmetric distributions in Section 7.2, as well as the concentration results used in the paper.

E.1.1 Proof of Lemma 7.2.5

1. Note that $R(f_{x,r})$ can be written as convolution of $\overline{P}$ with indicator function of B_r, both of which are unimodal and radially symmetric. The desired result then follows by Proposition 8 in Li et al. [LMM20], which implies that $R(f_{x,r})$ is also unimodal and radially symmetric.

2. This follows from the nonnegativity of the density.

3. As $\overline{P}$ is radially symmetric, let the density of $\overline{P}$ at x be given by $p(\|x\|)$. R_r^* can be written as $R_r^* = C\int_0^r p(s)s^{d-1}ds$ where C is a constant for a fixed dimension. Define $g(r) := \frac{R_r^*}{Cr^d} = \frac{\int_0^r p(s)s^{d-1}ds}{r^d}$ for $r > 0$. Property (iii) is equivalent to showing that $\frac{d}{dr}g(r) < 0$. By unimodality of $p(\cdot)$, it follows that $g(r) > \frac{p(r)}{d}$. Differentiating $g(\cdot)$, we get

$$\frac{d}{dr}g(r) = \frac{p(r)r^{d-1}r^d - dr^{d-1}\int_0^r p(s)s^{d-1}ds}{r^{2d}} = \frac{p(r) - dg(r)}{r} < 0.$$

4. Note that any r_1-packing of $B(0, r_2 - r_1)$ has the property that all balls in the packing must be entirely contained within the larger ball B_{r_2}. Furthermore, by Lemma 7.2.5(i) above, we know that $R(f_{x,r_1}) \geq R(f_{r_2,r_1})$ when $\|x\|_2 \leq r_2$. Hence, by summing up the densities of all balls in the packing, we obtain

$$R(f_{0,r_2}) \geq P(B_{r_2-r_1}, r_1)R(f_{r_2,r_1}),$$

from which the first inequality follows.

To obtain the second inequality, we use the sphere-packing lower bound

$$P(B_{r_2-r_1}, r_1) \geq N(B_{r_2-r_1}, 2r_1) \geq \left(\frac{r_2 - r_1}{2r_1}\right)^d,$$

where $N(\cdot, \cdot)$ denotes the covering number (cf. Proposition 4.2.12 of Vershynin [Ver18]).

5. The proof of the first inequality is the same as the proof of the corresponding statement in Lemma E.2.1. The second inequality follows by noting that $\mathbb{E}\,\|X_i - \mu\|_2^2 = \mathrm{Tr}(\Sigma_i) = d\sigma_i^2$. By Chebyshev's inequality, we have

$$\tilde{R}_i(f_{0,2\sigma_i\sqrt{d}}) = \mathbb{P}(\|X_i - \mu\|_2 \le 2\sqrt{d}\sigma_i) \ge \frac{3}{4},$$

for each i. Thus, $B_{2\sigma_{(2k)}\sqrt{d}}$ covers at least $\frac{3}{4}$ of the mass of at least $2k$ distributions, implying the desired result.

E.1.2 Proof of Lemma 7.2.8

Recall that $\tilde{R}_i(f) = \mathbb{E}\, f(X_i)$. We define the random variables

$$Y_{f,i} := f(X_i) - \tilde{R}_i(f).$$

Note that $\mathbb{E}_i[Y_{f,i}] = 0$ and $|Y_{f,i}| \le 1$. Furthermore, the variables $(Y_{f,i})_{i=1}^n$ are independent for each fixed f. Let

$$Z := \sup_{f \in \mathcal{H}_r} (R_n(f) - R(f)) = \sup_{f \in \mathcal{H}_r} \frac{1}{n} \sum_{i=1}^n Y_{f,i}.$$

We will apply Lemma E.8.2 to obtain a high-probability upper bound on Z. Here $V = d + 1$, the VC dimension of balls.

Since its application requires a bound on the expectation, we first derive the following lemma:

Lemma E.1.1. *If $nR_r^* \ge 1300V \log n$ with both $n > 1$ and $d \ge 1$, then*

$$\mathbb{E}\, Z \le 72\sqrt{V\frac{R_r^* \log n}{2n}}.$$

Proof. We will use Theorem E.8.3 from Appendix E.8, with $\sigma^2 = \sup_{x,r' \le r} R(f_{x,r'}) = R_r^*$.

In particular, note that since $n\sigma^2 \geq 1300V \log n$, we have

$$\log\left(\frac{4e^2}{\sigma}\right) = \frac{1}{2}\log\left(\frac{16e^4}{\sigma^2}\right) \leq \frac{1}{2}\log\left(\frac{16e^4 n}{1300V\log n}\right) \leq \frac{\log n}{2},$$

so

$$\left(24\sqrt{\frac{V}{5n}\log\left(\frac{4e^2}{\sigma}\right)}\right)^2 = \frac{576V}{5n}\log\left(\frac{4e^2}{\sigma}\right) \leq \frac{576V}{5n}\cdot\frac{\log n}{2} = 57.6V\frac{\log n}{n} \leq \sigma^2.$$

Thus, Theorem E.8.3 is applicable and leads to the following bound:[36]

$$\mathbb{E}\, Z \leq 72\frac{\sqrt{R_r^*}}{\sqrt{n}}\sqrt{V\log\left(\frac{4e^2}{\sigma}\right)} \leq 72\sqrt{\frac{VR_r^*\log n}{2n}}.$$

□

We now apply Theorem 12.9 from Boucheron et al.[BLM13] (stated in Lemma E.8.2 in Appendix E.8) with $W_{i,s} = Y_{i,f}$ and

$$\rho^2 = \sup_{f\in\mathcal{H}_r}\sum_{i=1}^n \mathbb{E}\, Y_{i,f}^2 = \sup_{f\in\mathcal{H}_r}\sum_{i=1}^n \mathbf{Var}[f(X_i)] \leq \sup_{f\in\mathcal{H}_r}\sum_{i=1}^n \mathbb{E}[f(X_i)] = \sup_{f\in\mathcal{H}_r} nR(f) = nR_r^*,$$

where the inequality holds because the variance of a Bernoulli random variable is bounded by its expectation. Hence, using Lemma E.1.1 and the assumption $nR_r^* \geq 1300V\log n$, we have

$$\begin{aligned} v &= 2n\,\mathbb{E}\, Z + \rho^2 \leq 2n\,\mathbb{E}\, Z + nR_r^* \\ &\leq 144\sqrt{0.5VnR_r^*\log n} + nR_r^* \leq nR_r^*\left(144\sqrt{\frac{0.5V\log n}{nR_r^*}} + 1\right) \\ &\leq nR_r^*\left(144\sqrt{\frac{0.5V\log n}{1300V\log n}} + 1\right) < 6nR_r^*. \end{aligned}$$

[36] Note that the definition of Z in Theorem E.8.3 has a factor of $1/\sqrt{n}$ as opposed to the factor of $1/n$ here.

Thus, $\frac{ntR_r^*}{2v} > \frac{t}{12}$, so

$$\log\left(1+2\log\left(1+\frac{ntR_r^*}{2v}\right)\right) \geq \log\left(1+2\log\left(1+\frac{t}{12}\right)\right) \geq \frac{t}{50}, \tag{E.1}$$

using the fact that $t \leq 1$.

Now suppose $nR_r^* \geq C_t \frac{V}{2} \log n$ for the constant $C_t = \left(\frac{144}{t}\right)^2$. Note that for $t \leq 1$, we have $nR_r^* \geq 1300V \log n$, so all the previous results are also valid. Moreover, we have

$$\begin{aligned}\frac{\mathbb{E}\, Z}{0.5tR_r^*} = \frac{n\,\mathbb{E}\, Z}{0.5tnR_r^*} &\leq \frac{72\sqrt{0.5VnR_r^* \log n}}{0.5tnR_r^*} = \frac{144\sqrt{0.5V\log n}}{t\sqrt{nR_r^*}} \\ &\leq \frac{144\sqrt{0.5V\log n}}{t\sqrt{0.5C_tV\log n}} = \frac{144}{t\sqrt{C_t}} < 1.\end{aligned}$$

Now we have all the ingredients required for the application of Theorem 12.9 :

$$\begin{aligned}\mathbb{P}\{Z \geq tR_r^*\} &\leq \mathbb{P}\{Z \geq \mathbb{E}\, Z + 0.5tR_r^*\} \\ &\leq \exp\left(-\frac{ntR_r^*}{4}\log\left(1+2\log\left(1+\frac{ntR_r^*}{2v}\right)\right)\right) \\ &\leq \exp\left(-\frac{1}{200}nt^2R_r^*\right),\end{aligned}$$

where the last inequality follows by inequality (E.1).

An identical argument can be used to upper-bound the quantity

$$\sup_{f\in\mathcal{H}_r}\left(R(f) - R_n(f)\right),$$

concluding the proof.

E.1.3 Proof of Lemma 7.6.1

We begin by proving inequality (7.20). First consider the following peeling lemma, an adaptation of Lemma 3 in Raskutti et al. [RWY10]:

Lemma E.1.2. *Let $A \subseteq \mathbb{R}^p$, and suppose $\{Y_x\}_{x\in A}$ is a collection of random variables indexed by x. Also suppose $g : \mathbb{R} \to \mathbb{R}_+$ is a strictly increasing function such that $\inf_{x\in A} g(h(\|x\|_2)) \geq \mu$, for some $\mu > 0$, and $h : \mathbb{R}_+ \to \mathbb{R}_+$ is a constraint function, and the tail bound*

$$\mathbb{P}\left(\sup_{x\in A: h(\|x\|_2)\leq s} Y_x \geq g(s)\right) \leq 2\exp\big(-cg(s)\big)$$

holds for all $s \in \mathrm{range}(h)$. *Then*

$$\mathbb{P}\left(Y_x \le 2g(h(\|x\|_2)), \quad \forall x \in A\right) \ge 1 - \frac{2\exp(-c\mu)}{1-\exp(-c\mu)}. \tag{E.2}$$

Proof. We define the sets

$$A_m := \left\{x \in A : 2^{m-1}\mu \le g(h(\|x\|_2)) \le 2^m\mu\right\},$$

for $m \ge 1$. By a union bound, we have

$$\mathbb{P}\left(\exists x \in A \text{ s.t. } Y_x > 2g(h(\|x\|_2))\right) \le \sum_{m=1}^{M} \mathbb{P}\left(\exists x \in A_m \text{ s.t. } Y_x > 2g(h(\|x\|_2))\right),$$

where $M = \sup_{m\ge 1} g^{-1}(2^{m-1}\mu) \in \mathrm{range}(h)$.

Further note that if $x \in A_m$ satisfies $Y_x > 2g(h(\|x\|_2))$, then $g(h(\|x\|_2)) \ge 2^{m-1}\mu$, so

$$\begin{aligned}
\mathbb{P}\left(\sup_{x\in A_m} Y_x > 2g(h(\|x\|_2))\right) &\le \mathbb{P}\left(\sup_{x\in A_m} Y_x > 2\cdot 2^{m-1}\mu\right) \\
&\le \mathbb{P}\left(\sup_{x\in A: g(h(\|x\|_2))\le 2^m\mu} Y_x > 2^m\mu\right) \\
&= \mathbb{P}\left(\sup_{x\in A: h(\|x\|_2)\le g^{-1}(2^m\mu)} Y_x > 2^m\mu\right) \\
&\le 2\exp\left(-c\cdot 2^m\mu\right),
\end{aligned}$$

if $m < M$. If $m = M$, the same logic shows that

$$\mathbb{P}\left(\sup_{x\in A_m} Y_x > 2g(h(\|x\|_2))\right) \le \mathbb{P}\left(\sup_{x\in A: h(\|x\|_2)\le \nu} Y_x > 2^m\mu\right),$$

where $\nu = \sup_{x\in A} h(\|x\|_2)$. Furthermore, $2^{m-1}\mu \le g(\nu) \le 2^m\mu$, so the last probability is upper-bounded by

$$\mathbb{P}\left(\sup_{x\in A: h(\|x\|_2)\le \nu} Y_x \ge g(\nu)\right) \le 2\exp(-cg(\nu)) \le 2\exp(-c\cdot 2^{m-1}\mu).$$

It follows that

$$\mathbb{P}\left(\sup_{x\in A_m} Y_x > 2g(h(\|x\|_2))\right) \le 2\exp(-c\cdot 2^{m-1}\mu),$$

for all $m \geq 1$, so summing up over m then gives

$$\mathbb{P}\left(\exists x \in A \text{ s.t. } Y_x > 2g(h(\|x\|_2))\right) \leq \sum_{m=1}^{\infty} 2\exp(-c \cdot 2^{m-1}\mu) \leq \frac{2\exp(-c\mu)}{1-\exp(-c\mu)},$$

implying inequality (E.2). □

We apply Lemma E.1.2 with $A = \{x : \|x\|_2 \leq \bar{r}\}$, and

$$Y_x = |R_n(f_{x,r}) - R(f_{x,r})|, \qquad h(\|x\|_2) = R(f_{x,r}), \qquad g(s) = ts,$$

for fixed values of $\bar{r}, r > 0$ and $t \in (0, 1]$. Clearly, g is monotonically increasing and satisfies $\inf_{x\in A} g(h(\|x\|_2)) \geq tR(f_{\bar{r},r})$. Note that for any $s \in \text{range}(h)$, we have $s = R(f_{x_s,r})$ for some x_s, and

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\in A: h(\|x\|_2)\leq s} |R_n(f_{x,r}) - R(f_{x,r})| \geq g(s)\right) \\
&= \mathbb{P}\left(\sup_{\|x_s\|_2 \leq \|x\|_2 \leq \bar{r}} |R_n(f_{x,r}) - R(f_{x,r})| \geq tR(f_{x_s,r})\right) \\
&\leq 2\exp(-cnR(f_{x_s,r})t^2) \\
&= 2\exp(-cntg(s)),
\end{aligned}$$

assuming $nR(f_{\bar{r},r}) \geq C_t d \log n$, where we use a slight modification of Lemma 7.2.8 where $\mathcal{H}_r$ is the set of balls centered around points in $\{\|x\|_2 \geq \|x_s\|_2\}$. Lemma E.1.2 then implies the desired concentration inequality.

To establish inequality (7.21), note that we can simply use a modification of Theorem 7.2.8, where $\mathcal{H}_r$ is now the set of balls centered around points in $\{\|x\|_2 > \bar{r}\}$.

E.2 Proofs for Univariate Estimators

We begin with the following lemma, also appearing as Lemma 1 in Pensia et al. [PJL19b].

Lemma E.2.1. *We have the following properties:*

(*i*) *For any $r > 0$ and $x, x' \in \mathbb{R}$, if $|x| < |x'|$, then $R(f_{x,r}) \geq R(f_{x',r})$.*

(*ii*) *For any $x \in \mathbb{R}$, if $r < r'$, then $R(f_{x,r}) \leq R(f_{x,r'})$.*

(*iii*) *If $0 < r < r'$, then $\frac{R^*_r}{r} > \frac{R^*_{r'}}{r'}$.*

(*iv*) *If* $0 < r, r'$, *then* $R(f_{r',r}) < \frac{r}{r'} R^*_{r'}$.

(*v*) *If* $1 \le k \le n$, *then* $\frac{k}{n} < R^*_{q_{(2k)}}$ *and* $\frac{k}{n} < R^*_{2\sigma_{(2k)}}$.

Proof. The proofs proceed using simple calculus and algebraic manipulations, relying only on the properties of symmetry and unimodality.

(i) Property (i) follows directly by unimodality and symmetry of $\overline{P}$.

(ii) Property (ii) is true by the non-negativity of density.

(iii) Let $p(x)$ be the density of $\overline{P}$. Then $R^*_x = 2\int_0^x p(y)dy$. Define $g(x) := \frac{R^*_x}{x}$ for $x > 0$. Property (iii) is equivalent to showing that $\frac{d}{dx} g(x) < 0$. By unimodality of $p(\cdot)$, we have $g(x) > 2p(x)$ for $x > 0$. By differentiation, we have

$$\frac{d}{dx} g(x) = \frac{2xp(x) - 2\int_0^x p(y)dy}{x^2} = \frac{2p(x) - g(x)}{x} < 0,$$

as wanted.

(iv) Note that r' can be written as $r' = (K + \alpha)r$, where $K \in \mathbb{N}$ and $\alpha \in [0, 1)$. We need to show that $R^*_{r'} > (K + \alpha) R(f_{r',r})$. We may write

$$\begin{aligned} R^*_{r'} &= 2\int_0^{r'} p(x)dx \\ &= 2\int_0^{\alpha r} p(x)dx + \sum_{k=1}^{K} 2\int_{r'-kr}^{r'-(k-1)r} p(x)dx. \end{aligned}$$

The second term is 0 if $K = 0$. By (iii) above, we have $R^*_{\alpha r} > \alpha R^*_r$. Therefore,

$$\begin{aligned} R^*_{r'} &> 2\alpha \int_0^r p(x)dx + \sum_{k=1}^{K} 2\int_{r'-kr}^{r'-(k-1)r} p(x)dx \\ &> \alpha \int_{r'-r}^{r'+r} p(x)dx + \sum_{k=1}^{K} \int_{r'-r}^{r'+r} p(x)dx \\ &= (\alpha + K) R(f_{r',r}), \end{aligned}$$

where the last inequality again uses unimodality of $\overline{P}$, and the second term is 0 if $K = 0$.

(v) Note that

$$R^*_{q_{(2k)}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(|X_i| \le q_{(2k)}) > \frac{1}{2} \cdot \frac{2k}{n} = \frac{k}{n}.$$

Let $\tilde{R}_i(f)$ be the expectation of f under P_i, i.e., $\tilde{R}_i(f) = \mathbb{E}\, f(X_i)$. For the second inequality, note that by Chebyshev's inequality,

$$\tilde{R}_i(f_{0,2\sigma_i}) = \mathbb{P}(|X_i - \mu| \leq 2\sigma_i) \geq \frac{3}{4},$$

for all i. Therefore, an interval of length $4\sigma_{(2k)}$ covers at least $\frac{3}{4}$ mass of at least $2k$ distributions, implying that

$$R^*_{2\sigma_{(2k)}} = R(f_{0,2\sigma_{(2k)}}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{R}_i(f) \geq \frac{1}{n} \cdot \frac{3 \times 2k}{4} > \frac{k}{n}.$$

□

Lemma E.2.1 shows that we can use $\overline{P}$ as a measure of distance between two intervals. In particular, if two intervals with the same center/radius are close under R, the respective radii/centers must also be close.

E.2.1 Proof of Theorem 7.3.1

We begin with the following result, which follows from Lemma 7.2.8:

Lemma E.2.2. *Let $t \in (0, 1]$, and let r be such that $R^*_r \geq C_{0.5t}\left(\frac{\log n}{n}\right)$. Then with probability at least $1 - 2\exp(-c' n R^*_r t^2)$, we have $R(f_{\widehat{\mu}_{M,r},r}) \geq (1-t)R^*_r$.*

Proof. This will follow from Lemma 7.2.8 by choosing $0.5t$ instead of t. If $R^*_r \geq C_{0.5t}\frac{\log n}{n}$, then with probability $1 - 2\exp(-cnR^*_r t^2/4)$, we have

$$|R_n(f) - R(f)| \leq \frac{tR^*_r}{2},$$

uniformly over $f \in \mathcal{H}_r$. Assume that this event happens. Note that $R(f_{0,r}) = R^*_r$ and $R_n(f_{\widehat{\mu}_{M,r},r}) \geq R_n(f_{0,r})$ by maximality of the modal interval estimator. Since $f_{\widehat{\mu}_{M,r},r}, f_{0,r} \in \mathcal{H}_r$, we have

$$\begin{aligned} R(f_{\widehat{\mu}_{M,r},r}) &\geq R_n(f_{\widehat{\mu}_{M,r},r}) - \frac{tR^*_r}{2} \geq R_n(f_{0,r}) - \frac{tR^*_r}{2} \\ &\geq R(f_{0,r}) - tR^*_r = R^*_r - tR^*_r, \end{aligned}$$

as wanted. □

Lemma E.2.2 states that if r is small, then $R(f_{x,r})$ behaves like a (scaled) density of the mixture distribution $\overline{P}$. Indeed, the density of $\overline{P}$ at the empirical mode, $\widehat{\mu}_{M,r}$, is within a constant factor of the density at μ^*.

Turning to the proof of the theorem, note that by Lemma E.2.1(i), we know that if $R(f_{r',r}) < R(f_{\widehat{\mu}_{M,r},r})$, then $|\widehat{\mu}_{M,r}| \le r'$. Furthermore, taking $t = \frac{1}{2}$ in Lemma E.2.2, we have $R(f_{\widehat{\mu}_{M,r},r}) \ge \frac{R^*_r}{2}$, with probability at least $1-2\exp(-c'nR^*_r/4)$. Thus, inequality (7.3) holds provided $R(f_{r',r}) < \frac{R^*_r}{2}$.

Now suppose Let $r' = \frac{2r}{R^*_r}$. By Lemma E.2.1(iv) and noting that $R^*_{r'} \le 1$, we have

$$R(f_{r',r}) < \frac{r}{r'}R^*_{r'} \le \frac{r}{r'} = \frac{r}{\frac{2r}{R^*_r}} = \frac{R^*_r}{2}.$$

This establishes inequality (7.4).

E.2.2 Proof of Theorem 7.3.3

The proof of Theorem 7.3.3 is similar in spirit to the proof of Theorem 7.3.1. We begin by proving a lemma, which replaces Lemma E.2.2:

Lemma E.2.3. *For $2k \ge C_{0.5t}\log n$ and $t \in (0,1]$, with probability at least $1-2\exp(-c'kt^2)$, we have*

$$R(f_{\widehat{\mu}_{S,k},r_{2k}}) \ge (1-t)R^*_{r_k} = (1-t)\frac{k}{n}.$$

Proof. By assumption, we have $nR^*_{r_{2k}} = 2k \ge C_{0.5t}\log n$. Applying Lemma 7.2.8 with $t = 0.5t$ and $r = r_{2k}$, we know that with probability at least $1-\exp(-c2kt^2/4)$, we have

$$\sup_{x,r\le r_{2k}} R_n(f_{x,r}) - R(f_{x,r}) < \frac{t}{2}R^*_{r_{2k}}.$$

Combined with the guarantee of Lemma 7.4.4, we conclude that

$$R_n(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) - R(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) < \frac{t}{2}R^*_{r_{2k}},$$

with probability at least $1-\exp(-ckt^2/2)-\exp(-k/8)$.

Furthermore, since all the distributions have densities, all the X_i's are distinct with probability 1, so $R_n(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) = \frac{k}{n}$. We thus conclude that

$$\frac{k}{n} - R(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) < \frac{t}{2} \cdot \frac{2k}{n},$$

so $R(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) > (1-t)\frac{k}{n} = (1-t)R^*_{r_k}$. Again using the fact that $\widehat{r}_k \le r_{2k}$, we can use Lemma E.2.1(ii) to conclude that $R(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) \le R(f_{\widehat{\mu}_{S,k},r_{2k}})$, so the required statement holds. □

Let $r' = \frac{2nr_{2k}}{k}$. Taking $t = \frac{1}{2}$ in Lemma E.2.3 and using Lemma E.2.1(i), it suffices to show that $R(f_{r',r_{2k}}) < \frac{k}{2n}$, which follows by Lemma E.2.1(iv) and the fact that $R^*_{r'} \le 1$.

E.2.3 Proof of Theorem 7.3.5

We first prove the following result:

Lemma E.2.4. *With probability at least* $1 - 4\exp(-ck^2/n)$*, both of the following statements hold:*

1. S_k *contains the origin in the sense that* $0 \in [\min(S_k), \max(S_k)]$.
2. *Diam*$(S_k) \le 2r_{2k}$

Proof. The k-median was defined using ψ_n. It is therefore instructive to study the properties of the population-level quantity $\psi(\theta) := \mathbb{E}\,\psi_n(\theta)$. For $\theta > 0$, we have

$$\begin{aligned}\psi(\theta) := \mathbb{E}\,\psi_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathrm{sign}(\theta - X_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(-\theta \le X_i < \theta) = R(f_{0,\theta}) = R^*_\theta.\end{aligned}$$

In particular $\psi(r_k) = R^*_{r_k} = \frac{k}{n}$. Similarly, for $\theta < 0$, we have $\psi(\theta) = -R^*_{|\theta|}$.

1. It suffices to show the events $\widehat{\theta}_{\mathrm{med},k} \le r_{2k}$ and $\widehat{\theta}_{\mathrm{med},-k} \ge -r_{2k}$ hold with the required probability. We will focus only on the error on the positive side, i.e., $\widehat{\theta}_{\mathrm{med},k} > r_{2k}$. The analysis for $\widehat{\theta}_{\mathrm{med},-k} < -r_{2k}$ is similar by symmetry. Recall that $\psi_n(\widehat{\theta}_{\mathrm{med},k}) = \frac{k}{n}$ a.s., so by monotonicity of ψ_n, it follows that

$$\mathbb{P}\left(\widehat{\theta}_{\mathrm{med},k} > r_{2k}\right) \le \mathbb{P}\left(\psi_n(r_{2k}) \le \frac{k}{n}\right)$$

$$= \mathbb{P}\left(\psi_n(r_{2k}) - \psi(r_{2k}) \leq -\frac{k}{n}\right).$$

Since $\psi_n(\cdot) - \psi(\cdot)$ is a centered sum of independent bounded random variables, we may apply Hoeffding's inequality on its negative tail. Therefore,

$$\mathbb{P}\left(\widehat{\theta}_{\text{med},k} > r_{2k}\right) \leq \exp\left(-cn\left(\frac{k}{n}\right)^2\right) \leq \exp(-ck^2/n).$$

2. We bound the probability that $\max(S_k) < 0$; the bound for $\min(S_k) > 0$ is analogous. If $\max(S_k) < 0$, then $\psi_n(0) \geq \frac{k}{n}$ by monontonicity of ψ_n and the fact that $\max(S_k) = \widehat{\theta}_{\text{med},k}$ and $\psi_n(\widehat{\theta}_{\text{med},k}) = \frac{k}{n}$. By Hoeffding's inequality, we then have

$$\begin{aligned}\mathbb{P}\left(\max(S_k) < 0\right) &\leq \mathbb{P}\left(\psi_n(0) \geq \frac{k}{n}\right) = \mathbb{P}\left(\psi_n(0) - \psi(0) \geq \frac{k}{n}\right) \\ &\leq \exp\left(-cn \cdot \frac{k^2}{n^2}\right) = \exp\left(-c\frac{k^2}{n}\right).\end{aligned}$$

□

By Lemma E.2.4 and Theorem 7.3.3, with probability with probability at least $1 - 4\exp(-c\log^2 n)$, both of the following events happen simultaneously:

1. $0 \in [\min(S_k), \max(S_k)]$.
2. $\text{Diam}(S_k) \leq 2r_{k_1}$.

As the set $[\min(S_k), \max(S_k)]$ is convex and 0 belongs to the set, $|\widehat{\mu}_{k_1,k_2}| \leq |\widehat{\mu}_{S,k_2}|$. As $\widehat{\mu}_{k_1,k_2} \in [\min(S_k), \max(S_k)]$, $|\widehat{\mu}_{k_1,k_2}|$ is less than the diameter of S_k. This proves the first inequality of the statement.

Let $r' := \frac{4\sqrt{n}\log n}{k_2} r_{2k_2}$. To prove the second inequality, we break down the analysis in two cases:

Case 1: Suppose $R^*_{r'} \geq \frac{2\log n}{\sqrt{n}}$. This implies that $r_{2k_1} \leq r'$ and thus desired holds. Since the final prediction is always within the set spanned by S_{k_1}, we must have $|\widehat{\mu}_{k_1,k_2}| \leq r'$ with probability at least $1 - 4\exp(-c\log^2 n)$.

Case 2: Suppose $R^*_{r'} < \frac{2\log n}{\sqrt{n}}$. We will first show that $|\widehat{\mu}_{S,k_2}| \leq r'$. Similar to the proof of Theorem 7.3.3, it suffices to show that $R(f_{r',r_{2k_2}}) < \frac{k_2}{2n}$. Indeed, we have by

Lemma E.2.1(iv)

$$R(f_{r',r_{2k_2}}) < \frac{r_{2k_2}}{r'} R^*_{r'} < \frac{1}{\frac{4\sqrt{n}\log n}{k_2}} \frac{2\log n}{\sqrt{n}} = \frac{k_2}{2n},$$

with probability at least $1 - 2\exp(-c'k_2)$.

Altogether, we conclude that $|\widehat{\mu}_{k_1,k_2}| \leq r'$, with probability at least $1 - 2\exp(-c'k_2) - 4\exp(-c\log^2 n)$.

E.3 Proofs for Examples

In this appendix, we provide the proofs for the propositions regarding the examples discussed in Section 7.3.1.

E.3.1 Proof of Proposition 7.3.9

Using the symmetry and unimodality of $\overline{p}$, we have the following relation:

$$2\overline{p}(0)r \geq R(f_{[0,r]}) \geq 2r\overline{p}(r).$$

Using the first inequality above and choosing $r = r_k$, we obtain $r_k \geq \frac{k}{2\overline{p}(0)}$. The second inequality implies that if $2\overline{p}(y)y \geq \frac{k}{n}$, then $r_k \leq y$. In the remainder of the proof, we will show the bounds for each example using this approach:

1. The lower bound follows by noting that the density at 0 is $\frac{1}{\sqrt{2\pi}\sigma}$. As a result, $r_{\log n} \geq \frac{\sqrt{2\pi}\sigma\log n}{2n}$. The upper bound follows by noting that density at $x = |\sigma|$ is within constant factor of the density at 0. Let $r = (\sigma\sqrt{2\pi e}\log n)/n$. For large enough n, we have that $r \leq \sigma$. Thus

$$2\overline{p}(r)r \geq 2\overline{p}(\sigma)r = 2\frac{e^{-1/2}}{\sqrt{2\pi}\sigma}\frac{\sigma\sqrt{2\pi e}\log n}{n} = \frac{2\log n}{n}.$$

   Therefore, $r_k \leq (\sigma\sqrt{2\pi e}\log n)/n$.

2. The lower bound follows by noting that the density at $x = 0$ is

$$\overline{p}(0) = \left(\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}cin}\right) = \Theta\left(\frac{\log n}{cn}\right),$$

where we use that $\log n \leq \sum_{i=1}^{n} i^{-1} \leq (\log n + 1)$. Thus $r_{\log n} \geq \frac{\log n}{2n\overline{p}(0)} = \Theta(1)$. The upper bound follows by noting that the density at $x = 1$ is

$$\overline{p}(1) = \left(\sum_{i=1}^{n} \frac{e^{-\frac{1}{i^2c^2}}}{\sqrt{2\pi}cin}\right) \geq \left(\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}cin}\left(1 - \frac{1}{i^2c^2}\right)\right) = \overline{p}(0) - \frac{1}{\sqrt{2\pi}c^3n}\sum_{i=1}^{n}\frac{1}{i^3},$$

where the inequality uses that for all $x \in \mathbb{R}$, $e^x \geq 1 + x$. As $\sum_{i=1}^{n} i^3$ converges, we let $C = \lim_n \sum_{i=1}^{n} \frac{1}{i^3}$. We thus have that

$$2\overline{p(1)}1 \geq 2\left(\overline{p}(0) - \frac{C}{c^3n}\right) \geq \frac{2\log n}{\sqrt{2\pi}n}\left(\frac{1}{c} - \frac{C}{c^3\log n}\right).$$

This last expression is greater than $(\log n)/n$, when c is less than (say) $\sqrt{1/2\pi}$ and n is large enough such that $\log n > 2C/c^2$.

3. We first consider the case $\alpha \geq 1$. The lower bound follows by noting that the density at 0 is

$$\frac{c\log n}{n}\frac{1}{\sqrt{2\pi}} + \frac{n - c\log n}{n}\frac{1}{n^\alpha} = \Theta\left(\frac{\log n}{n}\right).$$

The upper bound follows from the fact that at least $c\log n$ distributions have variance 1. Thus the interval $[-1, 1]$ contains more than 0.6 probability of at least $c\log n$ distributions. As $R(f_{0,1}) \geq 0.6c(\log n)/n$, which is larger than $(\log n)/n$ for $c \geq 5/3$, implying that $r_{\log n} \leq 1$.

We now consider the case when $\alpha < 1$. The density at 0 is

$$\frac{c\log n}{n}\frac{1}{\sqrt{2\pi}} + \frac{n - c\log n}{n}\frac{1}{n^\alpha} = \Theta\left(\frac{1}{n^\alpha}\right),$$

which implies the desired upper bound. For the desired lower bound, we note that the density at $x = 1$ is also $\Theta\left(\frac{1}{n^\alpha}\right)$. Using a similar calculation to that of Example 1 above, we get the desired upper bound on r_k.

E.3.2 Proof of Proposition 7.3.10

Since $r = r_{C\log n}$, we have $R_r^* = \frac{C\log n}{n}$. By inequality (7.4) of Theorem 7.3.1, we have

$$|\widehat{\mu}_{M,r}| \leq \frac{2nr_{C\log n}}{C\log n}, \tag{E.3}$$

w.h.p.

1. Analogously to Proposition 7.3.9, we have $r_{C\log n} = \Theta\left(\frac{C\sigma\log n}{n}\right)$. Inequality (E.3) then gives the result.

2. The bound of $\tilde{O}(n)$ follows by inequality (E.3) and noting that $r_{C\log n} = O1)$ for a fixed C and sufficiently small $c > 0$. We now focus on how to obtain the tighter bound of $On^{\epsilon})$ for an $\epsilon > 0$, using inequality (7.3).

   Let $\tilde{R}_i(f)$ be the expectation of f under P_i, i.e., $\tilde{R}_i(f) = \mathbb{E}\, f(X_i)$. Fix an $\epsilon > 0$. Let $r' = n^{\epsilon}$ and $r = 1$. Then it suffices to show that $R^*_r - R(f_{r',r}) \geq C' R^*_r$ where $C' > 0$ might depend on ϵ but not on n.

   We will show that

   a) $R^*_r - R(f_{r',r}) \geq c_1 \sum_{i\leq \frac{r'}{10c}} \tilde{R}_i(f_{0,r})$,

   b) $\sum_{i\leq \frac{r'}{5c}} \tilde{R}_i(f_{0,r}) \geq c_2 n R^*_r$.

   To derive the first inequality, note that

$$\begin{aligned}
nR^*_r - nR(f_{r',r}) &\geq \sum_{i\leq \frac{r'}{10c}} R(f_{0,1}) - R(f_{r',1}) \\
&\geq \sum_{i\leq \frac{r'}{10c}} 2\int_0^1 \frac{1}{\sqrt{2\pi}ci}\left(e^{-\frac{x^2}{2c^2i^2}} - e^{-\frac{(0.5r'+x)^2}{2c^2i^2}}\right) dx \\
&\geq \sum_{i\leq \frac{r'}{10c}} 2\int_0^1 \frac{(1 - e^{-\frac{0.25r'^2}{2c^2i^2}})}{\sqrt{2\pi}ci} e^{-\frac{x^2}{2c^2i^2}}\, dx \\
&\geq (1 - e^{-10}) \sum_{i\leq \frac{r'}{10c}} \tilde{R}_i(f_{0,r}).
\end{aligned}$$

   Now it remains to show that $\sum_{i\leq \frac{r'}{10c}} \tilde{R}_i(f_{0,r}) \geq c_2 R^*_r$. First note that $nR^*_r \leq \frac{\log n}{c}$. Hence,

$$\sum_{i\leq \frac{r'}{10c}} \tilde{R}_i(f_{0,1}) \geq \sum_{i:\frac{1}{c}<i\leq \frac{r'}{10c}} \tilde{R}_i(f_{0,1}) \geq \sum_{i:\frac{1}{c}<i\leq \frac{r'}{10c}} \frac{2e^{-0.5}}{\sqrt{2\pi}ci} \geq c_3 \log\left(\frac{r'}{10e}\right) \geq c_4 \log n^{\epsilon} \geq c_5 \epsilon n R^*_r.$$

3. For $\alpha < 1$, let $r' = \Theta\left(n^{\alpha}\right)$. Then it is easy that $R(f_{r',r}) \leq \frac{R^*_r}{2}$. This follows by observing that the density of a Gaussian distribution decreases by more than half at a distance of σ from the mean.

For $\alpha \geq 1$, let $r' = 10$. Then $R^*_r \geq 0.5\frac{C \log n}{n}$, as a Gaussian distribution contains about 0.68 mass within 1 standard deviation of the mean. Moreover,

$$R(f_{r',r}) \leq 0.1\frac{C \log n}{n} + \frac{n}{\sqrt{2\pi}n^\alpha} \leq 0.2\frac{C \log n}{n} \leq \frac{R^*_r}{2}.$$

Inequality (7.3) then implies the result.

E.3.3 Proof of Proposition 7.3.13

In the following, we will show the bounds on $r_{2\sqrt{n}\log n}$, which gives us the result:

1. As in the proof of Proposition 7.3.9, we have $r_k = \Theta\left(\frac{\sigma k}{n}\right)$ for small k.
2. By Lemma E.2.1(i), we have $r_{2\sqrt{n}\log n} \leq 2\sigma_{(4\sqrt{n}\log n)} = O\left(\sqrt{n}\log n\right)$.
3. Note that for any fixed k, the value of r_k for Example 7.3.8 is smaller than the value of r_k for Example 7.3.6 with $\sigma = n^\alpha$. Thus, we have $r_{2\sqrt{n}\log n} = O\left(\frac{n^\alpha\sqrt{n}\log n}{n}\right) = O\left(n^{\alpha-0.5}\log n\right)$.

E.3.4 Proof of Proposition 7.3.16

We first provide the main steps of the proof. Proofs of supporting lemmas are contained in further sub-sections.

E.3.4.1 Main Argument

Proof. (Proof of Proposition 7.3.16) Let W be a generic random variable with distribution Q_n as defined in Example 7.3.15. Let $A = [-2, 2]$. Consider two disjoint set of hypothesis classes $\mathcal{K}$ and $\mathcal{J}$, with $\mathcal{K} = \{f_{x,1} : x \in A\}$ and $\mathcal{J} = \{f_{x,1} : x \notin A\}$. The hypothesis class $\mathcal{J}$ contains the intervals that are far from 0. Define the following random variables:

$$Z_1 = \sup_{f\in\mathcal{K}} R_n(f), \qquad Z_2 = \sup_{f\in\mathcal{J}} R_n(f).$$

We would show that with constant non-zero probability: (i) $Z_1 < Z_2$ and (ii) the maximum is achieved in Z_2 at intervals that are far from 0.

Note that $R^*_1 = \sup_{f\in\mathcal{K}} R(f) = \Theta\left(n^{-\alpha}\right)$. Define $R^*_{\mathcal{J}} := \sup_{f\in\mathcal{J}} R(f)$. Note that supremum is achieved in both the cases and $R^*_{\mathcal{J}} < R^*_1$. Moreover, we have the following straightforward relations:

1. $2R_1^* \geq \mathbb{P}(W \in A) \geq R_1^*$.
2. $nR_{\mathcal{J}}^* = \Theta\left(n^{1-\alpha}\right)$.
3. $\mathbb{P}(W \in A)\sqrt{nR_{\mathcal{J}}^*} = O1)$.
4. For every constant C', there exists another constant $C > 0$ such that

$$R_{\mathcal{J}}^* + C\left(\sqrt{\frac{R_{\mathcal{J}}^*}{n}}\right) \geq R_1^* + C'\left(\sqrt{\frac{R_1^*}{n}}\right).$$

These relations suffice for showing that $Z_1 < Z_2$ with constant probability. To this end, we would show that with constant probability both (1) $Z_1 = R_1^* + O\left(\sqrt{\frac{R_1^*}{n}}\right)$, and (2) $Z_2 \geq R_{\mathcal{J}}^* + C\left(\sqrt{\frac{R_{\mathcal{J}}^*}{n}}\right)$, for any $C > 0$. Note that these events are dependent and thus we'd use the following lemma, which shows that conditioned on the inclusion of points in each of two disjoint intervals, the distributions of the histograms on each of the intervals behave independently:

Lemma E.3.1. *Let $\{x_1, \ldots, x_n\}$ be i.i.d. draws from a distribution with density p_i. Consider two disjoint intervals A and B. For any two disjoint subsets $S, T \subseteq \{1, \ldots, n\}$, we use x_S to denote the vector $(x_i : i \in S)$, and we define x_T similarly. Let E denote the event that $x_i \in A$ for all $i \in S$, and $x_i \in B$ for all $\in T$. Then for $x_S \subseteq A$ and $x_T \subseteq B$, we have*

$$p_{S,T}(x_S, x_T \mid E) = p_S(x_S \mid E)p_T(x_T \mid E).$$

Furthermore,

$$p_S(x_S \mid E) = \prod_{i \in S} \frac{p_i(x_i)}{\mathbb{P}(X_i \in A)}, \quad \textit{and}$$
$$p_T(x_T \mid E) = \prod_{i \in T} \frac{p_i(x_i)}{\mathbb{P}(X_i \in B)}$$

are the joint densities of independent draws from the renormalized distributions of the points lying in each interval.

Let $S \subset \{1, \ldots, n\}$ be an index set. For a fixed index set S, let the event E_S be $E_S = \{X_S \subset A, X_{S^c} \subset A^c\}$, where X_S is the vector $(X_i : i \in S)$ and A is defined above.

Conditioned on E_S, Lemma E.3.1 states that X_i's are independent. Thus conditioned on E_S, the random variables Z_1 and Z_2 are independent.

Lemma E.3.2. *Consider the setting in Proposition 7.3.16. Let $S \subset [n]$ be such that $|S| \leq n\,\mathbb{P}(A)$. Conditioned on the event E_S, we have that for some $C' > 0$,*

$$Z_1 \leq R_1^* + C'\sqrt{\frac{R_1^*}{n}}$$

with a constant nonzero probability.

Lemma E.3.3. *Consider the setting in Proposition 7.3.16. Let $S \subset [n]$ be such that $|S^c| \geq n\,\mathbb{P}(A^c)$. Conditioned on the event E_S, we have that for all $C > 0$,*

$$Z_2 \geq R_{\mathcal{J}_n}^* + C\left(\sqrt{\frac{R_{\mathcal{J}_n}^*}{n}}\right)$$

with a constant, nonzero probability depending on the constant C.

Lemma E.3.4. *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, where P is a uniform distribution over $[-b, -a] \bigcup [a, b]$ for some $0 \leq a < b$. For a $r > 0$, let $Z = \sup_{f \in \mathcal{H}_r} R_n(f)$ and $k \in \mathbb{N}$ such that $E = \{Z = k\}$ is an event of nonzero probability. If $\frac{b-a}{r} > C$, then*

1. $\mathbb{P}(|\widehat{\mu}_{M,r}| \geq \frac{b-a}{2}) \geq c > 0$.
2. $\mathbb{P}(|\widehat{\mu}_{M,r}| \geq \frac{b-a}{2} | Z \geq k) \geq c > 0$.

Lemmas E.3.2, E.3.3, and E.3.4 give us the required lower bound on the probability of error. Let $\widehat{\mu}_{M,1,\mathcal{J}} := \arg\max_{f \in \mathcal{J}} R_n(f)$. Clearly, we can write

$$\begin{aligned}
\mathbb{P}\left\{|\widehat{\mu}_{M,1}| \geq \frac{n^\alpha}{2}\right\} &= \mathbb{P}\left\{Z_1 < Z_2, |\widehat{\mu}_{M,1,\mathcal{J}}| \geq \frac{n^\alpha}{2}\right\} \\
&= \sum_{S \subset [n]} \mathbb{P}(E_S)\,\mathbb{P}\left(Z_1 \leq Z_2, |\widehat{\mu}_{M,1,\mathcal{J}}| \geq \frac{n^\alpha}{2} \middle| E_S\right) \\
&\geq \sum_{S \subset [n]: |S| \leq n\,\mathbb{P}(A)} \mathbb{P}(E_S)\,\mathbb{P}\left(Z_1 \leq nR_1^* + C\sqrt{nR^*}, Z_2 \geq nR_1^* + C\sqrt{nR_1^*}, |\widehat{\mu}_{M,1,\mathcal{J}}| \geq \frac{n^\alpha}{2} \middle| E_S\right).
\end{aligned}$$

Furthermore, note that since Z_1 is computed over the points lying in A and Z_2 and $\widehat{\mu}_{M,1,\mathcal{J}}$ is computed over the points lying in A^c, Lemma E.3.1 implies that

$$\begin{aligned}
&\mathbb{P}\left(Z_1 \leq nR_1^* + C\sqrt{nR_1^*}, Z_2 \geq nR_1^* + C\sqrt{nR_1^*}, |\widehat{\mu}_{M,1,\mathcal{J}}| \geq \frac{n^\alpha}{2} \middle| E_S\right) \\
&\quad = \mathbb{P}\left(Z_1 \leq nR_1^* + C\sqrt{nR_1^*} \middle| E_S\right)\mathbb{P}\left(Z_2 \geq nR_1^* + C\sqrt{nR_1^*}, |\widehat{\mu}_{M,1,\mathcal{J}}| \geq \frac{n^\alpha}{2} \middle| E_S\right)
\end{aligned}$$

$$= \mathbb{P}\left(Z_1 \le nR_1^* + C\sqrt{nR_1^*}\middle| E_S\right)\mathbb{P}\left(Z_2 \ge nR_1^* + C\sqrt{nR_1^*}\middle| E_S\right)$$
$$\cdot\,\mathbb{P}\left(|\widehat{\mu}_{M,1,\mathcal{J}}| \ge \frac{n^\alpha}{2}\middle| Z_2 \ge nR_1^* + C\sqrt{nR_1^*}, E_S\right).$$

Finally, note that conditioned on E_S, the points in A^c are certainly still uniformly distributed by the construction. Hence, we can apply Lemmas E.3.2, E.3.3, and E.3.4 to lower-bound each of the three factors by a constant. We conclude that

$$\mathbb{P}\left\{|\widehat{\mu}_{M,r}| \ge \frac{n^\alpha}{2}\right\} \ge \sum_{S\subset[n]:|S|\le n\,\mathbb{P}(A)} \mathbb{P}(E_S)\Theta(1) = \Theta(1),$$

where the final equality uses the fact that for $X \sim \text{Bin}(n,p)$, we have $\mathbb{P}(X \le \mathbb{E}\,X) = \Theta(1)$. This concludes the proof of Proposition 7.3.16. □

E.3.4.2 Proof of Lemma E.3.1

Proof. (Proof of Lemma E.3.1) Clearly, we have

$$p_{S,T}(x_S, x_T \mid E) = \frac{p_{S,T}(x_S,x_T)}{\mathbb{P}(E)} = \frac{\prod_{i\in S} p_i(x_i)\prod_{j\in T} p_i(x_j)}{\mathbb{P}(E)}.$$

Similarly, we may write

$$p_S(x_S \mid E) = \frac{p_i(x_S)\prod_{j\in T}\mathbb{P}(X_j\in B)}{\mathbb{P}(E)},$$
$$p_T(x_T \mid E) = \frac{p_i(x_T)\prod_{i\in S}\mathbb{P}(X_i\in A)}{\mathbb{P}(E)}.$$

Using the fact that

$$\mathbb{P}(E) = \prod_{i\in S}\mathbb{P}(X_i\in A)\prod_{j\in T}\mathbb{P}(X_j\in B)$$

implies the desired statements. □

E.3.4.3 Proof of Lemma E.3.2

Proof. (Proof of Lemma E.3.2) Throughout the whole proof, we will condition on the set E_S. Conditioned on E_S, Lemma E.3.1 states that X_S is a vector of $|S|$ i.i.d. points with distribution, say, $Q_{n|A}$. Under $Q_{n|A}$, $\sup_{f\in\mathcal{K}} R(f) = \frac{R_1^*}{\mathbb{P}(W\in A)} \ge \frac{1}{2}$.

Using Theorem E.8.4 (Theorem 8.3.23 in Vershynin[Ver18]), we get that

$$\mathbb{E}\left[\left|\sup_{f\in\mathcal{K}}\sum_{i\in S} f(X_i) - \mathbb{E}[f(X_i)|E_S]\right|\right] \leq C\sqrt{|S|} \leq C\sqrt{2|S|\frac{R_1^*}{\mathbb{P}(W\in A)}},$$

where the last step uses that $2R_1^* \geq \mathbb{P}(W \in A)$. Thus, with constant positive probability,

$$\begin{aligned} Z_1 = \sup_{f\in\mathcal{K}}\sum_i f(X_i) &\leq |S|\frac{R_1^*}{\mathbb{P}(A)} + C'\sqrt{|S|\frac{R_1^*}{\mathbb{P}(A)}} \\ &\leq nR_1^* + C'\sqrt{nR_1^*}, \end{aligned}$$

where we use Markov's inequality and the assumption that $|S| \leq n\,\mathbb{P}(A)$.

□

E.3.4.4 Proof of Lemma E.3.3

Proof. (Proof of Lemma E.3.3) We will condition on the event E_S throughout the proof. Once we have conditioned on E_S, there are $|S^c|$ points distributed over A^c according to Lemma E.3.1, i.e., i.i.d. with a uniform distribution, say, $Q_{n|A^c}$. Consider a fixed function $f \in \mathcal{J}$. As the distribution is uniform, $R(f) = R_{\mathcal{J}}^*$.

For each $i \in S^c$, let $Y_i = f(X_i) - \frac{R_{\mathcal{J}}^*}{\mathbb{P}(A^c)}$. Y_i's are centered i.i.d. Bernoulli random variables. We calculate the following quantities required for the Berry-Esseen Theorem,

$$\begin{aligned} \mathbb{E}[Y_i] &= 0 \\ \mathbf{Var}[Y_i] &= \frac{R_{\mathcal{J}}^*}{\mathbb{P}(A^c)}\left(1 - \frac{R_{\mathcal{J}}^*}{\mathbb{P}(A^c)}\right) \geq \frac{R_{\mathcal{J}}^*}{2\,\mathbb{P}(A^c)} \\ \mathbb{E}\,|Y_i|^3 &= \frac{R_{\mathcal{J}}^*}{\mathbb{P}(A^c)}\left|1 - \frac{R_{\mathcal{J}}^*}{\mathbb{P}(A^c)}\right|^3 + \left(1 - \frac{R_{\mathcal{J}}^*}{\mathbb{P}(A^c)}\right)\left|\frac{R_{\mathcal{J}}^*}{\mathbb{P}(A^c)}\right|^3 \\ &\leq \frac{R_{\mathcal{J}}^*}{\mathbb{P}(A^c)} + \left(\frac{R_{\mathcal{J}}^*}{\mathbb{P}(A^c)}\right)^3 \leq 2\frac{R_{\mathcal{J}}^*}{\mathbb{P}(A^c)} \end{aligned}$$

By the Berry-Esseen Theorem [Ver18], we have

$$\begin{aligned} \mathbb{P}\left\{\frac{\sum_{i\in S^c} Y_i}{\sqrt{|S^c|\,\mathbf{Var}[Y_i]}} \geq t\right\} &\geq \phi(t) - \frac{\mathbb{E}\,|Y_i|^3}{\sqrt{\mathbf{Var}[Y_i]^3|S^c|}} \geq \phi(t) - \frac{\frac{2R_{\mathcal{J}}^*}{\mathbb{P}(A^c)}}{\sqrt{\frac{\left(R_{\mathcal{J}}^*\right)^3}{8\,\mathbb{P}(A^c)^3}n\,\mathbb{P}(A^c)}} \\ &\geq \phi(t) - \frac{c'}{\sqrt{nR_{\mathcal{J}}^*}} = \phi(t) - o_n(1), \end{aligned}$$

where $\phi(t) := \mathbb{P}(g \le t)$ and $g \sim \mathcal{N}(0,1)$. Therefore,

$$\begin{aligned}
\mathbb{P}\left\{Z_2 \ge R^*_{\mathcal{J}} + C\left(\sqrt{\frac{R^*_{\mathcal{J}}}{n}}\right)\right\} &\ge \mathbb{P}\left\{\sum_{i\in S^c} f(X_i) \ge nR^*_{\mathcal{J}} + C\sqrt{nR^*_{\mathcal{J}}}\right\} \\
&= \mathbb{P}\left\{\sum_{i\in S^c} Y_i \ge |S|R^*_{\mathcal{J}} + C\sqrt{nR^*_{\mathcal{J}}}\right\} \\
&= \mathbb{P}\left\{\frac{1}{\sqrt{|S^c|\,\mathbf{Var}[Y_i]}}\sum_{i\in S^c} Y_i \ge \frac{|S|R^*_{\mathcal{J}} + C\sqrt{nR^*_{\mathcal{J}}}}{\sqrt{|S^c|\,\mathbf{Var}[Y_i]}}\right\} \\
&\ge \phi\left(\frac{|S|R^*_{\mathcal{J}} + C\sqrt{nR^*_{\mathcal{J}}}}{\sqrt{|S^c|\,\mathbf{Var}[Y_i]}}\right) - o_n(1) \\
&\ge \phi\left(\frac{n\,\mathbb{P}(A)R^*_{\mathcal{J}} + C\sqrt{nR^*_{\mathcal{J}}}}{\sqrt{n\,\mathbb{P}(A^c)\frac{R^*_{\mathcal{J}}}{2\,\mathbb{P}(A^c)}}}\right) - o_n(1) \\
&\ge \phi\left(\mathbb{P}(A)\sqrt{nR^*_{\mathcal{J}}} + \sqrt{2}C\right) - o_n(1) \ge c\phi\left(C' + \sqrt{2}C\right)
\end{aligned}$$

where we use that for $\alpha \ge \frac{1}{3}$, $\mathbb{P}(A)\sqrt{nR^*_{\mathcal{J}}} = \Theta\left(n^{-\alpha+\frac{1-\alpha}{2}}\right) = O1)$. □

E.3.4.5 Proof of Lemma E.3.4

Proof. (Proof of Lemma E.3.4) Let $\mathcal{H}$ be the set of intervals of width equal to $2r$. Currently, the intervals near the endpoints have less probability mass. We will replace such intervals with bigger intervals to make the process symmetric. First consider the intervals near $\pm a$ which have less probability mass: we can instead focus on bigger intervals to include the middle interval $[-a, a]$. Let $\mathcal{J} := \{⊮_{[x,y]} : |x - y| = 2r + 2(b-a), |x + a| \le 2r\}$. Next we can consider warping the number line and "joining" the two endpoints, i.e., let $\mathcal{K} := \{⊮_{[-\infty,x]\cup[y,\infty]} : 0 \le b - y \le 2r, 0 \le x + b \le 2r, y - x = 2b - 2r\}$.

Let $\mathcal{H}' := \mathcal{J} \cup \mathcal{K} \cup \mathcal{H} \setminus \{f \in \mathcal{H} : R(f) < \frac{2r}{2(b-a)}\}$ and $\widehat{\mu}'_{M,r} = \arg\max_{f\in\mathcal{H}'} R_n(f)$. Note that every function in $\mathcal{H}'$ contains equal mass and the distribution is uniform. Moreover, for $|x| \in [\frac{b-a}{2}, \frac{3(b-a)}{4}]$, $f_{x,r} \in \mathcal{H}' \cap \mathcal{H}$ because $b - a \ge Cr$. Thus we have not removed a lot of functions from $\mathcal{H}$.

The problem of the location of $\widehat{\mu}'_{M,r}$ is equivalent to a uniform distribution on a circle of circumference $2(b-a)$, where we form the circle by joining $-a$ and a at a single point, and join $-b$ to b. By symmetry, we obtain that $|\widehat{\mu}'_{M,r}|$ is uniform on $[a, b]$. Thus

$\mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in [\frac{b-a}{2}, \frac{3(b-a)}{4}]\right) = \frac{1}{4}$.

$$\begin{aligned}\mathbb{P}\left(|\widehat{\mu}_{M,r}| \geq \frac{b-a}{2}\right) &\geq \mathbb{P}\left(|\widehat{\mu}_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right]\right) \\ &\geq \mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right]\right) = \frac{1}{4}.\end{aligned}$$

This proves the first statement. Now, we consider the case when we condition on the value of Z. Note that if $|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right]$, then $Z = Z'$.

$$\begin{aligned}\mathbb{P}\left(|\widehat{\mu}_{M,r}| \geq \frac{b-a}{2}\middle| Z \geq k\right) &\geq \mathbb{P}\left(|\widehat{\mu}_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right]\middle| Z \geq k\right) \\ &\geq \mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right]\middle| Z \geq k\right) \\ &= \frac{\mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right], Z \geq k\right)}{\mathbb{P}\left(Z \geq k\right)} \\ &\geq \frac{\mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right], Z \geq k\right)}{\mathbb{P}\left(Z' \geq k\right)} \\ &= \frac{\mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right], Z' \geq k\right)}{\mathbb{P}\left(Z' \geq k\right)} \\ &= \mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right]\middle| Z' \geq k\right) = \frac{1}{4}\end{aligned}$$

where we use the following Lemma E.3.5 for independence of $\widehat{\mu}'_{M,r}$ and Z'. □

Lemma E.3.5. *Suppose $X_1, \ldots, X_n$ are i.i.d. uniform points on a circle. Let E be the event that the maximum number of points contained in an arc of a certain length is equal to k. Then the joint distribution $p(x_1, \ldots, x_n)$ is rotationally invariant.*

Proof. Suppose without loss of generality that the circle has circumference 1. Note that the law of $(X_1, \ldots, X_n)$ can be equivalently generated as follows: First generate $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} Unif[0, 1]$. Next, generate $R \sim Unif[0, 1]$, and define $X_i = Y_i + R$ for all $1 \leq i \leq n$, where the addition is taken modulo 1. We want to show that

$$p(x_1, \ldots, x_n \mid E) = p(x_1 + r, \ldots, x_n + r \mid E) \tag{E.4}$$

for any $r \in [0, 1]$, where addition is again taken modulo 1. Clearly, it suffices to consider configurations $(x_1, \ldots, x_n)$ that are consistent with E.

We can calculate

$$p(x_1, \ldots, x_n \mid E) = \frac{\int_{E'} p(x_1, \ldots, x_n, y_1, \ldots, y_n) dy}{P(E)},$$

where the integral is taken over the region of $[0,1]^n$ containing points $(y_1, \ldots, y_n)$ that can be obtained from $(x_1, \ldots, x_n)$ via some rotation. Importantly, note that

$$p(x_1, \ldots, x_n, y_1, \ldots, y_n) = p(x_1, \ldots, x_n \mid y_1, \ldots, y_n) p(y_1, \ldots, y_n) = p(y_1, \ldots, y_n),$$

since R is uniform, so we have

$$p(x_1, \ldots, x_n \mid E) = \frac{\int_{E'} p(y_1, \ldots, y_n) dy}{P(E)}.$$

Similarly, we can write

$$p(x_1 + r, \ldots, x_n + r \mid E) = \frac{\int_{E'} p(x_1 + r, \ldots, x_n + r, y_1, \ldots, y_n) dy}{P(E)} = \frac{\int_{E'} p(y_1, \ldots, y_n) dy}{P(E)}.$$

This establishes the desired equality (E.4) and completes the proof. □

E.4 Proofs for Multivariate Estimators

In this appendix, we provide proofs of the various theorems and lemmas for multivariate mean estimation.

E.4.1 Proof of Theorem 7.4.1

The initial steps in the proof parallel the proof of Theorem 7.3.1, where Lemma E.2.2 is proved using the concentration inequality in Lemma 7.2.8. It then follows that if we choose r such that $R_r^* \geq C_{0.5}\left(\frac{(d+1)\log n}{n}\right)$, we have $R(f_{\widehat{\mu}_{M,r},r}) \geq \frac{R_r^*}{2}$, w.h.p.

Now let $r_2 = 4r\left(\frac{2}{R_r^*}\right)^{\frac{1}{d}}$. By Lemma 7.2.5(i), the desired result will follow if we can show that $R(f_{r_2,r}) \leq \frac{R_r^*}{2}$. By Lemma 7.2.5(iv), we have

$$R(f_{r_2,r}) \leq \frac{R_r^*}{2} \cdot R_{r_2}^* \leq \frac{R_r^*}{2}.$$

To obtain inequality (7.7), note that using Lemma 7.2.5(v), we know that $r = 2\sqrt{d}\sigma_{(2Cd\log n)}$ satisfies the assumption on R_r^*. Plugging into inequality (7.6) then pro-

duces the desired bound.

E.4.2 Proof of Theorem 7.4.3

Let $j' := \min\{j \in \mathcal{J} : r_j \geq r^*\}$. Then

$$\begin{aligned}\mathbb{P}(j_* > j') &= \mathbb{P}\left(\bigcup_{i \in \mathcal{J}: i > j'} \left\{\|\widehat{\mu}_{M,r_i} - \widehat{\mu}_{M,r_{j'}}\|_2 > 8r_i \left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right\}\right) \\ &\leq \mathbb{P}\left(\|\widehat{\mu}_{M,r_{j'}}\|_2 > 4r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right) \\ &\quad + \sum_{i \in \mathcal{J}: i > j'} \mathbb{P}\left(\|\widehat{\mu}_{M,r_i}\|_2 > 4r_i \left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right),\end{aligned}$$

using a union bound and the triangle inequality. We may use Theorem 7.4.1 to bound each individual term, so that the probability of the bad event

$$\begin{aligned}E := &\bigcup_{i \in \mathcal{J}: i > j'} \left\{\|\widehat{\mu}_{M,r_i}\|_2 > 4r_i \left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right\} \\ &\bigcup \left\{\|\widehat{\mu}_{M,r_{j'}}\|_2 > 4r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right\}\end{aligned}$$

is bounded by

$$\begin{aligned}\mathbb{P}(E) &\leq (1 + |\mathcal{J}|) \cdot 2\exp(-c'd\log n) \\ &\leq 2\left(1 + \log_2\left(\frac{2r_{\max}}{r_{\min}}\right)\right)\exp(-c'd\log n).\end{aligned}$$

Finally, note that on the event E^c, we have $j_* \leq j'$ (establishing that j_* is finite), so

$$\|\widehat{\mu}_{M,r_{j_*}} - \widehat{\mu}_{M,r_{j'}}\|_2 \leq 8r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}.$$

Combined with the inequality $\|\widehat{\mu}_{M,r_{j'}}\|_2 < 4r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}$, we conclude that

$$\begin{aligned}\|\widehat{\mu}_{M,r_{j_*}}\|_2 &\leq 8r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d} + 4r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d} \\ &\leq 12r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\end{aligned}$$

$$\leq 24r^* \left(\frac{2n}{C_{0.5}(d+1)\log n} \right)^{1/d},$$

using the fact that $r_{j'} < 2r^*$.

E.4.3 Proof of Lemma 7.4.4

We first prove the upper bound. Note that $R(f_{0,r_{2k}}) = R^*_{r_{2k}} = \frac{2k}{n}$. It suffices to show that this ball contains at least k points, with high probability. By the multiplicative form of the Chernoff bound (Lemma E.8.1 in Appendix E.8),

$$\begin{aligned}\mathbb{P}\left(R_n(f_{0,r_{2k}}) \leq \frac{k}{n} \right) &= \mathbb{P}\left(R_n(f_{0,r_{2k}}) \leq \frac{1}{2} R(f_{0,r_{2k}}) \right) \\ &\leq \exp\left(-n \cdot \frac{k}{n} \cdot \frac{1}{8} \right) = \exp(-k/8).\end{aligned}$$

Therefore, with probability at least $1 - \exp(-k/8)$, a ball of radius r_{2k} contains at least k points, implying that the shortest gap, $\widehat{r}_k \leq r_{2k}$.

We now turn to verifying the lower bound. We will prove that with high probability, no ball of radius $r_{k/2}$ contains at least k points, so that $\widehat{r}_k > r_{k/2}$. By definition, $nR^*_{r_{k/2}} = \frac{k}{2}$. Thus, assuming $k \geq C_{0.5} d \log n$, we may apply Lemma 7.2.8 to conclude that

$$\sup_{f \in \mathcal{H}_{r_{k/2}}} R_n(f) - R(f) \leq \frac{R^*_{r_{k/2}}}{2},$$

with probability at least

$$1 - \exp\left(-\frac{cn}{4} R^*_{r_{k/2}} \right) = 1 - \exp(-ck/8).$$

This implies that

$$\sup_{f \in \mathcal{H}_{r_{k/2}}} R_n(f) \leq \frac{3}{2} \cdot R^*_{r_{k/2}} = \frac{3}{2} \cdot \frac{k}{2n} < \frac{k}{n},$$

which is exactly what we want.

E.4.4 Proof of Theorem 7.4.5

We parallel the proof of Theorem 7.3.3. Note that the guarantees of Lemma 7.4.4 and Lemma E.2.3 continue to hold in d dimensions, except that we have the lower bound $k \geq 2C_{0.5}(d+1)\log n$ instead. We then conclude that $R(f_{\widehat{\mu}_{S,k},r_{2k}}) \geq \frac{k}{2n}$, with probability at least $1 - 2\exp(-c'd\log n)$.

Setting $r' = 4r_{2k}\left(\frac{2n}{k}\right)^{1/d}$, it thus suffices to show that $R(f_{r',r_{2k}}) \leq \frac{k}{2n}$. By Lemma 7.2.5(iv), we have

$$R(f_{r',r_{2k}}) \leq \frac{k}{2n} \cdot R^*_{r'} \leq \frac{k}{2n},$$

as wanted.

E.4.5 Proof of Theorem 7.4.6

We begin with the following result, which can be proved directly via a union bound on Lemma E.2.4:

Lemma E.4.1. *With probability at least* $1 - 4d\exp(-ck^2/n)$*:*

(*i*) *The cuboid* S^∞_k *contains the origin.*

(*ii*) *We have the bound* $\mathrm{Diam}(S^\infty_k) \leq 2\sqrt{d}r_{2k,1}$.

Lemma E.4.1 will be critical in our analysis of the hybrid estimator proposed below. In particular, the estimator will consist of projecting the modal interval/shorth estimator onto the cuboid S^∞_k, and Lemma E.4.1(i) guarantees that the estimation error of the projected estimator will be no larger than the estimation error of the initial estimator without projection. On the other hand, Lemma E.4.1(ii) bounds the error of an estimator based on the k-median alone.

We first derive an upper bound of $\sqrt{d}r_{2\sqrt{n}\log n,1}$. We begin by deriving the following lemma, relating the statistics of marginal distributions to the statistics of the overall distribution:

Lemma E.4.2. *We have that* $r_{\frac{k}{2},1} \leq \frac{C}{\sqrt{d}}r_k$*, for some absolute constant* $C > 0$ *and any* $k \leq n$.

Proof. Consider a uniform distribution on a sphere (or shell) of radius r in $\mathbb{R}^d$. Theorem 3.4.6 in Vershynin [Ver18] provides a concentration result which states that most of the probability of such a distribution lies close to the equator; i.e., the set $\left[-\frac{Cr}{\sqrt{d}}, \frac{Cr}{\sqrt{d}}\right] \times \mathbb{R}^{d-1}$

contains at least half the probability for some absolute constant $C > 0$. Notice that a radially symmetric distribution is simply a weighted sum of uniform distributions on spheres. Thus, given a radially symmetric distribution restricted to the ball of radius r, the set $\left[-\frac{Cr}{\sqrt{d}}, \frac{Cr}{\sqrt{d}}\right] \times \mathbb{R}^{d-1}$ will contain at least half the total probability assigned to the ball.

By our definition of r_k, the ball of radius r_k centered at origin, $\mathbb{B}_{r_k}$, contains $\frac{k}{n}$ probability mass. The above argument implies that the set $\left[-\frac{Cr_k}{\sqrt{d}}, \frac{Cr_k}{\sqrt{d}}\right] \times \mathbb{R}^{d-1}$ will contain at least half the probability of the total probability contained in $\mathbb{B}_{r_k}$. Equivalently, $r_{\frac{k}{2},1} \leq \frac{C}{\sqrt{d}} r_k$. □

Since the output of the hybrid algorithm must lie within the cuboid $S^{\infty}_{\sqrt{n}\log n}$, it is clear that we have the error bound

$$\|\widehat{\mu}_{k_1,k_2}\|_2 \leq \sqrt{n}^{1/d} \cdot \sqrt{d} r_{2\sqrt{n}\log n,1}.$$

To obtain the second upper bound expression, we parallel the proof of Theorem 7.3.5, by splitting into two cases:

Case 1: $r_{4\sqrt{n}\log n} \leq \sqrt{n}^{1/d} r_{8d\log n}$. By Lemma E.4.2, we therefore have

$$r_{2\sqrt{n}\log n,1} \leq \frac{C}{\sqrt{d}} r_{4\sqrt{n}\log n} \leq \frac{C}{\sqrt{d}} \cdot \sqrt{n}^{1/d} r_{8d\log n}.$$

By Lemma E.4.1, w.h.p., the cuboid $S^{\infty}_{\sqrt{n}\log n}$ is entirely contained in the ℓ_2-ball of radius $\sqrt{d} r_{2\sqrt{n}\log n,1}$ around the origin. This ball in turn lies inside the ℓ_2-ball of radius $C\sqrt{n}^{1/d} r_{8d\log n}$ around the origin. Since the output of the hybrid algorithm must also lie within this ball, the desired result follows.

Case 2: $r_{4\sqrt{n}\log n} > \sqrt{n}^{1/d} r_{8d\log n}$. Denoting $r' = \sqrt{n}^{1/d} r_{8d\log n}$, we therefore have the relation $R^*_{r'} < \frac{4\sqrt{n}\log n}{2n}$. In particular, since

$$R(f_{\widehat{\mu}_{S,8d\log n}, r_{8d\log n}}) \geq R_n(f_{\widehat{\mu}_{S,8d\log n}, r_{8d\log n}}) - \frac{1}{2} R^*_{r_{8d\log n}} = \frac{8d\log n}{4n},$$

w.h.p., by Lemma 7.2.8, we have

$$R(f_{r', r_{8d\log n}}) \leq \left(\frac{1}{\sqrt{n}^{1/d}}\right)^d R^*_{r'} < \frac{1}{\sqrt{n}} \cdot \frac{2\log n}{\sqrt{n}} = \frac{8d\log n}{4n} \leq R(f_{\widehat{\mu}_{S,8d\log n}, r_{8d\log n}}).$$

This implies that $\widehat{\mu}_{S,8d\log n}$ is within r' of the origin.

Finally, we need to show that projecting the shorth estimator on the cuboid does not increase its distance from the origin. Note that ℓ_2-projection onto a cuboid is simply a componentwise operation of projection on each interval defining an edge of the cuboid. Furthermore, Lemma E.4.1 guarantees that the origin lies within the cuboid, w.h.p., in which case each interval contains 0. As argued in the proof of Theorem 7.3.5, the distance from the shorth estimator to the origin computed along any dimension will not increase after the projection. Therefore, the ℓ_2-norm of the projected estimator is also upper-bounded by r'.

Hence, if we take $C' = \max\{C, 1\}$, we have the desired bound in both cases. This concludes the proof.

E.4.6 Proof of Theorem 7.6.2

We begin by deriving the proof for the modal interval estimator. Let $s_1 = \frac{r}{2}$, and define s_2 such that $R(f_{s_2,r}) = \frac{1}{3}R(f_{s_1,r})$. Note that

$$R(f_{s_1,r}) \geq R(f_{0,r/2}) \geq \frac{3C_{1/6}d\log n}{n},$$

so $R(f_{s_2,r}) \geq \frac{C_{1/6}d\log n}{n}$. Applying Lemma 7.6.1 with $\bar{r} = s_1$ and $t = \frac{1}{6}$, we conclude that

$$R_n(f_{x,r}) \geq \frac{2}{3}R(f_{x,r}) \geq \frac{2}{3}R(f_{s_1,r}), \tag{E.5}$$

uniformly over $\|x\|_2 \leq s_1$, with probability at least $1 - \frac{2\exp(-cnR(f_{s_1,r})/36)}{1-\exp(-cnR(f_{s_1,r})/36)}$, which is in turn lower-bounded by $1 - 4\exp(-c_1 d\log n)$.

Furthermore, inequality (7.21) implies that

$$R_n(f_{x,r}) \leq R(f_{x,r}) + \frac{1}{3}R(f_{s_2,r}) \leq \frac{4}{3}R(f_{s_2,r}) = \frac{4}{9}R(f_{s_1,r}), \tag{E.6}$$

uniformly over $\|x\|_2 > s_2$, with probability at least $1 - 2\exp(-cnR(f_{s_2,r})/9) \geq 1 - 2\exp(-c_2 d\log n)$. Thus, combining inequalities (E.5) and (E.6), we conclude that

$$\sup_{\|x\|_2 > s_2} R_n(f_{x,r}) < \inf_{\|x\|_2 \leq s_1} R_n(f_{x,r}), \tag{E.7}$$

with probability at least $1 - 6\exp(-c_3 d\log n)$.

Now note that by inequality (E.5), we also have $R_n(f_{0,s_1}) \geq \frac{2}{3}R(f_{0,s_1}) > 0$, implying

that $\{x_1, \ldots, x_n\} \cap B(0, s_1) \neq \emptyset$. In particular,

$$\sup_{x \in \{x_1, \ldots, x_n\}} R_n(f_{x,r}) \geq \inf_{\|x\|_2 \leq s_1} R_n(f_{x,r}).$$

Together with inequality (E.7), we conclude that $\|\widetilde{\mu}_{M,r}\|_2 < s_2$.

Finally, we claim that $s_2 \leq 4r \left(\frac{n}{C_{1/6} d \log n}\right)^{1/d}$. To see this, let $\tilde{s}_2 := 4r \left(\frac{n}{C_{1/6} d \log n}\right)^{1/d}$, and note that by Lemma 7.2.5(iv), we have

$$R(f_{\tilde{s}_2,r}) \leq \frac{C_{1/6} d \log n}{n} \cdot R^*_{\tilde{s}_2} \leq \frac{C_{1/6} d \log n}{n}.$$

Since the last quantity is upper-bounded by $R(f_{s_2,r})$, we conclude that $s_2 \leq \tilde{s}_2$, as claimed.

Turning to the analysis of the computationally efficient shorth estimator, we adapt the argument in the proof of Theorem 7.3.3. By Lemma 7.2.8, if $R^*_{2r_{2k}} \geq \frac{C_{0.5}(d+1) \log n}{n}$, we have

$$\sup_x \sup_{r \leq 2r_{2k}} (R_n(f_{x,r}) - R(f_{x,r})) < \frac{t}{2} R^*_{2r_{2k}},$$

with probability at least $1 - 2\exp(-cnR^*_{2r_{2k}} t^2) \geq 1 - 2\exp\left(-cnt^2 \cdot \frac{2k}{n}\right)$.

We know that $\frac{k}{n} = R_n(f_{\widetilde{\mu}_{S,k}, \widetilde{r}_k}) \leq R_n(f_{\widetilde{\mu}_{S,k}, 2r_{2k}})$. Let s be defined such that $R(f_{s,2r_{2k}}) = \frac{k}{2n}$. By inequality (7.21), we know that

$$\sup_{\|x\|_2 \geq s} |R_n(f_{x,2r_{2k}}) - R(f_{x,2r_{2k}})| \leq \frac{1}{2} R(f_{s,2r_{2k}}),$$

with probability at least $1 - 2\exp(-ck)$, implying that for $\|x\|_2 \geq s$, we have

$$R_n(f_{x,2r_{2k}}) \leq R(f_{x,2r_{2k}}) + \frac{1}{2} R(f_{s,2r_{2k}}) \leq \frac{3}{2} R(f_{s,2r_{2k}}) = \frac{3k}{4n}.$$

Since this is strictly smaller than $R_n(f_{\widetilde{\mu}_{S,k}, 2r_{2k}})$, we conclude that $\|\widetilde{\mu}_{S,k}\|_2 \leq s$, w.h.p. , which also implies that $R(f_{\widetilde{\mu}_{S,k}, 2r_{2k}}) \geq \frac{k}{2n}$.

Finally, let $r' = 4r_{2k} \left(\frac{2n}{k}\right)^{1/d}$. By Lemma 7.2.5(iv), we have

$$R(f_{r', 2r_{2k}}) < \frac{k}{2n} \cdot R^*_{r'} \leq \frac{k}{2n} < R(f_{\widetilde{\mu}_{S,k}, 2r_{2k}}).$$

Applying Lemma 7.2.5(i), we conclude that $\|\widetilde{\mu}_{S,k}\|_2 \leq r'$.

E.5 Proofs for Expected Error Bounds

In this appendix, we prove the results stated in Section 7.5.

E.5.1 Proof of Proposition 7.5.2

The proof sketch is that we will show that with finite probability, no interval contains more than one low-variance point, and all the high-variance points lie far from origin. Conditioned on this event, the modal interval estimator incurs a high error.

Let $E = A \cap B$, where we define the events

$$\begin{aligned} A &= \{R_n(f_{x,1}) \leq 1, \quad \forall x : |x| \leq 3C \log n\}, \\ B &= \{X_i \notin [-4C \log n, 4C \log n], \quad \forall i > C \log n\}. \end{aligned}$$

Hence, on the event E, no interval overlapping with $[-3C \log n, 3C \log n]$ contains two low-variance points or a single high-variance point. Then $\mathbb{P}(E)$ is lower-bounded by

$$\begin{aligned} \mathbb{P}(E) &\geq \left(\prod_{i=1}^{C \log n} \mathbb{P}\{X_i \in [3i-3, 3i-2]\} \right) \left(\prod_{i > C \log n} \mathbb{P}\{X_i \notin [-4C \log n, 4C \log n]\} \right) \\ &= \left(\prod_{i=1}^{C \log n} \frac{1}{6i} \right) \left(\prod_{i > C \log n} (1 - n^{-\alpha} - h_n(8C \log n - 2)) \right) \\ &\geq \frac{1}{6^{C \log n} \Gamma(3C \log n)} e^{-cn^{1-\alpha}} \\ &\geq \exp\left(-cn^{1-\alpha} - O\left(\log^2 n \right) \right), \end{aligned}$$

assuming $h_n \log n \ll n^{-\alpha}$, which happens for $q_n = \Omega(n)$.

However, conditioned on E, the points $\{X_i\}_{i > C \log n}$ are i.i.d. with the following distribution:

$$p_{i,E}(x) = \begin{cases} 0, & |x| \leq 4C \log n, \\ \frac{h_n}{(1 - n^{-\alpha} - h_n(8C \log n - 2))}, & 4C \log n < |x| \leq q_n, \\ 0, & \text{otherwise}. \end{cases}$$

We can now apply the symmetry arguments of Lemma E.3.4. Note that no interval lying inside $[-3C \log n, 3C \log n]$ can contain more than one point. Thus unless a tie occurs, the mode will be located outside the interval $[-3C \log n, 3C \log n]$, and hence a distance

of $\Theta(q_n)$ away from the mean in expectation. Even if we were to break ties randomly, a large error would occur with probability at least $\frac{1}{n}$, since at most n ties can occur. Thus,

$$\mathbb{E}[|\widehat{\mu}_{M,1}||E] \geq \mathbb{P}(E)\,\mathbb{E}[|\widehat{\mu}_{M,1}||E] \geq \exp(-cn^{1-\alpha})\Theta(q_n).$$

The bounds in high probability follow from Lemma E.2.2, by noting that $nR_r^* = \Omega\left(n^{-\alpha}\right) = \Omega(\log n)$. Moreover, the density drops by at least half at $x > 1$.

E.5.2 Proof of Theorem 7.5.3

We begin by proving (i). By Theorem 7.4.1, we have

$$\|\widehat{\mu}_{M,r}\|_2 = O\left(r\left(\frac{c}{R_r^*}\right)^{1/d}\right),$$

with probability at least most $1 - O\exp\left(-c'nR_r^*\right)$. In the worst case, the modal interval estimator returns the point which is furthest from the origin, which has expected value bounded as

$$\mathbb{E}\left[\max_i \|X_i\|_2\right] \leq \mathbb{E}\left[\sqrt{\sum_{i=1}^n \|X_i\|_2^2}\right] \leq \sqrt{\sum_{i=1}^n \mathbb{E}[\|X_i\|_2^2]} \leq \sqrt{n \cdot d\sigma_{(n)}^2}.$$

Using the assumption that $\sigma_n \leq r\exp(CnR_r^*)$, for some constant $C > 0$, we then have

$$\begin{aligned}
\mathbb{E}\,\|\widehat{\mu}_{M,r}\|_2 &\leq O\left(r\left(\frac{c}{R_r^*}\right)^{1/d}\right) + O\exp\left(-c'nR_r^*\right))\sqrt{nd}\sigma_{(n)} \\
&\leq O\left(r\left(\frac{c}{R_r^*}\right)^{1/d}\right) + O\left(\exp\left(-c'nR_r^*\right)r\sqrt{nd}\exp(CnR_r^*)\right) \\
&= O\left(r\left(\frac{c}{R_r^*}\right)^{1/d}\right),
\end{aligned}$$

where in the last inequality, we use the facts that

$$\exp(-c'nR_r^*)\sqrt{nd} = O(\exp(-c''nR_r^*))$$

and $nR_r^* = \Omega\left(d\log n\right)$, and choose $C < c''$.

Turning to (ii), we first prove the following concentration result, which may be viewed as a refinement of Lemma 7.2.8 that is suitable for our settings. For example,

note that if $R^*_{\mathcal{J}} = O\left(\frac{1}{n}\right)$, the derivations from Lemma 7.2.8 would not be meaningful since $R^*_{\mathcal{J}} = o\left(\frac{\log n}{n}\right)$. On the other hand, if $KR^*_{\mathcal{J}} = \Theta\left(\frac{\log n}{n}\right)$, Lemma E.5.1 gives a vanishing upper bound.

Lemma E.5.1. *Let $\mathcal{J}$ be a set of intervals and define $R^*_{\mathcal{J}} := \sup_{f\in\mathcal{J}} R(f)$. If $R^*_{\mathcal{J}} \le \frac{1}{3}$, then for any $K \ge 8$, we have*

$$\mathbb{P}\left\{\sup_{f\in\mathcal{J}} R_n(f) \ge KR^*_{\mathcal{J}}\right\} \le \frac{2}{R^*_{\mathcal{J}}} \exp\left(-cnR^*_{\mathcal{J}} K \log K\right).$$

Proof. For a given $f \in \mathcal{J}$, the desired bound follows from Chernoff's inequality. We want to upper-bound the probability that any one interval in $\mathcal{J}$ has too many points. In general, the set $\mathcal{J}$ may be infinite, so a direct union bound is not feasible. We thus create a new finite set of intervals $\mathcal{F}$, not necessarily a subset of $\mathcal{J}$, satisfying the following properties:

1. For each $f \in \mathcal{F}$, we have $\frac{R^*_{\mathcal{J}}}{2} \le R(f) \le R^*_{\mathcal{J}}$.
2. $|\mathcal{F}| \le \frac{2}{R^*_{\mathcal{J}}}$.
3. $\mathcal{F}$ covers $\mathcal{J}$ in the sense that $\forall f \in \mathcal{J}, \exists f_1, f_2 \in \mathcal{F} : f(x) \le f_1(x) + f_2(x)$.

It follows that if any interval in $\mathcal{J}$ contains at least k points, then at least one interval in $\mathcal{F}$ contains at least $\frac{k}{2}$ points. We construct $\mathcal{F}$ of cardinality $|\mathcal{F}| = \lceil \frac{1}{R^*_{\mathcal{J}}} \rceil \le \frac{2}{R^*_{\mathcal{J}}}$, as follows: To create the first interval ($i = 1$), define $x_1 \in \mathbb{R}$ such that $R(\mathbb{1}_{(-\infty,x_1]}) = \frac{1}{|\mathcal{F}|}$. (Such an x_1 exists because $\overline{P}$ is assumed to have a density.) Then iteratively, for each $i \ge 1$, define x_i such that $R(\mathbb{1}_{(x_{i-1},x_i]}) = \frac{1}{|\mathcal{F}|}$. For the final interval, add $\mathbb{1}_{[x_{i-1},\infty)}$ to $\mathcal{F}$ and terminate the construction. Note that for each $f \in \mathcal{F}$, we have $R(f) = \frac{1}{\lceil 1/R^*_{\mathcal{J}} \rceil}$, which clearly lies in $\left[\frac{R^*_{\mathcal{J}}}{2}, R^*_{\mathcal{J}}\right]$ under the assumptions.

We are now ready to apply the union bound on $\mathcal{F}$ using Lemma E.8.1(ii):

$$\begin{aligned}
\mathbb{P}\left\{\sup_{f\in\mathcal{J}} R_n(f) \ge KR^*_{\mathcal{J}}\right\} &\le \mathbb{P}\left\{\sup_{f\in\mathcal{F}} R_n(f) \ge \frac{KR^*_{\mathcal{J}}}{2}\right\} \\
&\le |\mathcal{F}|\,\mathbb{P}\left\{R_n(f) \ge \frac{KR^*_{\mathcal{J}}}{2} \text{ for a fixed } f \text{ with } R(f) \le R^*_{\mathcal{J}}\right\} \\
&\le \frac{2}{R^*_{\mathcal{J}}} \exp\left(-cnR^*_{\mathcal{J}} K \log K\right).
\end{aligned}$$

□

For an $s \geq 0$, let $\mathcal{J}_s = \{f_{x,r} : \|x\|_2 \geq s\}$, i.e., the set of intervals which incur large error. By assumption, the support of at least CnR_r^* points is contained in $[-r, r]$, implying that $R_n(f_{0,r}) \geq CR_r^*$, a.s. If $\|\widehat{\mu}_{M,r}\|_2 \geq s$, then $\sup_{f\in\mathcal{J}_s} R_n(f) \geq CR_r^*$. However as s increases, the quantity $R_{\mathcal{J}_s}^* := \sup_{f\in\mathcal{J}_s} R(f) = R(f_{s,r})$ decreases. We can then use Lemma E.5.1 to control this probability of error.

For $s \geq \frac{Kr}{CR_r^*}$, it follows from Lemma 7.2.5(iv) that $R_{\mathcal{J}_s}^* = R(f_{s,r}) \leq \frac{CR_r^*}{K}$. Taking $K \geq C'$, we then have

$$\begin{aligned}
\mathbb{P}\{|\widehat{\mu}_{M,r}| \geq s\} &\leq \mathbb{P}\left(\sup_{f\in\mathcal{J}_s} R_n(f) \geq CR_r^*\right) \\
&= \mathbb{P}\left(\sup_{f\in\mathcal{J}_s} R_n(f) \geq \frac{CR_r^*}{R_{\mathcal{J}_s}^*} R_{\mathcal{J}_s}^*\right) \\
&\leq \frac{2}{R_{\mathcal{J}_s}^*} \exp\left(-cnR_{\mathcal{J}_s}^* \frac{CR_r^*}{R_{\mathcal{J}_s}^*} \log\left(\frac{CR_r^*}{R_{\mathcal{J}_s}^*}\right)\right) \\
&= \frac{2}{R_r^*} \exp\left(-cCnR_r^* \log\left(\frac{CR_r^*}{R_{\mathcal{J}_s}^*}\right) + \log\left(\frac{R_r^*}{R_{\mathcal{J}_s}^*}\right)\right) \\
&\leq \frac{2}{R_r^*} \exp\left(-c'nR_r^* \log\left(\frac{R_r^*}{R_{\mathcal{J}_s}^*}\right)\right),
\end{aligned}$$

where we have applied Lemma E.5.1 in the second inequality. Thus,

$$\begin{aligned}
\mathbb{E}\,|\widehat{\mu}_{M,r}| &\leq \frac{4r}{CR_r^*} + \int_{\frac{4r}{CR_r^*}}^{\infty} \mathbb{P}\{|\widehat{\mu}_{M,r}| \geq s\} ds \\
&\leq O\left(\frac{r}{R_r^*}\right) + \frac{2}{R_r^*} \int_{\frac{4r}{CR_r^*}}^{\infty} \exp\left(-c'nR_r^* \log\left(\frac{R_r^*}{R_{\mathcal{J}_s}^*}\right)\right) ds \\
&\leq O\left(\frac{r}{R_r^*}\right) + \frac{2}{R_r^*} \int_{\frac{4r}{CR_r^*}}^{\infty} \exp\left(-c'nR_r^* \log\left(\frac{sR_r^*}{r}\right)\right) ds \\
&\leq O\left(\frac{r}{R_r^*}\right) + \frac{r}{R_r^*}\frac{2}{R_r^*} \int_{4/C}^{\infty} \exp\left(-c'nR_r^* \log s_1\right) ds_1 \\
&= O\left(\frac{r}{R_r^*}\right) + \frac{r}{R_r^*}\frac{2}{R_r^*} \int_{4/C}^{\infty} s_1^{-c'nR_r^*} ds_1 \\
&\leq O\left(\frac{r}{R_r^*}\right) + \frac{r}{R_r^*}\frac{2}{R_r^*} \cdot \frac{1}{c'nR_r^* - 1} (4/C)^{1-c'nR_r^*} \\
&= O\left(\frac{r}{R_r^*}\right),
\end{aligned}$$

where the third inequality uses the fact that $R_{\mathcal{J}_s}^* = R(f_{s,r}) \leq \frac{r}{s}$, and the last equality follows from an appropriately small choice of C.

E.5.3 Proof of Theorem 7.5.5

Note that for any $s > 0$, Markov's inequality gives

$$\min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathcal{P}(\sigma_1, \sigma_2, p)} \mathbb{E}[\|\widehat{\mu} - \mu\|_2] \geq \min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathcal{P}(\sigma_1, \sigma_2, p)} s \cdot \mathbb{P}(\|\widehat{\mu} - \mu\|_2 \geq s).$$

Clearly, the right-hand expression is lower-bounded by the maximum over any specific collection of distributions in the class $\mathcal{P}(\sigma_1, \sigma_2, p)$. In particular, let $\mathcal{P}_m^\mu$ be the collection of multivariate distributions where each distribution is either $N(\mu, \sigma_1^2 I)$ or $N(\mu, \sigma_2^2 I)$, with m distributions of the latter type. We then have

$$\begin{aligned}
\min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathcal{P}(\sigma_1, \sigma_2, p)} \mathbb{P}(\|\widehat{\mu} - \mu\|_2 \geq s) &\geq \min_{\widehat{\mu}} \max_{\mu} \max_{np \leq m \leq 2np} \mathbb{P}(\|\widehat{\mu} - \mu\|_2 \geq s \mid \{P_i\} = \mathcal{P}_m^\mu) \\
&\geq \min_{\widehat{\mu}} \max_{\mu} \sum_{np \leq m \leq 2np} \mathbb{P}\left(\|\widehat{\mu} - \mu\|_2 \geq s \mid \{P_i\} = \mathcal{P}_m^\mu\right) p_m,
\end{aligned}$$

where $\{p_m\}$ is any allocation of probabilities defined over $\{\mathcal{P}_{np}^\mu, \ldots, \mathcal{P}_{2np}^\mu\}$, such that $0 \leq p_m \leq 1$ for all m and $\sum_m p_m \leq 1$. In particular, consider the probability mass function $\{q_m\}_{m=1}^n$ over $\{\mathcal{P}_1^\mu, \ldots, \mathcal{P}_n^\mu\}$ corresponding to the Binomial(n, p) distribution, and define $p_m = q_m$ for all $np \leq m \leq 2np$.

Now let $\mathbb{P}_{\text{Bin}}^\mu$ denote the probability distribution when the P_i's are chosen i.i.d. in the following manner: with probability $p' := 1.5p$, the distribution is $N(\mu, \sigma_2^2 I)$, and with probability $1 - 1.5p$, the distribution is $N(\mu, \sigma_1^2 I)$. Then

$$\mathop{\mathbb{P}}_{\text{Bin}}^{\mu}(\|\widehat{\mu} - \mu\|_2 \geq s) = \sum_{m=1}^{n} \mathbb{P}\left(\|\widehat{\mu} - \mu\|_2 \geq s \mid \{P_i\} = \mathcal{P}_m^\mu\right) q_m.$$

Hence,

$$\begin{aligned}
\left| \sum_{np \leq m \leq 2np} \mathbb{P}\left(\|\widehat{\mu} - \mu\|_2 \geq s \mid \{P_i\} = \mathcal{P}_m^\mu\right) p_m - \mathop{\mathbb{P}}_{\text{Bin}}^{\mu}(\|\widehat{\mu} - \mu\|_2 \geq s) \right| &\leq \sum_{m < np} q_m + \sum_{m > 2np} q_m \\
&\leq 2\exp(-cnp) \\
&\leq 2\exp(-c' \log n),
\end{aligned}$$

where second inequality follows from the multiplicative Chernoff bound (Lemma E.8.1) and the last inequality follows by the assumption $p = \Omega\left(\frac{\log n}{n}\right)$. Combining the inequali-

ties, we conclude that

$$\min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathcal{P}(s_1, s_2, p)} \mathbb{E}[\|\widehat{\mu} - \mu\|_2] \geq s \left(\min_{\widehat{\mu}} \max_{\mu} \overset{\mu}{\underset{\text{Bin}}{\mathbb{P}}} (\|\widehat{\mu} - \mu\|_2 \geq s) - 2 \exp(-c' \log n) \right).$$

Thus, it suffices to find s such that the expression $\min_{\widehat{\mu}} \max_{\mu} \mathbb{P}^{\mu}_{\text{Bin}}(\|\widehat{\mu} - \mu\|_2 \geq s)$ can be lower-bounded by a constant.

For part (i), using standard techniques [Tsy09; Wai19], we may obtain such a lower bound via Fano's inequality. In particular, if we can construct a set $\{\mu_1, \ldots, \mu_M\} \subseteq \mathbb{R}^d$ such that $\|\mu_j - \mu_k\|_2 \geq 2s$ and $KL(\mathbb{P}^{\mu_j}_{\text{Bin}}, \mathbb{P}^{\mu_k}_{\text{Bin}}) \leq \alpha$ for all $j \neq k$, then

$$\min_{\widehat{\mu}} \max_{\mu} \overset{\mu}{\underset{\text{Bin}}{\mathbb{P}}} (\|\widehat{\mu} - \mu\|_2 \geq s) \geq \left(1 - \frac{\alpha + \log 2}{\log M}\right).$$

Note that by tensorization and convexity of the KL divergence, we have the upper bound

$$KL(\overset{\mu_j}{\underset{\text{Bin}}{\mathbb{P}}}, \overset{\mu_k}{\underset{\text{Bin}}{\mathbb{P}}}) \leq n(1-p')KL\left(N(\mu_j, \sigma_1^2 I), N(\mu_k, \sigma_1^2 I)\right) + np'KL\left(N(\mu_j, \sigma_2^2 I), N(\mu_k, \sigma_2^2 I)\right), \tag{E.8}$$

where the KL divergences in the right-hand expression are computed with respect to single samples from the respective multivariate normal distributions. Furthermore, the right-hand side of inequality (E.8) is easily calculated to be

$$n(1-p') \cdot \frac{\|\mu_j - \mu_k\|_2^2}{2\sigma_1^2} + np' \cdot \frac{\|\mu_j - \mu_k\|_2^2}{2\sigma_2^2} = n\|\mu_j - \mu_k\|_2^2 \left(\frac{1-p'}{2\sigma_1^2} + \frac{p'}{2\sigma_2^2}\right).$$

In particular, suppose $\{\mu_1, \ldots, \mu_M\}$ is a $2s$-packing of the ball of radius $4s$ in ℓ_2-norm, with $s = C\sqrt{d} \min\left\{\frac{\sigma_1}{\sqrt{n}}, \frac{\sigma_2}{\sqrt{np'}}\right\}$. Then $\log M \geq cd$ and

$$KL(\overset{\mu_j}{\underset{\text{Bin}}{\mathbb{P}}}, \overset{\mu_k}{\underset{\text{Bin}}{\mathbb{P}}}) \leq 4ns^2 \left(\frac{1-p'}{2\sigma_1^2} + \frac{p'}{2\sigma_2^2}\right) \leq 4C^2 d := \alpha.$$

For a sufficiently small choice of C, we conclude that $\min_{\widehat{\mu}} \max_{\mu} \mathbb{P}^{\mu}_{\text{Bin}}(\|\widehat{\mu} - \mu\|_2 \geq s) \geq \frac{1}{2}$. Hence, we arrive at the desired bound (7.10).

We now turn to part (ii). We derive the tighter lower bound (7.12) for the case $d = 1$ by evaluating $KL(\mathbb{P}^{\mu_1}_{\text{Bin}}, \mathbb{P}^{\mu_2}_{\text{Bin}})$ more directly. By Theorem 2.2 in Tsybakov [Tsy09], we know that if we have a pair $\mu_1, \mu_2 \in \mathbb{R}^d$ such that $\|\mu_1 - \mu_2\|_2 \geq 2s$ and

$$KL(\overset{\mu_1}{\underset{\text{Bin}}{\mathbb{P}}}, \overset{\mu_2}{\underset{\text{Bin}}{\mathbb{P}}}) \leq \alpha < \infty, \tag{E.9}$$

then

$$\min_{\widehat{\mu}} \max_{\mu} \overset{\mu}{\underset{\text{Bin}}{\mathbb{P}}} (\|\widehat{\mu} - \mu\|_2 \geq s) \geq \max \left\{ \frac{\exp(-\alpha)}{4}, \frac{1 - \sqrt{\alpha/2}}{2} \right\}.$$

Again, since the KL divergence tensorizes, it suffices to compute the KL divergence between a single sample from the distributions $\mathbb{P}_{\text{Bin}}^{\mu_1}$ and $\mathbb{P}_{\text{Bin}}^{\mu_2}$, which we denote by $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively.

We provide the details of the calculation for general d, with the assumption (7.11) replaced by the condition

$$\left(\frac{\sigma_1}{\sigma_2}\right)^d = O\left(\frac{1}{np^2}\right). \tag{E.10}$$

By a straightforward calculation, we have

$$\begin{aligned}
\log\left(\frac{d\,\mathbb{P}_1(x)}{d\,\mathbb{P}_2(x)}\right) &= \log\left(\frac{(1-p')\frac{1}{(\sqrt{2\pi}\sigma_1)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2}\right) + p'\frac{1}{(\sqrt{2\pi}\sigma_2)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}\right)}{(1-p')\frac{1}{(\sqrt{2\pi}\sigma_1)^d}\exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_1^2}\right) + p'\frac{1}{(\sqrt{2\pi}\sigma_2)^d}\exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_2^2}\right)}\right) \\
&= \left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2} + \frac{\|x-\mu_2\|_2^2}{2\sigma_1^2}\right) + \log\left(\frac{1+y}{1+z}\right),
\end{aligned}$$

where

$$\begin{aligned}
y &:= \frac{p'}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d \exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2} + \frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right), \\
z &:= \frac{p'}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d \exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{2\sigma_1^2}\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
KL(\mathbb{P}_1, \mathbb{P}_2) &= \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}\left[\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2} + \frac{\|x-\mu_2\|_2^2}{2\sigma_1^2}\right] + \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}\left[\log\left(\frac{1+y}{1+z}\right)\right] \\
&\leq \frac{\|\mu_1-\mu_2\|^2}{2\sigma_1^2} + \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[y] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z] + \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z^2],
\end{aligned}$$

using the fact that

$$\log\left(\frac{1+y}{1+z}\right) = \log\left(1 + \frac{y-z}{1+z}\right) \leq \frac{y-z}{1+z} \leq y - z + z^2,$$

since $y, z > 0$. We now write

$$\mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[y]$$

$$= \frac{p'}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d \left((1-p')\int \exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}+\frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)\frac{1}{(\sqrt{2\pi}\sigma_1)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)dx \right.$$
$$\left. + p'\int \exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}+\frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)\frac{1}{(\sqrt{2\pi}\sigma_2)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}\right)dx\right)$$
$$:= A_y + B_y,$$

and

$$\mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z]$$
$$= \frac{p'}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d \left((1-p')\int \exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_2^2}+\frac{\|x-\mu_2\|_2^2}{2\sigma_1^2}\right)\frac{1}{(\sqrt{2\pi}\sigma_1)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)dx \right.$$
$$\left. + p'\int \exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_2^2}+\frac{\|x-\mu_2\|_2^2}{2\sigma_1^2}\right)\frac{1}{(\sqrt{2\pi}\sigma_2)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}\right)dx\right)$$
$$:= A_z + B_z.$$

Now, we may calculate

$$A_y = p'\left(\frac{1}{\sqrt{2\pi}\sigma_2}\right)^d \int \exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}\right)dx = p',$$

and

$$B_y = \frac{(p')^2}{1-p'}\left(\frac{\sigma_1}{\sqrt{2\pi}\sigma_2^2}\right)^d \int \exp\left(\frac{-\|x-\mu_1\|_2^2}{\sigma_2^2}+\frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)dx$$
$$= \frac{(p')^2}{1-p'}\left(\frac{\sigma_1}{\sqrt{2\pi}\sigma_2^2}\right)^d \left(\frac{\pi}{\frac{1}{\sigma_2^2}-\frac{1}{2\sigma_1^2}}\right)^{d/2} \leq \frac{(p')^2}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d,$$

using the fact that $\frac{1}{2\sigma_1^2} \leq \frac{1}{2\sigma_2^2}$. Under the assumption (E.10), we get that $B_y = O(1/n)$.

For ease of calculation, we now set

$$\mu_1^\top = (\mu, 0, \ldots, 0),$$
$$\mu_2^\top = (-\mu, 0, \ldots, 0). \tag{E.11}$$

Using the formula

$$\int \exp\left(-x^\top Ax + b^\top x + c\right) dx = \sqrt{\frac{\pi^d}{\det(A)}} \exp\left(\frac{1}{4} b^\top A^{-1} b + c\right), \tag{E.12}$$

we have

$$\begin{aligned} A_z &= p' \left(\frac{1}{\sqrt{2\pi}\sigma_2}\right)^d \int \exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{2\sigma_1^2} - \frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right) dx \\ &= p' \exp\left(\frac{\sigma_2^2}{2}\left(\left\|\frac{\mu_2}{\sigma_2^2} - \frac{\mu_2}{\sigma_1^2} + \frac{\mu_1}{\sigma_1^2}\right\|_2^2 - \frac{\mu_2^\top \mu_2}{2\sigma_2^2} + \frac{\mu_2^\top \mu_2}{2\sigma_1^2} - \frac{\mu_1^\top \mu_1}{2\sigma_1^2}\right)\right) \\ &= p' \exp\left(-2\mu^2\left(\frac{1}{\sigma_1^2} - \frac{\sigma_2^2}{\sigma_1^4}\right)\right). \end{aligned}$$

In particular, using the fact that $\exp(-x) \geq 1 - x$ for $x \geq 0$, we have

$$A_y - A_z = p' - A_z \leq p' \cdot 2\mu^2 \left(\frac{1}{\sigma_1^2} - \frac{\sigma_2^2}{\sigma_1^4}\right) \leq \frac{2\mu^2}{\sigma_1^2}.$$

We can use the simple fact that $B_z \geq 0$ to ensure that $B_y - B_z \leq B_y = O1/n)$.

Combining the inequalities, we conclude that

$$\mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[y] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z] = O\left(\frac{\mu^2}{\sigma_1^2}\right) + O\left(\frac{1}{n}\right).$$

Finally, we compute

$$\begin{aligned} \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z^2] = \left(\frac{p'}{1-p'}\right)^2 \left(\frac{\sigma_1}{\sigma_2}\right)^{2d} \Bigg(& (1-p') \int \exp\left(\frac{-\|x-\mu_2\|_2^2}{\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{\sigma_1^2}\right) \\ & \cdot \frac{1}{(\sqrt{2\pi}\sigma_1)^d} \exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2}\right) dx \\ & + p' \int \exp\left(\frac{-\|x-\mu_2\|_2^2}{\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{\sigma_1^2}\right) \frac{1}{(\sqrt{2\pi}\sigma_2)^d} \exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}\right) dx \Bigg) \\ := A'_z + B'_z. \end{aligned}$$

Again using the designation (E.11) and the formula (E.12), we have

$$A'_z = \frac{(p')^2}{1-p'} \left(\frac{\sigma_1}{\sqrt{2\pi}\sigma_2^2}\right)^d \int \exp\left(\frac{-\|x-\mu_2\|_2^2}{\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{\sigma_1^2} - \frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right) dx$$

$$
\begin{aligned}
&= \frac{(p')^2}{1-p'} \left(\frac{\sigma_1}{\sqrt{2}\sigma_2^2 \sqrt{\frac{1}{\sigma_2^2} - \frac{1}{2\sigma_1^2}}} \right)^d \exp\left(\frac{1}{4\left(\frac{1}{\sigma_2^2} - \frac{1}{2\sigma_1^2}\right)} \left\| \frac{2\mu_2}{\sigma_2^2} - \frac{2\mu_2}{\sigma_1^2} + \frac{\mu_1}{\sigma_1^2} \right\|_2^2 \right) \\
&\quad \times \exp\left(-\frac{\mu_2^\top \mu_2}{\sigma_2^2} + \frac{\mu_2^\top \mu_2}{\sigma_1^2} - \frac{\mu_1^\top \mu_1}{2\sigma_1^2} \right) \\
&= \frac{(p')^2}{1-p'} \left(\frac{\sigma_1}{\sqrt{2}\sigma_2^2 \sqrt{\frac{1}{\sigma_2^2} - \frac{1}{2\sigma_1^2}}} \right)^d \exp\left(\frac{(-2\mu/\sigma_2^2 + 3\mu/\sigma_1^2)^2}{4\left(\frac{1}{\sigma_2^2} - \frac{1}{2\sigma_1^2}\right)} - \frac{\mu^2}{\sigma_2^2} + \frac{\mu^2}{2\sigma_1^2} \right) \\
&\leq \frac{(p')^2}{1-p'} \left(\frac{\sigma_1}{\sigma_2} \right)^d \exp\left(\frac{(-2\mu/\sigma_2^2 + 3\mu/\sigma_1^2)^2}{4\left(\frac{1}{\sigma_2^2} - \frac{1}{2\sigma_1^2}\right)} - \frac{\mu^2}{\sigma_2^2} + \frac{\mu^2}{2\sigma_1^2} \right),
\end{aligned}
$$

and

$$
\begin{aligned}
B_z' &= \frac{(p')^3}{1-p'} \left(\frac{\sigma_1^2}{\sqrt{2\pi}\sigma_2^3} \right)^d \int \exp\left(\frac{-\|x-\mu_2\|_2^2}{\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{\sigma_1^2} - \frac{\|x-\mu_1\|_2^2}{2\sigma_2^2} \right) dx \\
&= \frac{(p')^3}{1-p'} \left(\frac{\sigma_1^2}{\sqrt{2}\sigma_2^3 \sqrt{\frac{3}{2\sigma_2^2} - \frac{1}{\sigma_1^2}}} \right)^d \exp\left(\frac{1}{4\left(\frac{3}{2\sigma_2^2} - \frac{1}{\sigma_1^2}\right)} \left\| \frac{2\mu_2}{\sigma_2^2} - \frac{2\mu_2}{\sigma_1^2} + \frac{\mu_1}{\sigma_2^2} \right\|_2^2 \right) \\
&\qquad \times \exp\left(-\frac{\mu_2^\top \mu_2}{\sigma_2^2} + \frac{\mu_2^\top \mu_2}{\sigma_1^2} - \frac{\mu_1^\top \mu_1}{2\sigma_2^2} \right) \\
&= \frac{(p')^3}{1-p'} \left(\frac{\sigma_1^2}{\sqrt{2}\sigma_2^3 \sqrt{\frac{3}{2\sigma_2^2} - \frac{1}{\sigma_1^2}}} \right)^d \exp\left(\frac{(-\mu/\sigma_2^2 + 2\mu/\sigma_1^2)^2}{4\left(\frac{3}{2\sigma_2^2} - \frac{1}{\sigma_1^2}\right)} - \frac{3\mu^2}{2\sigma_2^2} + \frac{\mu^2}{\sigma_1^2} \right) \\
&\leq \frac{(p')^3}{1-p'} \left(\frac{\sigma_1}{\sigma_2} \right)^{2d} \exp\left(\frac{(-\mu/\sigma_2^2 + 2\mu/\sigma_1^2)^2}{4\left(\frac{3}{2\sigma_2^2} - \frac{1}{\sigma_1^2}\right)} - \frac{3\mu^2}{2\sigma_2^2} + \frac{\mu^2}{\sigma_1^2} \right).
\end{aligned}
$$

Considering the exponential terms in the expressions for A_z' and B_z', note that for A_z', we have

$$
\frac{(-2\mu/\sigma_2^2 + 3\mu/\sigma_1^2)^2}{4\left(\frac{1}{\sigma_2^2} - \frac{1}{2\sigma_1^2}\right)} - \frac{\mu^2}{\sigma_2^2} = \frac{\mu^2}{\sigma_2^2} \left(\frac{\left(2 - \frac{3\sigma_2^2}{\sigma_1^2}\right)^2}{4\left(1 - \frac{\sigma_2^2}{2\sigma_1^2}\right)} - 1 \right) < 0,
$$

assuming $\sigma_2 \leq \sigma_1$, whereas for B_z', we have

$$
\frac{(-\mu/\sigma_2^2 + 2\mu/\sigma_1^2)^2}{4\left(\frac{3}{2\sigma_2^2} - \frac{1}{\sigma_1^2}\right)} - \frac{3\mu^2}{2\sigma_2^2} = \frac{\mu^2}{\sigma_2^2} \left(\frac{\left(1 - \frac{2\sigma_2^2}{\sigma_1^2}\right)^2}{4\left(\frac{3}{2} - \frac{\sigma_2^2}{\sigma_1^2}\right)} - \frac{3}{2} \right) < 0,
$$

using the fact that $\sigma_2 \leq \sigma_1$. Thus, using the assumption (E.10), we obtain

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z^2] = A'_z + B'_z &= O\left(\frac{1}{n}\right)\exp\left(\frac{\mu^2}{2\sigma_1^2}\right) + O\left(\frac{1}{n^2 p}\right)\exp\left(\frac{\mu^2}{\sigma_1^2}\right) \\ &= O\left(\frac{1}{n}\right)\exp\left(\frac{\mu^2}{\sigma_1^2}\right).\end{aligned}$$

Finally, we take $\mu = \frac{\sigma_1}{\sqrt{n}}$ to obtain the desired bound (E.9). This completes the proof.

E.5.4 Proof of Theorem 7.5.7

By a similar argument used to derive the bound in Theorem 7.5.3, the following expected error bound may be derived from the high-probability bound in Theorem 7.4.6 for the hybrid estimator:

$$\mathbb{E}\,\|\widehat{\mu}_{k_1,k_2}\|_2 \leq \min\left\{\sqrt{d}r_{2k_1,1}, \sqrt{n}^{1/d} r_{k_2}\right\}. \tag{E.13}$$

In what follows, we will bound these expressions to obtain the desired results.

As shown in the proof of Lemma 7.2.5(v), a ball of radius $C\sigma_2\sqrt{d}$ around the origin will contain at least $\frac{1}{2}$ of the mass of np distributions. Thus, if $np \geq 2k_2$, we will have $r_{k_2} \leq C\sigma_2\sqrt{d}$.

We now claim that $r_{2k_1,1} \leq \frac{C\sigma_1 \log n}{\sqrt{n}} := r'$, which we will show by integrating the marginal densities on the interval $[-r', r']$. Note that $\nu_i \leq \sigma_1$ for all i. We consider two cases: if $\nu_i \geq r'$, then $q_i(r') \geq \frac{c}{\nu_i} \geq \frac{c}{\sigma_1}$, using inequality (7.13), so $\int_{[-r',r']} q_i(x)dx \geq \frac{2cr'}{\sigma_1} \geq \frac{2\log n}{\sqrt{n}}$ for large enough C. If $\nu_i < r'$, then $\int_{[-\nu_i,\nu_i]} q(x)dx \geq c' \geq \frac{2\log n}{\sqrt{n}}$, as well. Thus,

$$\sum_{i=1}^{n}\int_{[-r',r']} q_i(x)dx \geq \sum_{i=1}^{n}\frac{2\log n}{\sqrt{n}} \geq 2\sqrt{n}\log n = 2k_1. \tag{E.14}$$

Combining the results with inequality (E.13) proves inequality (7.15).

We now consider the special cases:

(a) In the case when $p = \Omega\left(\frac{\sqrt{n}\log n}{n}\right)$, we can use fact that at least $np = \Omega(\sqrt{n}\log n)$ points have marginal variance at most σ_2. Let $r' := \frac{C\sigma_2 \log n}{p\sqrt{n}}$. By similar reasoning as above, for at least np distributions, we have $\int_{[-r',r']} q_i(x)dx \geq \frac{\log n}{p\sqrt{n}}$. Thus, we can replace inequality (E.14) by

$$\sum_{i=1}^{n}\int_{[-r',r']} q_i(x)dx \geq np\cdot\frac{2\log n}{p\sqrt{n}} \geq 2\sqrt{n}\log n,$$

to conclude that $r_{2k_1,1} = O\left(\frac{\sigma_2 \log n}{p\sqrt{n}}\right)$. This leads to the stated bound.

(b) In this case, we will obtain a better bound by showing that $\|\widehat{\mu}_{S,k_2}\|_2 \le r_{2k_2}$, w.h.p., rather than the looser bound $\|\widehat{\mu}_{S,k_2}\|_2 \le C'\sqrt{n}^{1/d} r_{k_2}$ used to derive inequality (E.13) (cf. Theorem 7.4.6). Since $r_{2k_2} \le C\sigma_2\sqrt{d}$, the tighter bound will then follow.

Let $r' := C'\sqrt{d \log n}\sigma_2$. As argued in the proof of Theorem 7.4.5, it suffices to show that $R(f_{r',r_{2k}}) \le \frac{k}{2n}$, where $k = k_2$. We will deal with low-variance and high-variance points separately.

First, consider i such that $\nu_i = \Omega(\sigma_1) = \Omega(\sigma_2 n^{\frac{1}{d}}) \ge C''\sigma_2 n^{\frac{1}{d}}$ for large C'', and let v_d denote the volume of the ball of radius 1. Then

$$\mathbb{P}\left(X_i \in B(r', r_{2k})\right) \le \mathbb{P}\left(X_i \in B(0, r_{2k})\right) \le f_i(0) v_d r_{2k}^d \le \left(\frac{c'}{C''\sigma_2 n^{1/d}}\right)^d v_d \sigma_2^d C^d \sqrt{d}^d \le \frac{1}{n},$$

where we use condition (7.14) and the fact that $\frac{v_d\sqrt{d}^d}{\tilde{C}^d} \le 1$ for a sufficiently large constant $\tilde{C}$.

Now consider i such that $\nu_i \le \sigma_2$. By condition (7.14), we have

$$\mathbb{P}\left(X_i \in B(r', r_{2k})\right) \le \exp(-c_1 \log n) \le \frac{1}{n^{c_1}}.$$

For large enough C', we can ensure that $c_1 \ge 1$. Altogether, we conclude that

$$R(f_{r',r_{2k}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(X_i \in B(r', r_{2k})\right) \le \frac{1}{n} < \frac{k_2}{2n},$$

which concludes the proof.

E.5.5 Details for Table 7.2

1. Large Heterogeneity

   - Upper bound: We have $\frac{\sigma_1}{\sigma_2} = \Omega(n^{1/d})$. Since $\sigma_2 = 1$, Theorem 7.5.7(b) states that the error of the hybrid estimator is bounded as follows:

   $$\mathbb{E}\,\|\widehat{\mu} - \mu\|_2 \le C''_u\sqrt{d}\sqrt{\log n}.$$

   - Lower bound: As remarked after Theorem 7.5.7, the lower bounds for the class $\mathcal{P}(\sigma_1, \sigma_2, p)$ also hold for the class $\mathcal{Q}(\sigma_1, \sigma_2, p)$, because these families

share the class of distributions used in the proof of Theorem 7.5.5. Using Theorem 7.5.5(a), the error of any estimator $\widehat{\mu}$ is bounded from below as follows:

$$\mathbb{E}\left\|\widehat{\mu}-\mu\right\|_2 \geq C_\ell \sqrt{d} \min\left(\frac{1}{\sqrt{np}}, \frac{\sigma_1}{\sqrt{n}}\right) = C_\ell \frac{\sqrt{d}}{\sqrt{np}} \min\left(1, \sigma_1\sqrt{p}\right) = \Omega\left(\frac{\sqrt{d}}{\sqrt{np}}\right),$$

where we use the fact that $\sigma_1\sqrt{p} = \Omega(1)$ by assumption.

2. Mild Heterogeneity

   - Upper bound: Since $\sigma_2 = 1$, inequality (7.15) in Theorem 7.5.7 states that the error of the hybrid estimator is bounded as follows:

   $$\mathbb{E}\left\|\widehat{\mu}-\mu\right\|_2 \leq C_u'' \sqrt{d}\sigma_1 \frac{\log n}{\sqrt{n}}.$$

   - Lower bound: Using Theorem 7.5.5(a), the error of any estimator $\widehat{\mu}$ is bounded from below as follows:

   $$\mathbb{E}\left\|\widehat{\mu}-\mu\right\|_2 \geq C_\ell \sqrt{d} \min\left(\frac{1}{\sqrt{np}}, \frac{\sigma_1}{\sqrt{n}}\right) = C_\ell \frac{\sqrt{d}}{\sqrt{n}} \min\left(\frac{1}{\sqrt{p}}, \sigma_1\right) = \Omega\left(\frac{\sqrt{d}\sigma_1}{\sqrt{n}}\right),$$

   where we use the fact that $\sigma_1 = O(1/\sqrt{p})$ by assumption.

3. Large p

   - Upper bound: As $p = \Omega\left(\frac{\sqrt{n}\log n}{n}\right)$, Theorem 7.5.7(a) states that the error of the hybrid estimator is bounded as follows:

   $$\mathbb{E}\left\|\widehat{\mu}-\mu\right\|_2 \leq C_u' \sqrt{d}\log n \min\left(\frac{1}{p\sqrt{n}}, \frac{\sigma_1}{\sqrt{n}}\right).$$

   - Lower bound: The lower bound follows directly from Theorem 7.5.5(a).

E.6 Proofs for Alternative Conditions

In this appendix, we prove the statements of the results in Section 7.7.

E.6.1 Proof of Theorem 7.7.1

We first prove claim (i). Note that the result of Lemma E.2.2 will still hold, since it only depends on the uniform concentration bound and optimality of the modal interval estimator. Thus, $R(f_{\widehat{\mu}_{M,r},r}) \geq \frac{R^*_r}{2}$, w.h.p.

For a fixed value of r', define $\widehat{\mu}' = \frac{\widehat{\mu}_{M,r}}{\|\widehat{\mu}_{M,r}\|_2} \cdot r'$ to be the rescaled version of $\widehat{\mu}_{M,r}$. By condition (C1), we will have $\|\widehat{\mu}_{M,r}\|_2 \leq r'$ if we can show that $R(f_{\widehat{\mu}',r}) \leq R(f_{\widehat{\mu}_{M,r},r})$. Note that

$$R(f_{\widehat{\mu}',r}) \leq g(r', r),$$

so if we choose r' sufficiently large so that $g(r', r) < \frac{R^*_r}{2}$, the desired inequality will hold.

Turning to claim (ii), note that Lemma 7.4.4 continues to hold, since it only relies on the uniform concentration bound and a Chernoff bound. We thus conclude that $R(f_{\widehat{\mu}_{S,k},r_{2k}}) \geq \frac{k}{4n} = \frac{R^*_{2k}}{4}$, w.h.p. For a fixed value of r', we define $\widehat{\mu}' = \frac{\widehat{\mu}_{M,r_{2k}}}{\|\widehat{\mu}_{M,r_{2k}}\|_2} \cdot r'$. By condition (C1) (which we only need to assume holds for $r = r_{2k}$), if $R(f_{\widehat{\mu}',r_{2k}}) \leq R(f_{\widehat{\mu}_{M,r_{2k}},r_{2k}})$, then $\|\widehat{\mu}_{M,r_{2k}}\|_2 \leq r'$. Furthermore, $R(f_{\widehat{\mu}',r_{2k}}) \leq g(r', r_{2k})$, so we simply need to choose r' such that $g(r', r_{2k}) < \frac{1}{4}$.

For the hybrid estimator, note that Lemma E.4.1 shows that the output is always within $\sqrt{d} r_{4\sqrt{n \log n},1}$ of the output. Furthermore, the output of shorth estimator is always with r' of the origin by part (ii). If the shorth estimator lies outside the $S^\infty_{\sqrt{n \log n}}$, then its ℓ_2 projection on $S^\infty_{\sqrt{n \log n}}$ will only decrease its distance from the origin because (1) the origin belongs to $S^\infty_{\sqrt{n \log n}}$; and (2) $S^\infty_{\sqrt{n \log n}}$ is convex.

E.6.2 Proof of Proposition 7.7.3

We first show that for each $r > 0$, the functions $R_i(f_{x,r})$ are unimodal as functions of $x \in \mathbb{R}^d$. Let q be the uniform distribution on the Euclidean ball of radius r. Then $p_i \star q$, being a convolution of two log-concave densities, is also log-concave. Log-concave densities by definition are proportional to $e^{-\phi(x)}$ for some convex function ϕ, and therefore they are unimodal and monotonically decreasing along rays from the mode. Now note that if condition (C3) holds, then $R_i(f_{x,r})$ must also be symmetric around 0. Hence, if $R_i(f_{x,r})$ is unimodal, its unique mode must clearly occur at 0. This proves that conditions (C2) and (C3) together imply condition (C1).

For the second statement, it suffices to verify the inequality

$$\sup_{\|x\|_2 = a} R_i(f_{x,r}) \leq \frac{1}{\lfloor a/2r \rfloor}, \qquad \forall i. \tag{E.15}$$

Indeed, we would then have

$$g(a,r) = \sup_{\|x\|_2=a} \frac{1}{n}\sum_{i=1}^{n} R_i(f_{x,r}) \leq \frac{1}{n}\sum_{i=1}^{n} \sup_{\|x\|_2=a} R_i(f_{x,r}) \leq \frac{1}{\lfloor a/2r \rfloor}.$$

Thus, it remains to verify inequality (E.15). Focusing on a particular i, consider $x \in \mathbb{R}^d$ such that $\|x\|_2 = a$. We know that $R_i(f_{x,r})$ is decreasing on the ray from 0 to x. Furthermore, we can pack $\lfloor \frac{a}{2r} \rfloor$ balls of radius r on the ray, including the balls $B(x_i^*, r)$ and $B(x, r)$ at the endpoints. The total mass of these balls is clearly upper-bounded by 1. Hence,

$$\left\lfloor \frac{a}{2r} \right\rfloor \cdot R_i(f_{x,r}) \leq 1,$$

implying the desired result.

E.6.3 Proof of Proposition 7.7.5

Let X have an elliptically symmetric density defined as $p_X(x) = f(x^\top \Sigma^{-1} x)$ for a decreasing function $f : \mathbb{R} \to \mathbb{R}$. Consider a point $x_0 \in \mathbb{R}^d$ such that $\|x_0\|_2 = r_2$, and consider the ball $B(x_0, r_1) = \{x \in \mathbb{R}^d \ : \ \|x - x_0\| \leq r_1\}$. For analysis purposes, we first transform the elliptically symmetric density to a spherically symmetric, decreasing density. This may be achieved by applying the linear transformation $\Sigma^{-1/2} : \mathbb{R}^d \to \mathbb{R}^d$. Define $Y := \Sigma^{-1/2} X$, let $\Sigma^{-1/2} x_0 = y_0$, and let $\hat{B}$ be the image of $B(x_0, r_1)$ under the transformation $\Sigma^{-1/2}$. Note that

$$\hat{B} = \left\{ y \in \mathbb{R}^d \ : \ (y - y_0)^\top \Sigma (y - y_0) \leq r_1 \right\},$$

and further note that $R(f_{x_0,r_1})$ is equal to the integral of $p_Y(\cdot)$ over $\hat{B}$; i.e., $\mathbb{P}(Y \in \hat{B})$. It is easy to see that $\hat{B} \subseteq B\left(y_0, \frac{r_1}{\lambda_{\min}(\Sigma)}\right)$. Hence,

$$R(f_{x_0,r_1}) = \mathbb{P}(Y \in \hat{B}) \leq \mathbb{P}(Y \in B\left(y_0, \frac{r_1}{\lambda_{\min}(\Sigma)}\right).$$

We may now use the strategy from Lemma 7.2.5, to obtain

$$\begin{aligned}
1 &\geq \mathbb{P}(Y \in B(0, \|y_0\|_2)) \\
&\geq P\left(B(0, \|y_0\|_2), \frac{r_1}{\lambda_{\min}(\Sigma)}\right) \cdot \mathbb{P}\left(Y \in B\left(y_0, \frac{r_1}{\lambda_{\min}(\Sigma)}\right)\right) \\
&\geq P\left(B\left(0, \frac{r_2}{\lambda_{\max}(\Sigma)}\right), \frac{r_1}{\lambda_{\min}(\Sigma)}\right) \cdot R(f_{x_0,r_1}).
\end{aligned}$$

Since this inequality holds for any x_2 with $\|x_2\|_2 = r_2$, we conclude that

$$\begin{aligned} g(r_2, r_1) &\leq \frac{1}{P\left(B\left(0, \frac{r_2}{\lambda_{\max}(\Sigma)}\right), \frac{r_1}{\lambda_{\min}(\Sigma)}\right)} \\ &\leq C\left(\frac{r_1 \lambda_{\max}(\Sigma)}{r_2 \lambda_{\min}(\Sigma)}\right)^d. \end{aligned}$$

E.6.4 Proof of Proposition 7.7.8

We index the distributions so that $\{R_i\}_{i=1}^s$ are radially symmetric. Note that

$$g(a, r) = \sup_{\|x\|_2 = a} R(f_{x,r}) \leq \frac{1}{n} \sum_{i=1}^{n} \sup_{\|x\|_2 = a} R_i(f_{x,r}).$$

Furthermore, for each $1 \leq i \leq s$, we have

$$\sup_{\|x\|_2 = a} R_i(f_{x,r}) \leq \left(\frac{r}{a}\right)^d R_i(f_{0,a}) \leq \left(\frac{r}{a}\right)^d.$$

On the other hand, for $i > s$, we have

$$\sup_{\|x\|_2 = a} R_i(f_{x,r}) \leq \frac{r}{a}.$$

Hence,

$$g(a, r) \leq \frac{s}{n}\left(\frac{r}{a}\right)^d + \frac{n-s}{n}\left(\frac{r}{a}\right).$$

Now note that $R^*_{q_{(f(n))}} \geq \frac{f(n)}{2n}$. Thus,

$$g(r', r) \leq \frac{s}{n} \cdot \frac{1}{2^d n} + \frac{n-s}{n} \cdot \frac{1}{2n^{1/d}} \leq \frac{1}{n} + \frac{n-s}{n} \cdot \frac{1}{2n^{1/d}} < \frac{f(n)}{4n} \leq \frac{R^*_r}{2},$$

using the assumed lower bound on s.

E.7 Proofs for Regression

In this appendix, we provide the proofs of the statements in Section 7.8.

E.7.1 Proof of Proposition 7.8.1

We write

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}\left[1\left\{|y_i - x_i^\top \beta| \le r\right\}\right] &= \sum_{i=1}^{n} \mathbb{P}\left(|y_i - x_i^\top \beta| \le r\right) \\ &= \sum_{i=1}^{n} \mathbb{P}\left(|x_i^\top (\beta^* - \beta) + \epsilon_i| \le r\right).\end{aligned}$$

Note that conditioned on x_i, each summand is maximized uniquely when $x_i^\top(\beta^* - \beta) = 0$, since the distribution of ϵ_i is symmetric and unimodal. Since

$$\sum_{i=1}^{n} \mathbb{E}\left[1\left\{|y_i - x_i^\top \beta| \le r\right\}\right] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}\left[1\left\{|y_i - x_i^\top \beta| \le r\right\} \mid \{x_i\}_{i=1}^n\right]\right], \tag{E.16}$$

we see that the right-hand expression in equation (E.16) is therefore maximized when $\beta = \beta^*$. On the other hand, we can also argue that the maximizer is unique. Indeed, suppose $\beta \in \mathbb{R}^d$ were such that $\beta \neq \beta^*$. The set

$$\mathcal{S} := \left\{\{x_i\}_{i=1}^n \subseteq (\mathbb{R}^d)^n : x_i^\top(\beta - \widehat{\beta}) = 0 \quad \forall i\right\}$$

has Lebesgue measure 0. We can write

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}\left[1\left\{|y_i - x_i^\top \beta| \le r\right\} \mid \{x_i\}_{i=1}^n\right]\right] &= \int_{\{x_i\}\in\mathcal{S}} \mathbb{E}\left[1\left\{|y_i - x_i^\top \beta| \le r\right\} \mid \{x_i\}_{i=1}^n\right] d\,\mathbb{P}(\{x_i\}) \\ &\quad + \int_{\{x_i\}\notin\mathcal{S}} \mathbb{E}\left[1\left\{|y_i - x_i^\top \beta| \le r\right\} \mid \{x_i\}_{i=1}^n\right] d\,\mathbb{P}(\{x_i\}).\end{aligned}$$

Noting that

$$\begin{aligned}\mathbb{E}\left[1\left\{|y_i - x_i^\top \beta| \le r\right\} \mid \{x_i\}_{i=1}^n\right] &= \mathbb{E}\left[1\left\{|y_i - x_i^\top \beta^*| \le r\right\} \mid \{x_i\}_{i=1}^n\right], \quad \forall \{x_i\} \in \mathcal{S}, \\ \mathbb{E}\left[1\left\{|y_i - x_i^\top \beta| \le r\right\} \mid \{x_i\}_{i=1}^n\right] &< \mathbb{E}\left[1\left\{|y_i - x_i^\top \beta^*| \le r\right\} \mid \{x_i\}_{i=1}^n\right], \quad \forall \{x_i\} \notin \mathcal{S},\end{aligned}$$

completes the proof.

E.7.2 Proof of Theorem 7.8.3

The proof follows the same approach used to prove estimation error bounds for the modal interval estimator throughout the paper (e.g., Theorem 7.3.1). By Lemma 7.8.2, we know that $R_{\widehat{\beta}} \ge \frac{R_{\beta^*}}{2}$, w.h.p. We will be done if we can show that $R_\beta < \frac{R_{\beta^*}}{2}$ for all β

satisfying

$$\|\beta - \beta\|_2 > \frac{c' n \sigma_{(cd \log n)}}{\lambda_{\min}}. \tag{E.17}$$

First note that

$$R_{\beta^*} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(|\epsilon_i| \leq r).$$

Hence, as argued for mean estimation, we certainly have $r \leq C' \sigma_{(Cd \log n)}$.

Also note that for any $\beta \in \mathbb{R}^d$, we have

$$y_i - x_i^\top \beta = \epsilon_i + x_i^\top (\beta^* - \beta) \sim N\left((\beta^* - \beta)^\top \mu_i', (\beta^* - \beta)^\top \Sigma_i' (\beta^* - \beta)\right).$$

Let $\mathcal{J}$ denote the set of indices of the smallest $d \log n$ of the σ_i's. Note that

$$R_{\beta^*} \geq \frac{1}{n} \sum_{i \in \mathcal{J}} \mathbb{P}(|\epsilon_i| \leq r) \geq 2r \cdot \frac{c}{n} \sum_{i=1}^{d \log n} \frac{1}{\sqrt{2\pi} \sigma_{(i)}},$$

since the Gaussian pdf decreases by a factor of $\approx 68\%$ within one standard deviation of 0.

Now suppose $\beta \in \mathbb{R}^d$ satisfies inequality (E.17). We have

$$R_\beta \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}\left(|z_i| \leq r\right),$$

where $z_i \sim N\left(0, \sigma_i^2 + (\beta^* - \beta)^\top \Sigma_i' (\beta^* - \beta)\right)$. For $i \notin \mathcal{J}$, we write

$$\mathbb{P}(|z_i| \leq r) \leq 2r \cdot \frac{1}{\sqrt{2\pi} \sqrt{\sigma_i^2 + (\beta^* - \beta)^\top \Sigma_i' (\beta^* - \beta)}} \leq \frac{2r}{n \sigma_{(d \log n)} \sqrt{2\pi}},$$

since by the choice of β, we have

$$(\beta^* - \beta)^\top \Sigma_i' (\beta^* - \beta) \geq \lambda_{\min} \|\beta - \beta^*\|_2^2 \geq n^2 \sigma_{(d \log n)}^2.$$

For $i \in \mathcal{J}$, we write

$$\mathbb{P}(|z_i| \leq r) \leq 2r \cdot \frac{1}{\sqrt{2\pi} \sqrt{\sigma_i^2 + (\beta^* - \beta)^\top \Sigma_i' (\beta^* - \beta)}} \leq \frac{2r}{3\sigma_i^2 \sqrt{2\pi}},$$

since by the choice of β, we have

$$(\beta^* - \beta)^\top \Sigma_i'(\beta^* - \beta) \geq 2\sigma^2_{(d \log n)} \geq 2\sigma_i^2.$$

Thus, we conclude that

$$R_\beta \leq \frac{2r}{\sqrt{2\pi}} \cdot \frac{1}{n}\left(\sum_{i \in \mathcal{J}} \frac{1}{3\sigma_i^2} + \sum_{i \notin \mathcal{J}} \frac{1}{n\sigma_{(d \log n)}}\right) \leq \frac{R_{\beta^*}}{3} + \frac{c'}{n} < \frac{R_{\beta^*}}{2},$$

as wanted. This concludes the proof.

E.7.3 Proof of Theorem 7.8.4

For $i \in [n]$, consider the sets

$$U_i := \{\beta \subseteq \mathbb{R}^d : -r \leq x_i^\top \beta \leq +r\}.$$

The set U_i is sandwiched between the two hyperplanes $x_i^\top \beta = y_i - r$ and $x_i^\top \beta = y_i + r$. Denote these hyperplanes by $H_-(U_i)$ and $H_+(U_i)$, respectively. These $2n$ hyperplanes partition $\mathbb{R}^d$ into a finite number of (possibly unbounded) convex regions, which we denote by $\{R_1, \ldots, R_M\}$. Define the function $f(\beta) := \sum_{i=1}^n \mathbb{1}_{U_i}(\beta)$. Our goal is to find $\hat{\beta} = \operatorname{argmax}_{\beta \in \mathbb{R}^d} f(\beta)$, where $\mathbb{1}_{U_i}$ is the indicator function of U_i. It is easy to see that $f(\cdot)$ is constant when restricted to the interior of any fixed region R_j for $j \in [M]$. Also, since $\mathbb{1}_{U_i}$ is an upper-semicontinuous function for each $i \in [n]$, so is f. Thus, the value of $f(\cdot)$ at the vertices R_j is at least as large as the value of f in its interior. Thus, to find the maximum of $f(\cdot)$, we may only consider $\beta \in \mathbb{R}^d$ that correspond to vertices of R_j for $j \in [M]$. All such vertices may be obtained by choosing any d (mutually non-parallel) hyperplanes from among $\{H_-(U_1), \ldots, H_-(U_n), H_+(U_1), \ldots, H_+(U_M)\}$ and considering their point of intersection. The total number of such points is bounded above by $\binom{2n}{d}$, and our algorithm may simply list such points and evaluate f at each point in the list.

E.8 Auxiliary Results

This appendix contains several technical results invoked throughout the paper.

We will employ the following multiplicative Chernoff bound, which is standard (cf. Vershynin [Ver18] or Boucheron et al. [BLM13]):

Lemma E.8.1. *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameters $\{p_i\}$. Let $S_n = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[S_n]$.*

(i) For any $\delta \in (0, 1]$, we have

$$\mathbb{P}\left(S_n \geq (1+\delta)\mu\right) \leq \exp\left(-\frac{\mu\delta^2}{3}\right).$$

and

$$\mathbb{P}\left(S_n \leq (1-\delta)\mu\right) \leq \exp\left(-\frac{\mu\delta^2}{2}\right).$$

(ii) For $\delta \geq 4$, we have

$$\mathbb{P}\left(S_n \geq \delta\mu\right) \leq \exp\left(-c\mu\delta\log\delta\right).$$

We will also use the following result from Boucheron et al. [BLM13]:

Lemma E.8.2. (*Theorem 12.9 from Boucheron et al. [BLM13]*) *Let $W_1, \ldots, W_n$ be independent vector-valued random variables and let $Z = \sup_{s\in\mathcal{T}} \sum_{i=1}^n W_{i,s}$. Assume that for all $i \leq n$ and $s \in \mathcal{T}$, we have $\mathbb{E}\, W_{i,s} = 0$, and $|W_{i,s}| \leq 1$. Let*

$$v := 2\,\mathbb{E}\, Z + \rho^2,$$
$$\rho^2 := \sup_{t\in T} \sum_{i=1}^n \mathbb{E}\, W_{i,s}^2.$$

Then $\mathbf{Var}(Z) \leq v$ and

$$\mathbb{P}\{Z \geq \mathbb{E}\, Z + t\} \leq \exp\left(-\frac{t}{4}\log\left(1 + 2\log\left(1 + \frac{t}{v}\right)\right)\right).$$

We now state and prove a generalization of Theorem 13.7 from Boucheron et al.[BLM13]:

Theorem E.8.3. *Let $\mathcal{A} = \{A_t : t \in \mathcal{T}\}$ be a countable class of measurable subsets of $\mathcal{X}$ with VC dimension V, such that $A_0 = \emptyset \in \mathcal{A}$. Let $X_1, \ldots, X_n$ be independent random variables taking values in $\mathcal{X}$, with distributions $P_1, \ldots, P_n$, respectively. Assume that for some $\sigma > 0$, we have*

$$\frac{1}{n}\sum_{i=1}^n P_i(A_t) \leq \sigma^2, \text{ for every } t \in \mathcal{T}.$$

Let Z and Z^- be defined as follows:

$$Z = \frac{1}{\sqrt{n}} \sup_{t \in \mathcal{T}} \sum_{i=1}^{n} \left(\mathbb{1}_{X_i \in A_t} - P_i(A_t) \right), \quad \text{and}$$
$$Z^- = \frac{1}{\sqrt{n}} \sup_{t \in \mathcal{T}} \sum_{i=1}^{n} \left(P_i(A_t) - \mathbb{1}_{X_i \in A_t} \right).$$

If $\sigma \geq 24\sqrt{\frac{V}{5n} \log\left(\frac{4e^2}{\sigma}\right)}$, then

$$\max\left(\mathbb{E}\, Z, \mathbb{E}\, Z^-\right) \leq 72\sigma\sqrt{V \log \frac{4e^2}{\sigma}}.$$

Proof. The following proof is an adaptation of the proof of Theorem 13.7 in Boucheron et al. [BLM13]. The generalization from identical to non-identical distributions is possible because (1) independence suffices for symmetrization inequality; and (2) after conditioning on $X_1, \ldots, X_n$, it is no longer relevant whether the distributions of the random variables are identical. We include the initial steps of the proof for completeness and direct the reader to Boucheron et al. [BLM13] for more details.

By the symmetrization inequalities of Lemma 11.4 in Boucheron et al. [BLM13], we have

$$\begin{aligned} &\mathbb{E} \frac{1}{\sqrt{n}} \sup_{t \in \mathcal{T}} \sum_{i=1}^{n} \left(\mathbb{1}_{X_i \in A_t} - P(A_t) \right) \\ &\qquad \leq 2\, \mathbb{E}\left[\mathbb{E}\left[\frac{1}{\sqrt{n}} \sup_{t \in \mathcal{T}} \sum_{i=1}^{n} \epsilon_i \mathbb{1}_{X_i \in A_t} \middle| X_1, \ldots, X_n \right] \right], \end{aligned} \tag{E.18}$$

where the ϵ_i's are independent Rademacher variables. Define the random variable

$$\delta_n^2 = \max\left(\sup_{t \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \in A_t},\ \sigma^2 \right).$$

Clearly, $\delta_n^2 \leq \frac{Z}{\sqrt{n}} + \sigma^2$, so by Jensen's inequality,[37]

$$\mathbb{E}\, \delta_n \leq \sqrt{\mathbb{E}\left(\frac{Z}{\sqrt{n}} \right) + \sigma^2}.$$

[37] Note that both Z and Z^- are non-negative since $\phi \in \mathcal{A}$.

Now let $Z_t = \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i \mathbb{1}_{X_i \in A_t}$. Noting that the Rademacher averages are sub-Gaussian, conditioned on the X_i's, we have

$$\begin{aligned}
&\log \mathbb{E}\left[e^{\lambda(Z_t - Z_{t'})} \middle| X_1, \ldots, X_n \right] \\
&\quad \leq \frac{\lambda^2 \left(\frac{1}{n} \sum_{i=1}^n (\mathbb{1}_{X_i \in A_t} - \mathbb{1}_{X_i \in A_{t'}})^2 \right)}{2} \\
&\quad = \frac{\lambda^2 \left(\frac{1}{n} \sum_{i=1}^n (\mathbb{1}_{X_i \in A_t} \neq \mathbb{1}_{X_i \in A_{t'}}) \right)}{2}.
\end{aligned}$$

Let $d(t, t') = \sqrt{\frac{1}{n} \sum_{i=1}^n (\mathbb{1}_{X_i \in A_t} \neq \mathbb{1}_{X_i \in A_{t'}})}$, and let $H(\delta, \mathcal{T})$ denote the universal δ-metric entropy (with respect to $d(\cdot, \cdot)$). Since the zero function (corresponding to $\emptyset$) belongs to the function class, we have

$$\sup_{t \in \mathcal{T}} d(t, 0) = \sup_{t \in \mathcal{T}} \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \in A_t}} \leq \delta_n.$$

Therefore, we can apply the discrete version of Dudley's inequality (Lemma 13.1 in Boucheron et al. [BLM13]) with δ_n as the maximum radius. Since $\delta_n \geq \sigma$, we can upper-bound the random quantity $H(a\delta_n)$ by the fixed quantity $H(a\sigma)$, for any $a > 0$. This implies that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\sqrt{n}} \sup_{t \in \mathcal{T}} \sum_{i=1}^n \epsilon_i \mathbb{1}_{X_i \in A_t} \middle| X_1, \ldots, X_n \right] \\
&\leq 3 \sum_{j=0}^{\infty} \delta_n 2^{-j} \sqrt{H(\delta_n 2^{-j-1}, \mathcal{T})} \\
&\leq 3 \sum_{j=0}^{\infty} \delta_n 2^{-j} \sqrt{H(\sigma 2^{-j-1}, \mathcal{T})}.
\end{aligned}$$

Taking the expectation with respect to $X_1, \ldots, X_n$ and combining with inequality (E.18) we then obtain

$$\begin{aligned}
\mathbb{E}\, Z &\leq 6\, \mathbb{E}\, \delta_n \cdot \sum_{j=1}^{\infty} 2^{-j} \sqrt{H(\sigma 2^{-j-1}, \mathcal{T})} \\
&\leq 6 \sqrt{\mathbb{E}\left(\frac{Z}{\sqrt{n}} \right) + \sigma^2} \left(\sum_{j=1}^{\infty} 2^{-j} \sqrt{H(\sigma 2^{-j-1}, \mathcal{T})} \right).
\end{aligned}$$

From this step onward, the proof is identical to the proof of Theorem 13.7 in Boucheron

et al. [BLM13]. □

Theorem E.8.4. (*Theorem 8.3.23 in Vershynin*[*Ver18*]) *Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ) with finite VC dimension $V \geq 1$. Let $X, X_1, X_2, \ldots, X_n$ be independent random points in Ω distributed according to the law μ. Then*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)\right|\right] \leq C\sqrt{\frac{V}{n}}.$$

F APPENDIX TO CHAPTER 9

F.1 Omitted Proofs from Section 9.2: Technical Details Regarding Stability

Lemma 9.2.11 (Certificate Lemma). *Let G be an (ϵ, δ)-stable distribution with respect to $\mu \in \mathbb{R}^d$, for some $0 < \epsilon < 1/3$ and $\delta \geq \epsilon$. Let P be a distribution with $d_{\mathrm{TV}}(P, G) \leq \epsilon$. Denoting by $\mu_P, \mathbf{\Sigma}_P$ the mean and covariance of P, if $\lambda_{\max}(\mathbf{\Sigma}_P) \leq 1 + \lambda$, for some $\lambda \geq 0$, then $\|\mu_P - \mu\|_2 = O(\delta + \sqrt{\epsilon\lambda})$.*

Proof. Let $d_{\mathrm{TV}}(P, G) = \alpha$. By Fact 9.2.4 we can write $P = (1-\alpha)G_0 + \alpha B$. We may assume without loss of generality $\alpha = \epsilon$, since we can always treat a part of the inliers as outliers. Denoting by $\mu_P, \mathbf{\Sigma}_P$ the mean and covariance of P, and using $\mu_{G_0}, \mu_B, \mathbf{\Sigma}_{G_0}, \mathbf{\Sigma}_B$ for the corresponding quantities of the other two distributions, we have that

$$\mathbf{\Sigma}_P = (1-\epsilon)\mathbf{\Sigma}_{G_0} + \epsilon\mathbf{\Sigma}_B + \epsilon(1-\epsilon)(\mu_{G_0} - \mu_B)(\mu_{G_0} - \mu_B)^\top .$$

Letting v be the unit vector in the direction of $\mu_{G_0} - \mu_B$, we have that

$$1 + \lambda \geq v^\top \mathbf{\Sigma}_P v = (1-\epsilon)v^\top \mathbf{\Sigma}_{G_0} v + \epsilon v^\top \mathbf{\Sigma}_B v + \epsilon(1-\epsilon)(v^\top(\mu_{G_0} - \mu_B))^2 . \tag{F.1}$$

The second term of the left-hand side is nonzero and the third one is just $\epsilon(1-\epsilon)\|\mu_{G_0} - \mu_B\|_2^2$. We now focus on the first term, which by adding and subtracting μ (the vector realizing the definition of stability for G) can be written as

$$(1-\epsilon)\mathop{\mathbb{E}}_{X\sim G_0}[(v^\top(X - \mu_{G_0}))^2] = (1-\epsilon)\left(\mathop{\mathbb{E}}_{X\sim G_0}[(v^\top(X-\mu))^2] - (v^\top(\mu - \mu_{G_0}))^2\right) . \tag{F.2}$$

We note that in the decomposition of Fact 9.2.4, we can write $G_0(x) = w_0(x)G(x)$ with

$$w_0(x) = \frac{1}{1-\epsilon}\begin{cases} P(x)/G(x) , & \text{if } G(x) > P(x) \\ 1 , & \text{otherwise} . \end{cases}$$

Letting $h(x) := (1-\epsilon)w_0(x)$ we have that $h(x) \leq 1$ for all x and $\mathbb{E}_{X\sim G}[h(X)] = 1-\epsilon$, thus $G_0(x) = h(x)G(x)/(\int h(x)G(x)\mathrm{d}x) =: G_h(x)$. Returning to Equation (F.2), this means

that

$$\mathop{\mathbb{E}}_{X\sim G_0}[(v^\top(X-\mu))^2] = \mathop{\mathbb{E}}_{X\sim G_h}[(v^\top(X-\mu))^2] = v^\top \overline{\mathbf{\Sigma}}_{h,G} v \geq 1 - \frac{\delta^2}{\epsilon}\,, \tag{F.3}$$

by applying stability. Similarly, the other term in Equation (F.2) is $(v^\top(\mu - \mu_{G_0}))^2 \leq \delta^2$. Putting everything together, Equation (F.1) becomes

$$\begin{aligned} 1+\lambda &\geq (1-\epsilon)(1-\delta^2/\epsilon - \delta^2) + \epsilon(1-\epsilon)\|\mu_{G_0} - \mu_B\|_2^2 \\ &\geq 1 - 3\delta^2/\epsilon + (\epsilon/2)\|\mu_{G_0} - \mu_B\|_2^2\,, \end{aligned}$$

which yields $\|\mu_{G_0} - \mu_B\|_2 \lesssim \sqrt{\lambda/\epsilon} + \delta/\epsilon$. Then, writing $\mu_P = (1-\epsilon)\mu_{G_0} + \epsilon\mu_B$ and using stability follows that $\|\mu_P - \mu\|_2 \lesssim \delta + \sqrt{\lambda\epsilon}$. □

Lemma 9.2.12. *For any $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$, if a distribution G is $(2\epsilon, \delta)$-stable with respect to $\mu \in \mathbb{R}^d$, and P is an ϵ-corrupted version of G in total variation distance, there exist distributions G_0 and B such that $P = (1-\epsilon)G_0 + \epsilon B$ and G_0 is (ϵ, δ)-stable with respect to μ.*

Proof. By Fact 9.2.4, we have the decomposition $P = (1-\epsilon)G_0 + \epsilon B$, where $G_0(x) = \min\{G(x), P(x)\}/(1-\epsilon)$. We can write $G_0(x) = w_0(x)G(x)$, where

$$w_0(x) = \frac{1}{1-\epsilon}\begin{cases} P(x)/G(x)\,, & \text{if } G(x) > P(x) \\ 1\,, & \text{otherwise}\,. \end{cases}$$

To see why the final claim is true, we consider a weight function $w : \mathbb{R}^d \to [0,1]$ such that $\mathbb{E}_{X\sim G_0}[w(X)] \geq 1-\epsilon$ and examine the adjusted distribution G_{0w}. We have that

$$G_{0w}(x) = \frac{w(x)G_0(x)}{\int_{\mathbb{R}^d} w(x)G_0(x)\mathrm{d}x} = \frac{(1-\epsilon)w(x)w_0(x)G(x)}{\int_{\mathbb{R}^d}(1-\epsilon)w(x)w_0(x)G(x)\mathrm{d}x} = \frac{h(x)G(x)}{\mathbb{E}_{X\sim G}[h(X)]} = G_h(x)\,,$$

where we let $h(x) := (1-\epsilon)w(x)w_0(x)$. We have that $h(x) \leq 1$ point-wise and $\int_{\mathbb{R}^d} h(x)G(x)\mathrm{d}x = (1-\epsilon)\,\mathbb{E}_{X\sim G_0}[w(X)] \geq (1-\epsilon)^2 \geq 1-2\epsilon$. Recalling that G is $(2\epsilon,\delta)$-stable, the conclusion follows. □

Lemma 9.2.13. *Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. Let $w : \mathbb{R}^d \to [0,1]$ such that $\mathbb{E}_{X\sim G}[w(X)] \geq 1-\epsilon$ and let G be an (ϵ,δ)-stable distribution with respect to $\mu \in \mathbb{R}^d$. For any matrix $\mathbf{U} \in \mathbb{R}^{d\times d}$ and any vector $b \in \mathbb{R}^d$, we have that*

$$\mathop{\mathbb{E}}_{X\sim G_w}\left[\|\mathbf{U}(X-b)\|_2^2\right] = \|\mathbf{U}\|_F^2(1 \pm \delta^2/\epsilon) + \|\mathbf{U}(\mu - b)\|_2^2 \pm 2\delta\,\|\mathbf{U}\|_F^2\|\mu - b\|_2\,.$$

Proof. We can write

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim G_w}\left[\|\mathbf{U}(X-b)\|_2^2\right] &= \mathop{\mathbb{E}}_{X\sim G_w}[\|\mathbf{U}(X-\mu)\|_2^2 + \|\mathbf{U}(\mu-b)\|_2^2 + 2(X-\mu)^\top\mathbf{U}^\top\mathbf{U}(\mu-b)] \\ &= \mathop{\mathbb{E}}_{X\sim G_w}\left[\|\mathbf{U}(X-\mu)\|_2^2\right] + \|\mathbf{U}(\mu-b)\|_2^2 + 2(\mu_{w,G}-\mu)^\top\mathbf{U}^\top\mathbf{U}(\mu-b)\ . \end{aligned} \tag{F.4}$$

We now focus on the first term. Let the spectral decomposition $\mathbf{U}^\top\mathbf{U} = \mathrm{tr}(\mathbf{U}^\top\mathbf{U})\sum_{i=1}^d \alpha_i v_i v_i^\top$, where $\sum_{i=1}^d \alpha_i = 1$ and $\alpha_i \geq 0$. We have that

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim G_w}\left[\|\mathbf{U}(X-\mu)\|_2^2\right] &= \mathrm{tr}\left(\mathbf{U}^\top\mathbf{U}\mathop{\mathbb{E}}_{X\sim G_w}[(X-\mu)(X-\mu)^\top]\right) = \mathrm{tr}(\mathbf{U}^\top\mathbf{U})\sum_{i=1}^d \alpha_i \mathrm{tr}(v_i v_i^\top \overline{\mathbf{\Sigma}}_{w,G}) \\ &= \mathrm{tr}(\mathbf{U}^\top\mathbf{U})\sum_{i=1}^d \alpha_i v_i^\top\overline{\mathbf{\Sigma}}_{w,G}v_i = \mathrm{tr}(\mathbf{U}^\top\mathbf{U})(1\pm\delta^2/\epsilon) = \|\mathbf{U}\|_F^2(1\pm\delta^2/\epsilon)\ ,\end{aligned}$$

where the second from the end relation is due to stability. Regarding the last term of Equation (F.4), we have that

$$\begin{aligned}|(\mu_{w,G}-\mu)^\top\mathbf{U}^\top\mathbf{U}(\mu-b)| &= |\mathrm{tr}(\mathbf{U}^\top\mathbf{U}(\mu-b)(\mu_{w,G}-\mu)^\top)| \leq \mathrm{tr}(\mathbf{U}^\top\mathbf{U})\|(\mu-b)(\mu_{w,G}-\mu)^\top\|_2 \\ &= \|\mathbf{U}\|_F^2\|\mu-b\|_2\|\mu_{w,G}-\mu\|_2 \leq \|\mathbf{U}\|_F^2\delta\|\mu-b\|_2\ ,\end{aligned}$$

where the last inequality uses stability condition for the mean. □

Corollary 9.2.14. *Fix $0<\epsilon<1/2$ and $\delta\geq\epsilon$. Let G be an (ϵ,δ)-stable distribution with respect to $\mu\in\mathbb{R}^d$. Let a matrix $\mathbf{U}\in\mathbb{R}^{d\times d}$ and a function $w:\mathbb{R}^d\to[0,1]$ with $\mathbb{E}_{X\sim G}[w(X)]\geq 1-\epsilon$. For the function $\tilde{g}(x) = \|\mathbf{U}(x-b)\|_2^2$, we have that*

$$\begin{aligned}(1-\epsilon)\|\mathbf{U}\|_F^2(1-\delta^2/\epsilon-2\delta\|b-\mu\|_2) &\leq \mathop{\mathbb{E}}_{X\sim G}[w(X)\tilde{g}(X)] \\ &\leq \|\mathbf{U}\|_F^2\left(1+\delta^2/\epsilon+\|b-\mu\|_2^2+2\delta\|b-\mu\|_2\right)\ .\end{aligned}$$

Proof. Beginning with the upper bound, we have the following inequalities:

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim G}[w(X)\tilde{g}(X)] &= \mathop{\mathbb{E}}_{X\sim G}[w(X)]\mathop{\mathbb{E}}_{X\sim G_w}[\tilde{g}(X)] \\ &\leq \mathop{\mathbb{E}}_{X\sim G_w}[\tilde{g}(X)] \qquad\qquad \text{(Using } \tilde{g}(x)\geq 0 \text{ and } w(x)\leq 1\text{)} \\ &\leq \|\mathbf{U}\|_F^2(1+\delta^2/\epsilon) + \|\mathbf{U}(\mu-b)\|_2^2 + 2\delta\|\mathbf{U}\|_F^2\|b-\mu\|_2 \\ &\leq \|\mathbf{U}\|_F^2\left(1+\delta^2/\epsilon+\|b-\mu\|_2^2+2\delta\|b-\mu\|_2\right)\ ,\end{aligned}$$

where the second inequality from the end uses Lemma 9.2.13. The lower bound is derived similarly:

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim G}[w(X)\tilde{g}(X)] &= \mathop{\mathbb{E}}_{X\sim G}[w(X)] \mathop{\mathbb{E}}_{X\sim G_w}[\tilde{g}(X)] \\ &\geq (1-\epsilon) \mathop{\mathbb{E}}_{X\sim G_w}[\tilde{g}(X)] \\ &\geq (1-\epsilon)\|\mathbf{U}\|_F^2(1-\delta^2/\epsilon - 2\delta\|b-\mu\|_2) \ ,\end{aligned}$$

where we applied Lemma 9.2.13 in the last step. □

F.2 Omitted Proofs from Section 9.3

F.2.1 Johnson-Lindenstrauss Sketch

Lemma 9.3.5. *Fix a set of n points $x_1, \ldots, x_n \in \mathbb{R}^d$. For $t \in [K]$, define $g_t(x) := \|\mathbf{M}_t(x-\mu_t)\|_2^2$ and let $\tilde{g}_t(x), v_{t,j}$ as in Algorithm 6. If C is a sufficiently large constant and $L = C\log((n+d)K/\tau)$, with probability at least $1-\tau$, for every $t \in [K]$ we have the following:*

1. $0.8g_t(x_i) \leq \tilde{g}_t(x_i) \leq 1.2g_t(x_i)$ *for every* $i \in [n]$,
2. $0.8\|\mathbf{M}_t\|_F^2 \leq \left(\frac{1}{L}\sum_{j=1}^L \|v_{t,j}\|_2^2\right) \leq 1.2\|\mathbf{M}_t\|_F^2$.

Proof. We show that the claim holds for a fixed iteration t with probability τ/K. Recall that $\tilde{g}_t(x)$ from Algorithm 6, can be written as

$$\tilde{g}_t(x) = \|\mathbf{U}_t(x-\mu_t)\|_2^2 = \frac{1}{L}\sum_{j\in[L]}(v_{t,j}^\top(x-\mu_t))^2 = \frac{1}{L}\sum_{j\in[L]}(z_{t,j}^\top\mathbf{M}_t(x_i-\mu_t))^2 \ .$$

Applying Fact 9.2.7 with $\gamma = \tau/K$ and $u_i = \mathbf{M}_t(x_i-\mu_t)$, gives that choosing $L = C\log(nK/\tau)$ suffices to guarantee that $\tilde{g}_t(x_i)/g_t(x_i) \in [0.8, 1.2]$ for every x_i with probability $1-\tau$.

We now show the second claim. Again, fix a $t \in [K]$. Consider the orthonormal base $\{e_i\}_{i=1}^d$ of $\mathbb{R}^d$. We apply Fact 9.2.7 with $\gamma = \tau/K$ and $u_i = \mathbf{M}_t e_i$, $i \in [d]$. This yields that choosing $L = C\log(dK/\tau)$, we get that for all $i \in [d]$:

$$\begin{aligned}\frac{1}{L}\sum_{j=1}^L \mathrm{tr}(z_{t,j}z_{t,j}^\top\mathbf{M}_t e_i e_i^\top\mathbf{M}_t^\top) = \frac{1}{L}\sum_{j=1}^L(z_{t,j}^\top u_i)^2 &= [0.8,1.2]\frac{1}{L}\sum_{j=1}^L\|u_i\|^2 \\ &= [0.8,1.2]\mathrm{tr}(\mathbf{M}_t^\top\mathbf{M}_t e_i e_i^\top) \ ,\end{aligned}$$

with probability $1 - \tau/K$. Summing all these inequalities for $i = 1, \dots, d$ and noting that $\sum_{i=1}^{d} e_i e_i^\top = \mathbf{I}_d$ gives that

$$\frac{1}{L}\sum_{j=1}^{L} \mathrm{tr}(z_{t,j} z_{t,j}^\top \mathbf{M}_t^\top \mathbf{M}_t) = [0.8, 1.2]\mathrm{tr}(\mathbf{M}_t^\top \mathbf{M}_t) \ ,$$

which precisely means that $\frac{1}{L}\sum_{j=1}^{L} \|v_{t,j}\|^2 = [0.8, 1.2]\|\mathbf{M}_t\|_F^2$. To have both claims hold simultaneously, we can just apply Fact 9.2.7 for all $n + d$ points, giving the result. □

F.2.2 Proof of Lemma 9.3.6

It is more useful to think of the algorithm in the following equivalent form.

Algorithm 16 Downweighting Filter

function DownweightingFilter($P, w, \tilde{\tau}, R, T, \ell_{\max}$)
  $r \leftarrow CdR^{2+4\log d}$.
  $w' \leftarrow w, \ell \leftarrow 1$.
  while $\mathbb{E}_{X\sim P}[w'(X)\tilde{\tau}(X)] > 2T$ and $\ell \le \ell_{\max}$ **do**
    $\ell \leftarrow \ell + 1$
    $w'(x) \leftarrow w(x)(1 - \tilde{\tau}(x)/r)$
  end while
  return w'.
end function

Lemma 9.3.6. *Let $P = (1-\epsilon)G + \epsilon B$ be the empirical distribution on n samples, as in Algorithm 6. If $(1-\epsilon)\,\mathbb{E}_{X\sim G}[w(X)\tilde{\tau}(X)] \le T$, $\|\tilde{\tau}\|_\infty \le r$, and $\ell_{\max} > r/T$, then Algorithm 7 modifies the weight function w to w' such that*

1. $(1-\epsilon)\,\mathbb{E}_{X\sim G}[w(X) - w'(X)] < \epsilon\,\mathbb{E}_{X\sim B}[w(X) - w'(X)]$,
2. $\mathbb{E}_{X\sim P}[w'(X)\tilde{\tau}(X)] \le 2T$,

and the algorithm terminates after $O(\log(\ell_{\max}))$ iterations, each of which takes $O(n)$ time.

Proof. We show correctness of Algorithm 16. We denote by w_ℓ the weight function at the ℓ-th iteration of the filter, which is of the form $w_\ell(x) = w(x)(1-\tilde{\tau}(x)/r)^\ell$ for every $x \in \mathbb{R}^d$. To show the first claim, we fix an iteration ℓ for which the algorithm has not stopped yet and examine the loss in weight between that iteration and the $(\ell+1)$-th iteration. From the update rule $w_{\ell+1}(x) = w_\ell(x)(1 - \tilde{\tau}(x)/r)$ we get that $w_\ell(x) - w_{\ell+1}(x) = w_\ell(x)\tilde{\tau}(x)/r$.

Thus, the weight removed in that iteration from the good distribution is

$$(1-\epsilon)\mathop{\mathbb{E}}_{X\sim G}[w_\ell(X)-w_{\ell+1}(X)] = \frac{1-\epsilon}{r}\mathop{\mathbb{E}}_{X\sim G}[w_\ell(X)\tilde{\tau}(X)] \le \frac{1}{r}T\ ,$$

while the weight removed from the bad distribution is

$$\begin{aligned}\epsilon\mathop{\mathbb{E}}_{X\sim B}[w_\ell(X)-w_{\ell+1}(X)] &= \frac{1}{r}\epsilon\mathop{\mathbb{E}}_{X\sim B}[w_\ell(X)\tilde{\tau}(X)] \\ &= \frac{1}{r}\left(\mathop{\mathbb{E}}_{X\sim P}[w_\ell(X)\tilde{\tau}(X)] - (1-\epsilon)\mathop{\mathbb{E}}_{X\sim G}[w_\ell(X)\tilde{\tau}(X)]\right) \\ &> \frac{1}{r}T\ ,\end{aligned}$$

where the last inequality uses that $(1-\epsilon)\,\mathbb{E}_{X\sim G}[w_\ell(X)\tilde{\tau}(X)] \le (1-\epsilon)\,\mathbb{E}_{X\sim G}[w(X)\tilde{\tau}(X)] \le T$ and the fact that since the algorithm has not terminated in the ℓ-th iteration it must be true that $\mathbb{E}_{X\sim P}[w_\ell(X)\tilde{\tau}(X)] > 2T$. This completes the proof of the first claim.

Regarding runtime, it suffices to show that for any $\ell > \frac{r}{eT}$, $\mathbb{E}_{X\sim P}\left[w_\ell(X)\tilde{\tau}(X)\right] \le 2T$. This follows from the inequalities

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim P}[w_\ell(X)\tilde{\tau}(X)] &= \mathop{\mathbb{E}}_{X\sim P}\left[w(X)(1-\tilde{\tau}(X)/r)^\ell\tilde{\tau}(X)\right] \\ &\le \mathop{\mathbb{E}}_{X\sim P}[w(X)\exp(-\ell\tilde{\tau}(X)/r)\tau(X)] \\ &\le \frac{r}{e\cdot\ell}\mathop{\mathbb{E}}_{X\sim P}[w(X)] \le \frac{r}{e\cdot\ell} \le T\ ,\end{aligned}$$

where we used the fact that $xe^{-\alpha x} \le 1/(e\cdot\alpha)$ for all $x\in\mathbb{R}$. By noting that $w(x)(1-\tilde{\tau}(x)/r)^\ell$ is monotonically decreasing as ℓ grows, we can improve the running time by using a binary search implementation. This gives the logarithmic guarantee of our statement. □

F.2.3 Proof of Lemma 9.3.7

We state and prove a more general version of Lemma 9.3.7 so that it can be also used in Section 9.4. The difference is that we allow the scores to center points using a vector different from the true mean μ_t of P_t, so long as this vector is $O(\delta)$-close to μ_t in Euclidean norm. Lemma 9.3.7 is obtained by using Lemma F.2.1 below with $\widehat{\mu}_t = \mu_t$.

Lemma F.2.1. *Consider the setting of Algorithm 6 and the deterministic Condition 9.3.4.*

Moreover, let $\hat{\mu}_t$ be any vector in $\mathbb{R}^d$ with $\|\hat{\mu}_t - \mu_t\| = O(\delta)$ and define the functions

$$f_t(x) := \|\mathbf{M}_t(x - \hat{\mu}_t)\|_2^2, \qquad \tilde{f}_t(x) := \|\mathbf{U}_t(x - \hat{\mu}_t)\|_2^2 \tag{F.5}$$
$$h_t(x) := f_t(x)\,\mathbb{I}\{f_t(x) > C_3\|\mathbf{M}_t\|_F^2\lambda_t/\epsilon\}, \quad \tilde{h}_t(x) := \tilde{f}_t(x)\,\mathbb{I}\{\tilde{f}_t(x) > C_3\|\mathbf{U}_t\|_F^2\widehat{\lambda}_t/\epsilon\}\ .$$

We have that $\mathbb{E}_{X\sim G}[w_t(X)h_t(X)]$ and $\mathbb{E}_{X\sim G}[w_t(X)\tilde{h}_t(X)]$ are bounded from above by $c\lambda_t\|\mathbf{M}_t\|_F^2$ for some constant c of the form $c = C/C_2$, where C_2 is the constant used in Line 15 and C is some absolute constant.

We prove the result by using the facts from Section 9.2.2. For brevity, we will prove the results for $\tilde{h}_t$ by using the functions $\tilde{f}_t$ and the matrix $\mathbf{U}_t$; the results for h_t would follow by replacing $\tilde{f}_t$ and $\mathbf{U}_t$ with f_t and $\mathbf{M}_t$ respectively and using that $\|\mathbf{U}_t\|_F$ is close to $\|\mathbf{M}_t\|_F$ (Item 2 of the deterministic condition). We begin with Lemma F.2.2, which is a generalization of the following implication of stability: The (ϵ, δ)-stability of a distribution G implies that $\mathbb{E}_{X\sim G}[(v^\top(X-\mu))^2\,\mathbb{I}\{X \in L\}] \le 3\delta^2/\epsilon$ for any set L with mass $\mathbb{P}_{X\sim G}[X \in L] \le \epsilon$ (see, for example, Proposition C.3 of [PJL20b]). The following lemma generalizes this to having a matrix in place of v.

Lemma F.2.2. *Under the setting of Algorithm 6, the deterministic Condition 9.3.4, and using the notation of Equation (F.5), if $L_t \subseteq \mathbb{R}^d$ is a set with $\mathbb{E}_{X\sim G}[w_t(X)\,\mathbb{I}\{X \in L_t\}] \le \epsilon$, then we have that*

$$\mathop{\mathbb{E}}_{X\sim G}[w_t(X)\tilde{f}_t(X)\,\mathbb{I}\{X \in L\}] \le c\lambda_t\|\mathbf{U}_t\|_F^2\ ,$$

for some constant c of the form $c = C'/C_2$, where C' is a sufficiently large constant.

Proof. Define the new weight function $w_t'(x) = w_t(x)\,\mathbb{I}\{x \notin L_t\}$. We have assumed that the distribution G is $(C''\epsilon, \delta)$-stable. Let $\epsilon' := C''\epsilon$ for brevity. We have the following inequalities, which we explain below.

$$\begin{aligned}
\mathop{\mathbb{E}}_{X\sim G}[w_t(X)\tilde{f}_t(X)\,\mathbb{I}\{X \in L_t\}] &= \mathop{\mathbb{E}}_{X\sim G}[w_t(X)\tilde{f}_t(X)] - \mathop{\mathbb{E}}_{X\sim G}[w_t(X)\tilde{f}_t(X)\,\mathbb{I}\{X \notin L_t\}] \\
&= \mathop{\mathbb{E}}_{X\sim G}[w_t(X)\tilde{f}_t(X)] - \mathop{\mathbb{E}}_{X\sim G}[w_t'(X)\tilde{f}_t(X)] \\
&\le \|\mathbf{U}_t\|_F^2\left(1 + \frac{\delta^2}{\epsilon'} + \|\widehat{\mu}_t - \mu\|_2^2 + 2\delta\|\widehat{\mu}_t - \mu\|_2\right) \\
&\quad - (1-2\epsilon)\|\mathbf{U}_t\|_F^2\left(1 - \frac{\delta^2}{\epsilon'} - 2\delta\|\widehat{\mu}_t - \mu\|_2\right) \\
&\le \|\mathbf{U}_t\|_F^2\left(3\frac{\delta^2}{\epsilon'} + 4\delta\|\widehat{\mu}_t - \mu\|_2 + \|\widehat{\mu}_t - \mu\|_2^2\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \|\mathbf{U}_t\|_F^2 \left(3\frac{\delta^2}{\epsilon'} + 4\delta\left(\|\widehat{\mu}_t - \mu_t\|_2 + \|\mu_t - \mu\|_2\right) + 4\|\widehat{\mu}_t - \mu_t\|_2^2 + 4\|\mu_t - \mu\|_2^2\right) \\
&\leq \|\mathbf{U}_t\|_F^2 \left(3\frac{\delta^2}{\epsilon'} + \delta O(\delta + \sqrt{\epsilon'\lambda_t}) + O(\delta^2 + \epsilon'\lambda_t)\right) \\
&\leq c\|\mathbf{U}_t\|_F^2 \lambda_t \ .
\end{aligned}
$$

We note that the third line above follows by applying Corollary 9.2.14 with $\mathbf{U} = \mathbf{U}_t$ and $b = \mu_t$ on both terms of the previous line. The fifth line uses the triangle inequality. The sixth line uses the assumption that $\|\widehat{\mu}_t - \mu_t\|_2 = O(\delta)$ as well as the certificate lemma (Lemma 9.2.11). Note that the required assumption $d_{\mathrm{TV}}(P_t, P) = O(\epsilon)$ from that lemma is satisfied because $\mathbb{E}_{X\sim G}[w_t'(X)] \geq \mathbb{E}_{X\sim G}[w_t(X)] - \epsilon \geq 1 - O(\epsilon)$ (see Claim 9.3.13).

Regarding that last line, we recall Line 15 of Algorithm 6, which implies that $\lambda_t \gtrsim C_2\delta^2/\epsilon$. Thus the terms δ^2/ϵ' can be bounded as $\delta^2/\epsilon' \lesssim 1/C_2$. Using that, it can be seen that we can choose $c = C'/C_2$ for some constant $C' > 0$. □

We are now ready to prove our result.

Proof of Lemma F.2.1. We first show the following.

Claim F.2.3. *Consider the setting of Algorithm 6, the notation of Equation (F.5), and assume that the deterministic Condition 9.3.4 holds. Let $L_t = \{x : \tilde{f}_t(x) > C_3\|\mathbf{U}_t\|_F^2\widehat{\lambda}_t/\epsilon\}$. We have that $\mathbb{E}_{X\sim G}[w_t(X)\,\mathbb{I}\{X \in L_t\}] \leq \epsilon$.*

Proof. Let $u^* := \arg\max\{u : \mathbb{E}_{X\sim G}[w_t(X)\,\mathbb{I}\{\tilde{f}_t(X) > u\}] \geq \epsilon\}$ and the set $L_t^* = \{x : \tilde{f}_t(x) > u^*\}$. It suffices to show that $u^* \leq C_3\|\mathbf{U}_t\|_F^2\hat{\lambda}_t/\epsilon$ (because this would mean that $L_t \subseteq L_t^*$).

By Lemma F.2.2, we have that $\mathbb{E}_{X\sim G}[w_t(X)\tilde{f}_t(X)\,\mathbb{I}\{X \in L_t^*\}] \leq (C'/C_2)\lambda_t\|\mathbf{U}_t\|_F^2$. If we define the new weights $w_t'(x) = w_t(x)\,\mathbb{I}\{x \in L_t^*\}$ and use them to normalize the distribution, we get that

$$
\mathop{\mathbb{E}}_{X\sim G_{w_t'}}[\tilde{f}_t(X)] = \frac{1}{\mathbb{E}_{X\sim G}[w_t(X)\,\mathbb{I}\{X \in L_t^*\}]} \mathop{\mathbb{E}}_{X\sim G}[w_t(X)\tilde{f}_t(X)\,\mathbb{I}\{X \in L_t^*\}] \leq (C'/C_2)\|\mathbf{U}_t\|_F^2\frac{\hat{\lambda}_t}{\epsilon} \ ,
$$

where we used that the denominator is ϵ. The fact that $\mathbb{E}_{X\sim G_{w_t'}}[\tilde{f}_t(X)] \leq (C'/C_2)\|\mathbf{U}_t\|_F^2\hat{\lambda}_t/\epsilon$ means that at least one point in L_t^* has $\tilde{f}_t(X) \leq (C'/C_2)\|\mathbf{U}_t\|_F^2\hat{\lambda}_t/\epsilon$, which shows that $u^* \leq (C'/C_2)\|\mathbf{U}_t\|_F^2\hat{\lambda}_t/\epsilon$. Since the algorithm uses the value $C_3 = (C'/C_2)$, the proof is completed. □

Lemma F.2.1 now follows by combining Claim F.2.3 with Lemma F.2.2:

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim G}[w_t(X)\tilde{h}_t(X)] &= \mathop{\mathbb{E}}_{X\sim G}\left[w_t(X)\tilde{f}_t(X)\,\mathbb{I}\left\{\tilde{f}_t(X) > C_3\|\mathbf{U}_t\|_F^2\frac{\hat{\lambda}_t}{\epsilon}\right\}\right] \\ &\le (C'/C_2)\lambda_t\|\mathbf{U}_t\|_F^2 \le 2(C'/C_2)\lambda_t\|\mathbf{M}_t\|_F^2\,,\end{aligned}$$

where the last inequality is due to Item 2 of Condition 9.3.4. Letting $C = 2C'$ we have that $\mathbb{E}_{X\sim G}[w_t(X)\tilde{h}_t(X)] \le C/C_2$ as claimed. As mentioned earlier, the same techniques lead to a similar bound on $\mathbb{E}_{X\sim G}[w_t(X)h_t(X)]$. □

F.2.4 Proof of Claim 9.3.11

Claim 9.3.11. *Let the fraction of outliers be $\epsilon < 1/10$ and a parameter $0 < \tau < 1$. Let the distribution $P = (1-\epsilon)G + \epsilon B$. Let $R > 0, \mu \in \mathbb{R}^d$ be such that $\mathbb{P}_{X\sim G}[\|X-\mu\|_2 > R] \le \epsilon$. There is an estimator $\widehat{\mu}$ on $k = O(\log(1/\tau))$ samples from P such that $\|\widehat{\mu}-\mu\|_2 \le 4R$ with probability at least $1-\tau$. Furthermore, $\widehat{\mu}$ can be computed in time $O(k^2d)$ and memory $O(kd)$.*

We can use the following well-known fact (see, e.g., [DHL19], for a proof):

Fact F.2.4. *There is an algorithm* NaivePrune *with the following guarantees. Let $\epsilon' < 1/2$, and $\tau > 0$. Let $S \subset \mathbb{R}^d$ be a set of n points so that there exists a $\mu \in \mathbb{R}^d$ and a subset $S' \subseteq S$ so that $|S'| \ge (1-\epsilon')n$, and $\|x-\mu\|_2 \le R$ for all $x \in S'$. Then* $\mathrm{NaivePrune}(S, R, \tau)$ *runs in time $O(nd\log(1/\tau))$, uses memory $O(nd)$, and with probability $1-\tau$ outputs a set of points $T \subset S$ so that $S' \subseteq T$, and $\|x-\mu\|_2 \le 4R$ for all $x \in T$.*

Proof of Claim 9.3.11. The estimator draws a set $S = \{X_1, \ldots, X_k\}$ of k samples from the distribution P. Letting $Y_i := \mathbb{I}\{\|X_i - \mu\|_2 > R\}$ we have that $\mathbb{E}[Y_i] \le 2\epsilon \le 1/5$. Thus, using Hoeffding bound,

$$\mathbb{P}\left[\frac{1}{k}\sum_{i=1}^k Y_i > 1/4\right] \le \mathbb{P}\left[\frac{1}{k}\sum_{i=1}^k Y_i < \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^k Y_i\right] + 0.05\right] \le 2e^{-2k(0.05)^2}\,.$$

Choosing $k = 200\log(2/\tau)$ makes this probability at most τ. Conditioning on that event, the fraction of points outside the ball is at most $\epsilon' := 1/4$, thus running $\mathrm{NaivePruning}(S, R, \tau)$ algorithm of Fact F.2.4 and outputting any point from the returned set yields the desired estimator.

□

F.2.5 Omitted Proofs from Section 9.3.4

Claim 9.3.13. *Under Condition 9.3.4, Algorithm 6 maintains the following invariant:* $\mathbb{E}_{X\sim G}[w_t(X)] \geq 1-3\epsilon$. *In particular, if* $\epsilon \leq 1/8$, *then* $d_{\mathrm{TV}}(P_t, P) \leq 9\epsilon$.

Proof. For every iteration, we denote by $\Delta w_t = w_t - w_{t+1}$, that is for every point $x \in \mathbb{R}^d$, $\Delta w_t(x) = w_t(x) - w_{t+1}(x)$ is the difference between the weights for the two consecutive iterations t and $t+1$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{X\sim G}[w_t(X)] &= \mathop{\mathbb{E}}_{X\sim G}[w_1(X)] - \sum_{i=1}^{t-1} \mathop{\mathbb{E}}_{X\sim G}[\Delta w_i(X)] \\
&\geq 1-\epsilon - \sum_{i=1}^{t-1} \mathop{\mathbb{E}}_{X\sim G}[\Delta w_i(X)] && \text{(Claim 9.3.12)} \\
&\geq 1-\epsilon - \frac{\epsilon}{1-\epsilon} \sum_{i=1}^{t-1} \mathop{\mathbb{E}}_{X\sim B}[\Delta w_i(X)] && \text{(Lemma 9.3.6)} \\
&\geq 1-\epsilon - \frac{\epsilon}{1-\epsilon} \left(\mathop{\mathbb{E}}_{X\sim B}[w_1(X)] - \mathop{\mathbb{E}}_{X\sim B}[w_t(X)] \right) \\
&\geq 1-\epsilon - \frac{\epsilon}{1-\epsilon} \\
&\geq 1-3\epsilon \,,
\end{aligned}$$

where the last line uses that $\epsilon \leq 1/2$. The proof of the second conclusion follows from Claim F.2.5 (stated below). □

Claim F.2.5. *Let* $\epsilon \leq 1/8$. *If* $\mathbb{E}_{X\sim G}[w_t(X)] \geq 1-3\epsilon$, *then* $d_{\mathrm{TV}}(P_t, P) \leq 9\epsilon$.

Although we work with discrete distributions (the empirical distributions on the samples) in Section 9.3, we prove the claim for continuous distributions because it will be useful in Section 9.4.

Proof. By definition $P_t(x) = w_t(x)P(x)/\mathbb{E}_{X\sim P}[w_t(X)]$. Letting $L := \int_{\mathbb{R}^d} w_t(x)P(x)\mathrm{d}x = \mathbb{E}_{X\sim P}[w_t(x)]$, we have that

$$\int_{\mathbb{R}^d} |P_t(x) - P(x)|\mathrm{d}x = (1-\epsilon)\int_{\mathbb{R}^d} G(x)\left|\frac{w_t(x)-L}{L}\right|\mathrm{d}x + \epsilon\int_{\mathbb{R}^d} B(x)\left|\frac{w_t(x)-L}{L}\right|\mathrm{d}x \,.$$

We note that $1 \geq L \geq (1-\epsilon)\,\mathbb{E}_{X\sim G}[w_t(x)] \geq 1-4\epsilon$ using Claim 9.3.13. The second term can be bounded as

$$\epsilon\int_{\mathbb{R}^d} B(x)\left|\frac{w_t(x)-L}{L}\right|\mathrm{d}x \leq \frac{\epsilon}{L}\int_{\mathbb{R}^d} B(x)(w_t(x)+L) \leq \frac{2\epsilon}{1-4\epsilon} \leq 4\epsilon \,.$$

For the first term, we have that

$$(1-\epsilon)\int_{\mathbb{R}^d} G(x)\left|\frac{w_t(x)-L}{L}\right|\mathrm{d}x \le \frac{1-\epsilon}{L}\int_{\mathbb{R}^d} G(x)(|1-w_t(x)|+|1-L|)\mathrm{d}x$$
$$\le \frac{1}{1-4\epsilon}\left(1-\mathop{\mathbb{E}}_{X\sim G}[w_t(X)]+4\epsilon\right) \le 14\epsilon\,.$$

□

Claim 9.3.14. *Under Condition 9.3.4, if $C_1 \ge 22$, $\mathbf{B}_t \succeq (0.5C_1\delta^2/\epsilon)\mathbf{I}_d$ for every $t \in [K]$.*

Proof. Using the simple fact that for random variables X, Y it holds $\mathbf{Var}(Y) \ge \mathbb{E}_X[\mathbf{Var}(Y|X)]$, we get that

$$\begin{aligned}
\mathbf{\Sigma}_t &\succeq (1-\epsilon)\mathop{\mathbb{E}}_{X\sim G_{w_t}}\left[(X-\mu_{G_{w_t}})(X-\mu_{G_{w_t}})^\top\right]\\
&= (1-\epsilon)\left(\mathop{\mathbb{E}}_{X\sim G_{w_t}}\left[(X-\mu)(X-\mu)^\top\right]-(\mu_{G_{w_t}}-\mu)(\mu_{G_{w_t}}-\mu)^\top\right)\\
&\succeq (1-\epsilon)\left((1-\delta^2/\epsilon)\mathbf{I}_d-\delta^2\mathbf{I}_d\right) \qquad \text{(by stability and Claim 9.3.13)}\\
&\succeq (1-\epsilon)(1-2\delta^2/\epsilon)\mathbf{I}_d\\
&\succeq (1-3\delta^2/\epsilon)\mathbf{I}_d\,,
\end{aligned}$$

where we used that $\epsilon \le \delta$. We recall the definition $\mathbf{B}_t = (\mathbb{E}_{X\sim P_t}[w_t(X)])^2\mathbf{\Sigma}_t - (1-C_1\delta^2/\epsilon)\mathbf{I}_d$ and bound the first term as follows

$$\begin{aligned}
\left(\mathop{\mathbb{E}}_{X\sim P_t}[w_t(X)]\right)^2\mathbf{\Sigma}_t &\succeq (1-3\epsilon)^2(1-3\delta^2/\epsilon)\mathbf{I}_d \qquad \text{(Claim 9.3.13)}\\
&\succeq (1-4\delta^2/\epsilon-6\epsilon-27\delta^2\epsilon)\mathbf{I}_d\\
&\succeq (1-11\delta^2/\epsilon)\mathbf{I}_d\,,
\end{aligned}$$

where the last line uses $\epsilon < 1/6$ and $\epsilon \le \delta$. Therefore, if we choose $C_1 \ge 22$, we get that $\mathbf{B}_t \succeq (0.5C_1\delta^2/\epsilon)\mathbf{I}_d$. □

Claim 9.3.10. *In the setting of Algorithm 6 and under the Condition 9.3.4, if $x \in S$, we have that $\tau_t(x) \le 1.25\tilde{\tau}_t(x) + 3C_3(\lambda_t/\epsilon)\mathrm{tr}(\mathbf{M}_t^2)$, where C_3 is the constant used in Algorithm 6.*

Proof. By Condition 9.3.4 we have that for all the n samples, $\tilde{g}_t(x) \ge 0.8g_t(x)$. Recall the definitions $\tilde{\tau}_t(x) = \tilde{g}(x)\,\mathbb{I}\{\tilde{g}(x) > C_3\|\mathbf{U}_t\|_F^2\hat{\lambda}_t/\epsilon\}$ and $\tau_t(x) = g(x)\,\mathbb{I}\{g(x) > C_3\|\mathbf{M}_t\|_F^2\lambda_t/\epsilon\}$. We split into cases based on whether each of $g_t, \tilde{g}_t$ has been zeroed by its thresholding operation:

- If $\tau_t(x)$ has been zeroed, (i.e., $g_t(x) < C_3\|\mathbf{M}_t\|_F^2\lambda_t/\epsilon$), the claim trivially holds since the left-hand side is zero.
- If none of $\tilde{\tau}_t(x), \tau_t(x)$ has been zeroed, then $\tilde{\tau}_t(x) = \tilde{g}_t(x)$ and $\tau_t(x) = g_t(x)$, thus the claim holds by the aforementioned fact that $\tilde{g}_t(x) \geq 0.8 g_t(x)$.
- If $\tilde{\tau}_t(x)$ has been zeroed but $\tau_t(x)$ has not, then the worst case is $g_t(x) = (1/0.8)\tilde{g}_t(x)$. This means that in this case:

$$\tau_t(x) \leq \frac{1}{0.8} C_3 \frac{\hat{\lambda}_t}{\epsilon} |\mathbf{U}_t\|_F^2 < 3C_3 \frac{\lambda_t}{\epsilon} |\mathbf{M}_t\|_F^2 \,,$$

  where we used that $\hat{\lambda} \leq 1.2\lambda_t$ and $\|\mathbf{U}_t\|_F^2 \leq 1.2\|\mathbf{M}_t\|_F^2$, due to Condition 9.3.4.

□

F.3 Omitted Proofs from Section 9.4

F.3.1 Omitted Proofs from Section 9.4.2

Lemma 9.4.6. *In the context of Algorithm 8, if* $(1-\epsilon)\,\mathbb{E}_{X\sim G}[w(X)\tilde{\tau}(X)] \leq T$, $\|\tilde{\tau}\|_\infty \leq r$, *and* $\ell_{\max} > r/T$, *then Algorithm 9 modifies the weight function* w *to* w' *such that* (i) $(1-\epsilon)\,\mathbb{E}_{X\sim G}[w(X) - w'(X)] < \epsilon\,\mathbb{E}_{X\sim B}[w(X) - w'(X)]$, *and* (ii) *upon termination we have* $\mathbb{E}_{X\sim P}[w'(X)\tilde{\tau}(X)] \leq 54T$. *Furthermore, if the estimator of Line 3 is set to be that of Lemma 9.4.16, the algorithm terminates after* $O(\log(\ell_{\max}))$ *iterations, each of which uses* $O((R^2\epsilon/\delta^2)\log(1/\tau))$ *samples, takes* $O(nd)$ *time and memory* $O(\log(1/\tau))$.

Proof. Let the set $L^* = \{\ell \in [\ell_{\max}] \,:\, 6T \leq \mathbb{E}_{X\sim P}[w(X)(1-\tilde{\tau}(X)/r)^\ell \tilde{\tau}(X)] \leq 18T\}$. The invariant is that throughout Algorithm 9, the set L maintained has non-zero intersection with L^*. This can be seen by examining cases about ℓ in Line 6. If $\ell \in L^*$, then ℓ is kept in L. If $\ell \notin L^*$, then all elements discarded are not members of L^* (for example if $\mathbb{E}_{X\sim P}[w(X)(1-\tilde{\tau}(X)/r)^\ell\tilde{\tau}(X)] > 18T$, then by Lemma 9.4.16 $f(\ell) > 9T$ and we discard the lower half of L). Thus, at the end, L has at most two elements with at least one of them belonging in L^*. This element would satisfy $3T \leq f(\ell) \leq 27T$. Thus the algorithm will definitely return some element. On the other hand, any element returned will satisfy $2T < \mathbb{E}_{X\sim P}[w(X)(1-\tilde{\tau}(X)/r)^\ell\tilde{\tau}(X)] < 54T$. This has already shown part (ii) of the lemma. For part (i), it is more convenient to imagine let ℓ increased by one at each step until it reaches the value finally returned by the algorithm and consider the loss in

weight between that and the next iteration, exactly as in the proof of Lemma 9.3.6. That proof was using only the facts that for all $\ell' \leq \ell$, $2T < \mathbb{E}_{X\sim P}[w(X)(1-\tilde{\tau}(X)/r)^{\ell'}\tilde{\tau}(X)]$ (which we just showed above) and $\mathbb{E}_{X\sim G}[w(X)(1-\tilde{\tau}(X)/r)^{\ell'}\tilde{\tau}(X)] < T$ (which is true by assumption). The reason why $\ell_{\max} = r/T$ suffices is also shown identically to the proof of Lemma 9.3.6. □

F.3.2 Omitted Proofs from Section 9.4.2.1

We now focus on showing Lemma 9.4.9. In order to avoid confusion with the fraction of outliers ϵ, we use ϵ' for our accuracy parameter. We will use a uniform convergence result from [AB99] combined with a powerful VC-dimension bound from [GJ95] for the class of functions that are computable by a small number of arithmetic operations. [GJ95] considers the class of concepts parameterized by k real numbers, $\mathcal{F} = \{h_a\}_{a\in\mathbb{R}^k}$, for which there exists an algorithm $\mathcal{A}$ for calculating $h_a(x)$ that takes as input x, a and each line of $\mathcal{A}$ is one of the following:

- an arithmetic operation $+, -, \times$, and $/$ on two inputs or previously computed values,
- a jump to a different line of the algorithm conditioned on whether an input or a previously calculated value is greater than or equal to zero,
- output zero or one.

The parameters a and the inputs x consist of real numbers, and the model of computation assumed allows for arithmetic operations and the comparisons between reals to be done in constant time. We refer the reader to [GJ95, Section 2] for more details and relation with algebraic decision trees with bounded depths. The result from [GJ95] is that $\mathrm{VCdim}(\mathcal{F}) = O(mk)$ where k is the size of the parameterization and m is the runtime of the algorithm $\mathcal{A}$. Using the bound on the VC dimension, we have the following result for the uniform convergence:

Proposition F.3.1 ([GJ95; AB99])**.** *Let $\mathcal{F}$ be a class of functions of the form $\mathcal{F} = \{h_a : \mathbb{R}^d \to [0,1] \mid a \in \mathbb{R}^k\}$, where for any $(a,x) \in \mathbb{R}^k \times \mathbb{R}^d$, $h_a(x)$ can be computed by an algorithm $\mathcal{A}$ with runtime m that takes as input a, x and is allowed to perform conditional jumps (conditioned on equality and inequality of real values) and execute the standard arithmetic operations on real numbers $(+, -, \times, /)$ in constant time. For any distribution D on $\mathbb{R}^d$ and any $\epsilon' \in (0,1)$, there*

exist $N = O\left(\frac{1}{(\epsilon')^2}(\log(km) + km\log(1/\epsilon'))\right)$ *points* $x_1, \ldots, x_N$ *in* $\mathbb{R}^d$ *such that*

$$\sup_{h\in\mathcal{F}}\left|\mathop{\mathbb{E}}_{X\sim D}[h(X)] - \frac{1}{N}\sum_{i=1}^{N} h(x_i)\right| \leq \epsilon' \,. \tag{F.6}$$

For completeness, we show at the end of this section how this is derived from the statements of [AB99] and [GJ95]. We now apply this result to our case. We need to specify a family $\mathcal{F}$ of functions broad enough to capture every $w_{t+1}\tilde{\tau}_t$ and $w_{t+1}\tau_t$ that could be encountered during the execution of Algorithm 8. The factor r used in the statement below is a normalization factor to make sure that the functions are in $[0,1]$.

Lemma F.3.2. *Consider the setting of Algorithm 8. Let* $r' := (CdR^2 + 1 + C_1\delta^2/\epsilon)^{C\log d}$. *There exists a family* $\mathcal{F}$ *of functions from* $\mathbb{R}^d$ *to* $[0,1]$ *such that:*

1. *For every iteration* t *of Algorithm 8, we have that* $\frac{1}{r'}w_{t+1}\tau_t \in \mathcal{F}$ *and* $\frac{1}{r'}w_{t+1}\tilde{\tau}_t \in \mathcal{F}$.
2. *Functions of* $\mathcal{F}$ *are parameterized by at most* $k = O(dK\max(L,d))$ *real numbers, that is,* $\mathcal{F}$ *has the form* $\mathcal{F} = \{h_a : \mathbb{R}^d \to [0,1] \mid a \in \mathbb{R}^k\}$.
3. *For every* $h_a \in \mathcal{F}$ *and* $x \in \mathbb{R}^d$, $h_a(x)$ *can be in* $m = dK\max(L,d)(dR\epsilon/\delta^2)^{O(\log d)}$ *steps in the model that allows conditional jumps and standard arithmetic operations on real numbers.*

Proof. Let $L_2' := \max(L,d)$. Every function in our family will be parameterized by $2K+1$ scalars, $\{u_t \in \mathbb{R} \;:\; t \in [K+1]\} \cup \{\ell_t \in \mathbb{R} : t \in [K]\}$, and $(K+1)(L_2'+1)$ vectors in $\mathbb{R}^d$, $\{a_t : t \in [K+1]\} \cup \{v_{t,j} : t \in [K+1], j \in [L_2']\}$. For brevity, we denote by $\mathbf{V}$ the tensor in $\mathbb{R}^{(K+1)\times L_2'\times d}$ having all the vectors $\mathbf{V}_{t,j} = v_{t,j}$ and by $\mathbf{A}$ the tensor in $\mathbb{R}^{(K+1)\times d}$ with $\mathbf{A}_t = a_t$, $t \in [K+1]$. Similarly, denote by u the vector $(u_1, \ldots, u_{K+1})$ and let $\ell = (\ell_1, \ldots \ell_K)$. We define our class to be

$$\mathcal{F} = \left\{h_{\ell,u,\mathbf{V},\mathbf{A}} : \mathbb{R}^d \to [0,1] \;:\; u \in \mathbb{R}^{K+1}, \ell \in \mathbb{R}^K, \mathbf{V} \in \mathbb{R}^{(K+1)\times L_2'\times d}, \mathbf{A} \in \mathbb{R}^{(K+1)\times d}\right\} \,,$$

which includes all functions of the form $h_{\ell,u,\mathbf{V},\mathbf{A}}(x) = \tilde{h}_{\ell,u,\mathbf{V},\mathbf{A}}\,\mathbb{I}\{\tilde{h}_{\ell,u,\mathbf{V},\mathbf{A}}(x) \in (0,1)\}$, where

$$\begin{aligned}\tilde{h}_{\ell,u,\mathbf{V},\mathbf{A}}(x) = &\mathbb{I}\left\{\|x-\hat{\mu}\|_2 \leq 5R\right\}\frac{1}{r}\cdot\frac{\sum_{j=1}^{L_2'}(v_{K+1,j}^\top(x-a_{K+1}))^2}{L_2'}\,\mathbb{I}\left\{\frac{\sum_{j=1}^{L_2'}(v_{K+1,j}^\top(x-a_{K+1}))^2}{L_2'} > u_{K+1}\right\}\\ &\cdot\prod_{t=1}^{K}\left(1 - \frac{1}{r}\frac{\sum_{j=1}^{L_2'}(v_{t,j}^\top(x-a_t))^2}{L_2'}\,\mathbb{I}\left\{\frac{\sum_{j=1}^{L_2'}(v_{t,j}^\top(x-a_t)))^2}{L_2'} > u_t\right\}\right)^{\ell_t} \,. \end{aligned}\tag{F.7}$$

We note that the radius $r' := (CdR^2 + 1 + C_1\delta^2/\epsilon)^{C\log d}$ is an upper bound on the value that the functions $w_{t+1}\tau_t$ and $w_{t+1}\tilde{\tau}_t$ in Algorithm 8 can take: For $\tau_t(x)$ we have

$$\tau_t(x) \le \|\mathbf{M}_t(x-\mu_t)\|_2^2 \le \|\mathbf{M}_t\|_2^2\|x-\mu_t\|_2^2 \le \|\mathbf{\Sigma}_t\|_2^{2\log d}R^2 = O(dR^{2+4\log d}) \ ,$$

while for $\tilde{\tau}(x)$ we have the bounds

$$\begin{aligned}
\tilde{\tau}_t(x) &\le \tilde{g}_t(x) \le \|\mathbf{U}_t(x-\mu_t)\|_2^2 \lesssim R^2\|\mathbf{U}_t\|_2^2 \lesssim R^2\|\mathbf{U}_t\|_F^2 \\
&\le R^2\frac{1}{L}\sum_{j=1}^{L}\|\mathbf{M}_t z_{t,j}\|_2^2 \le dR^2\|\mathbf{M}_t\|_2^2 \le dR^2\|\mathbf{B}_t\|_2^{2\log d} \\
&\le dR^2\left(\|\mathbf{\Sigma}_t\|_2 + 1 + C_1\delta^2/\epsilon\right)^{2\log d} \le dR^2\left(CR^2 + 1 + C_1\delta^2/\epsilon\right)^{2\log d} \\
&\le \left(CdR^2 + 1 + C_1\delta^2/\epsilon\right)^{O(\log d)} \ .
\end{aligned}$$

We check that $\mathcal{F}$ can indeed implement the functions $w_{t+1}\tilde{\tau}_t$ used in Algorithm 8 for any $t \in [K]$: Note that the scores $\tilde{g}_t$ used in the algorithm are means of the form $\frac{1}{L}\sum_{j=1}^{L}(v_{t,j}^\top(x-a_t))^2$. Thus, the first line of Equation (F.7) implements $\mathbb{I}\{\|x-\hat{\mu}\|_2 \le 5R\}\frac{1}{r}\tilde{\tau}_t$. The purpose of the second line in Equation (F.7) is to match the operation of the Downweighting filter, which, in the t-th round multiplies w_t with $(1-\tilde{\tau}_t(x)/r)^{\ell_t}$ for some power ℓ_t. Finally, we note that $w_{t+1}\tau_t$ are implemented in $\mathcal{F}$ by taking $v_{t,j}$ to be the rows of the matrix $\mathbf{M}_t$ (this is why we need the sums to be on $L' = \max(L,d)$ terms in Equation (F.7)).

We need to specify the arithmetic complexity m and the dimension of the parameterization k of our family $\mathcal{F}$. For the first, we have that for any $h \in \mathcal{F}$ and $x \in \mathbb{R}^d$, the value $h(x)$ can be computed using $O(KdL'\ell_{\max})$ standard arithmetic operations and jumps, where $\ell_{\max}$ is the maximum exponent that ℓ_t can have and is set to be $\ell_{\max} := \left(\frac{dR}{\delta^2/\epsilon}\right)^{C\log d}$ in Line 22 of Algorithm 8. The dL' comes from the computation of the means $(1/L')\sum_{j=1}^{L'}(v_{t,j}^\top(x-a_t))^2$ and the K comes from the fact that we have K factors in the expression of h.

Regarding the other parameter k, we have that every $h \in \mathcal{F}$ is parameterized by $O(K)$ scalars and $O(KL')$ d-dimensional vectors. Thus, $k = O(KL'd)$. □

We are now ready to prove Lemma 9.4.9.

Lemma 9.4.9. *Consider the setting of Algorithm 8, where B is the distribution of outliers supported in a ball of radius R around μ. Let $r' := (CdR^2 + 1 + C_1\delta^2/\epsilon)^{C\log d}$ for sufficiently large constant C. Denote by ϵ the contamination rate and let an arbitrary $\epsilon' \in (0,1)$. There*

exists a set S_{cover} *of* $N = \frac{1}{\epsilon'^3} d^4 K^2 L^2 (dR\epsilon/\delta^2)^{O(\log d)}$ *points* $x_1, \ldots, x_N$ *lying in the ball of radius* R *around* μ*, such that for all* $t \in [K]$*, for all choices of the vectors* $z_{t,j}$ *of Line 24 of Algorithm 8 it holds*

$$\left| \mathop{\mathbb{E}}_{X \sim B} \left[\frac{1}{r'} w_{t+1}(X) \tilde{\tau}_t(X) \right] - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{r'} w_{t+1}(x_i) \tilde{\tau}_t(x_i) \right| \leq \epsilon'$$

$$\textit{and} \quad \left| \mathop{\mathbb{E}}_{X \sim B} \left[\frac{1}{r'} w_{t+1}(X) \tau_t(X) \right] - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{r'} w_{t+1}(x_i) \tau_t(x_i) \right| \leq \epsilon' \ .$$

Proof of Lemma 9.4.9. We use Proposition F.3.1 for the family $\mathcal{F}$ of Lemma F.3.2 and plug the upper bounds for the arithmetic complexity m and the dimension of the parameters k. Proposition F.3.1 states that N can be chosen to be a multiple of

$$\frac{1}{\epsilon'^2} (\log(km) + km \log(1/\epsilon')) \ .$$

Taking the much looser bound $N = \Theta(\frac{km}{\epsilon'^3})$ suffices for our purposes. Plugging in $k = O(dK \max(L, d))$, $m = dK \max(L, d)(dR\epsilon/\delta^2)^{O(\log d)}$ from Lemma F.3.2, we get $km = d^2 K^2 \max(d^2, L^2)(dR\epsilon/\delta^2)^{O(\log d)} \lesssim d^4 K^2 L^2 (dR\epsilon/\delta^2)^{O(\log d)}$. □

For completeness, we provide the proof of Proposition F.3.1.

Proof of Proposition F.3.1. We derive the result from the statements of [AB99] without explaining all of the definitions of the notions involved. Please see [AB99] for more details. Applying Theorem 17.7 [AB99] with the loss function $\ell_h(x, y) = h(x)$ we obtain

$$\mathbb{P} \left[\sup_{h \in \mathcal{F}} \left| \mathop{\mathbb{E}}_{X \sim D} [h(X)] - \frac{1}{N} \sum_{i=1}^{N} h(X_i) \right| > \epsilon' \right] \leq 4 \mathcal{N}_1(\epsilon'/8, \mathcal{F}, 2N) \exp(-\epsilon'^2 N/32) \ , \tag{F.8}$$

where the probability is taken over a set of N i.i.d. points $X_1, \ldots, X_N$ drawn from D.

To bound from above the covering number $\mathcal{N}_1(\epsilon'/8, \mathcal{F}, 2N)$, we use Theorem 18.4 from [AB99] which gives that $\mathcal{N}_1(\epsilon'/8, \mathcal{F}, 2N) \leq e(d' + 1)(16e/\epsilon')^{d'}$ where $d' = \mathrm{Pdim}(\mathcal{F})$ is the pseudo-dimension of $\mathcal{F}$. From that, we conclude that choosing any

$$N > \frac{32}{\epsilon'^2} \log(4e(d' + 1)) + \frac{32 d'}{\epsilon'^2} \log(16e/\epsilon')$$

makes the probability in Equation (F.8) less than 1.

It remains to bound d' from above, which can be done as follows. Define the *subgraph*

class associated to the family $\mathcal{F}$

$$\mathcal{B}_{\mathcal{F}} := \{B_h \mid h \in \mathcal{F}\} \ ,$$

where for any $h \in \mathcal{F}$, $B_h : \mathbb{R}^{d+1} \to \{0,1\}$ is defined as $B_h(x,y) = \mathbb{I}\{h(x) \geq y\}$. The pseudo-dimension is defined to be $\mathrm{Pdim}(\mathcal{F}) = \mathrm{VCdim}(\mathcal{B}_{\mathcal{F}})$ (see Section 11.2 in [AB99]). By Theorem 2.3 in [GJ95], we have that $\mathrm{VCdim}(\mathcal{B}_{\mathcal{F}}) = O(km)$ since it $\mathcal{B}_{\mathcal{F}}$ functions that are parametrized by vectors of $\mathbb{R}^k$ (same as for family $\mathcal{F}$) and the functions of $B_h(x,y)$ can be computed using at most $m+2$ operations (m to compute h and two to do the comparison with y and threshold). Putting everything together, it suffices to choose

$$N = C\frac{1}{\epsilon'^2}(\log(km) + km\log(1/\epsilon'))$$

in order to make the probability in Equation (F.8) less than 1. In that case, by probabilistic argument, there exists at least one set of N points satisfying the desired event. □

Claim 9.4.10. *Let S be the cover of Lemma 9.4.9 with r' and ϵ' as defined above. Suppose that the deterministic condition Condition 9.4.5 holds. If $x \in S_{cover}$, then $\tau_t(x) \leq 5\tilde{\tau}_t(x) + (18C_3 + 12/C_2)(\lambda_t/\epsilon)\|\mathbf{M}_t\|_F^2$, where C_3 and C_2 are the constants used in Algorithm 8.*

Proof. By Condition 9.4.5 we have that for all the N samples of the cover, $\tilde{g}_t(x) \geq 0.2g_t(x) - 0.8(\delta^2/\epsilon^2)\|\mathbf{M}_t\|_F^2$. Recall the definitions

$$\tilde{\tau}_t(x) = \tilde{g}(x)\,\mathbb{I}\{\tilde{g}(x) > C_3\|\mathbf{U}_t\|_F^2\hat{\lambda}_t/\epsilon\}, \tau_t(x) = g(x)\,\mathbb{I}\{g(x) > C_3\|\mathbf{M}_t\|_F^2\lambda_t/\epsilon\}\,.$$

We split into cases based on whether each of $g_t, \tilde{g}_t$ has been zeroed by their thresholding operation:

- If $\tau_t(x)$ has been zeroed, (i.e., $g_t(x) < C_3\|\mathbf{U}_t\|_F^2\lambda_t/\epsilon$), the claim trivially holds since the left-hand side is zero.
- If none of $\tilde{\tau}_t(x), \tau_t(x)$ has been zeroed, then $\tilde{\tau}_t(x) = \tilde{g}_t(x)$ and $\tau_t(x) = g_t(x)$, thus the claim holds by the aforementioned fact that $\tilde{g}_t(x) \geq 0.2g_t(x) - 0.8(\delta^2/\epsilon^2)\|\mathbf{M}_t\|_F^2$.
- If $\tilde{\tau}_t(x)$ has been zeroed but $\tau_t(x)$ has not, then the worst case is $\tilde{g}_t(x) = 0.2g_t(x) - 0.8(\delta^2/\epsilon^2)\|\mathbf{M}_t\|_F^2$. This means that in this case:

$$\tau_t(x) \leq \frac{1}{0.2}C_3\frac{\hat{\lambda}_t}{\epsilon}\|\mathbf{U}_t\|_F^2 + 4\frac{\delta^2}{\epsilon^2}\|\mathbf{M}_t\|_F^2$$

$$
\begin{aligned}
&< 18C_3\frac{\lambda_t}{\epsilon}\|\mathbf{M}_t\|_F^2 + 4\frac{\delta^2}{\epsilon^2}\|\mathbf{M}_t\|_F^2 \\
&\leq 18C_3\frac{\lambda_t}{\epsilon}\|\mathbf{M}_t\|_F^2 + \frac{12}{C_2}\frac{\lambda_t}{\epsilon}\|\mathbf{M}_t\|_F^2 \ ,
\end{aligned}
$$

where in the second inequality we used that $\hat{\lambda}_t \leq 3\lambda_t$ and $\|\mathbf{U}_t\|_F^2 \leq 1.2\|\mathbf{M}_t\|_F^2$ due to Condition 9.4.5, and in the last inequality we used that $\delta^2/\epsilon < \hat{\lambda}_t/C_2$ and $\hat{\lambda}_t \leq 3\lambda_t$ (Condition 9.4.5 again).

□

Remark F.3.3 (On the choice of K and L)**.** *We comment on how the values for the number of iterations K and L that are used in Algorithm 8 are derived. First, the derivation of $K = C\log d\log(dR/(\delta^2/\epsilon))$ for large enough constant C is identical to that of Section 9.3.4 (see Equation (9.3)). We will thus focus on L. We note that in the proof of Lemma 9.4.7 we use Lemma 9.4.9 with $\epsilon' \gtrsim \frac{(\delta^2/\epsilon)^{2\log d}}{\epsilon(CdR^2+1+C_1\delta^2/\epsilon)^{C\log d}}$. This means that the cover S_{cover} of that lemma has size bounded by*

$$
|S_{cover}| \leq \frac{1}{\epsilon'^3}d^4K^2L^2\left(\frac{dR}{\delta^2/\epsilon}\right)^{O(\log d)} \lesssim \frac{(CdR^2+1+C_1\delta^2/\epsilon)^{O(\log d)}}{(\delta^2/\epsilon)^{O(\log d)}}L^2 \ .
$$

The analog of Lemma 9.3.5 thus requires that L is multiple of $\log\left(\frac{|S_{cover}|+d}{\tau}\right)$, where τ is the desired probability of failure. Note that we have the following (rough) bounds

$$
\begin{aligned}
\log\left(\frac{|S_{cover}|+d}{\tau}\right) &\lesssim \log\left(\frac{(L(CdR^2+1+C_1\delta^2/\epsilon)^{O(\log d)}}{\tau(\delta^2/\epsilon)^{O(\log d)}}\right) \\
&\lesssim \log^2(d)\log(CdR^2+1+C_1\delta^2/\epsilon)\log\left(\frac{1}{\tau\epsilon}\right)\log(L) \ .
\end{aligned}
$$

Thus, we want to choose L such that it holds $L \geq C\log^2(d)\log(CdR^2+1+C_1\delta^2/\epsilon)\log\left(\frac{1}{\tau\epsilon}\right)\log L$. Using the basic fact that for any $a > 0$, $x \geq 2a\log a \Rightarrow x \geq a\log x$ with $a = C\log^2(d)\log(CdR^2+1+C_1\delta^2/\epsilon)\log\left(\frac{1}{\tau\epsilon}\right)$, it suffices to choose any L satisfying the following

$$
\begin{aligned}
L \geq C\log^2(d)\log&\left(CdR^2+1+C_1\frac{\delta^2}{\epsilon}\right)\log\left(\frac{1}{\tau\epsilon}\right) \\
&\times\log\left(\log^2(d)\log\left(CdR^2+1+C_1\frac{\delta^2}{\epsilon}\right)\log\left(\frac{1}{\tau\epsilon}\right)\right) \ .
\end{aligned}
$$

We see that the choice in Algorithm 8 satisfies this condition.

F.3.3 Omitted Proofs from Section 9.4.3

Lemma 9.4.14. *Let $\mathbf{A}, \mathbf{B}, \mathbf{B}_1, \ldots, \mathbf{B}_p$ be symmetric $d \times d$ matrices and define $\mathbf{M} = \mathbf{B}^p, \mathbf{M}_S = \prod_{i=1}^p \mathbf{B}_i$. If $\|\mathbf{B}_i - \mathbf{B}\|_2 \le \delta\|\mathbf{B}\|_2$, then $\|\mathbf{M}_S - \mathbf{B}^p\|_2 \le p\delta(1+\delta)^p\|\mathbf{B}\|_2^p$.*

Proof. We have the following:

$$\begin{aligned}\mathbf{B}^p - \prod_{i=1}^p \mathbf{B}_i &= \sum_{i=0}^{p-1}\left(\left(\prod_{j=1}^{i}\mathbf{B}_j\right)\mathbf{B}^{p-i} - \left(\prod_{j=1}^{i+1}\mathbf{B}_j\right)\mathbf{B}^{p-i-1}\right)\\ &= \sum_{i=0}^{p-1}\left(\left(\prod_{j=1}^{i}\mathbf{B}_j\right)(\mathbf{B}-\mathbf{B}_{i+1})\,\mathbf{B}^{p-i-1}\right)\ .\end{aligned}$$

Using that $\|\mathbf{B}_j\|_2 \le (1+\delta)\|\mathbf{B}\|_2$, we obtain the following bound:

$$\begin{aligned}\left\|\mathbf{B}^p - \prod_{i=1}^p \mathbf{B}_i\right\|_2 &\le \sum_{i=0}^{p-1}\left\|\left(\prod_{j=1}^{i}\mathbf{B}_j\right)(\mathbf{B}-\mathbf{B}_{i+1})\,\mathbf{B}^{p-i-1}\right\|_2\\ &\le \sum_{i=0}^{p-1}\left(\left(\prod_{j=1}^{i}\|\mathbf{B}_j\|\right)\|\mathbf{B}-\mathbf{B}_{i+1}\|_2\|\mathbf{B}\|_2^{p-i-1}\right)\\ &\le \sum_{i=0}^{p-1}\|\mathbf{B}\|_2^p(1+\delta)^i\delta \le p\delta(1+\delta)^p\|\mathbf{B}\|_2^p.\end{aligned}$$

□

F.3.4 Omitted Proofs from Section 9.4.3.1

Lemma F.3.4. *For any $\delta, \tau \in (0,1)$ and any distribution D on $\mathbb{R}^d$ with mean μ and covariance matrix $\mathbf{\Sigma}$, there exists an estimator $\widehat{\mu}$ on $n = O\left((\mathrm{tr}(\mathbf{\Sigma})/\delta^2)\log(1/\tau)\right)$ i.i.d. samples from D, such that $\|\widehat{\mu} - \mu\|_2 = O(\delta)$. Moreover, this $\widehat{\mu}$ can be computed in time $O(nd\log(1/\tau))$ and using memory $O(d\log(1/\tau))$.*

Proof. Let $X_1, \ldots, X_m$ be independent samples from D. We first show that the empirical mean $Y := (1/m)\sum_{i=1}^m X_i$, is δ-accurate with constant probability.

$$\mathbb{E}[\|Y-\mu\|_2^2] = \sum_{j=1}^d \mathbb{E}[(Y_j-\mu_j)^2] = \frac{1}{m}\sum_{j=1}^d \mathbf{\Sigma}_{jj} = \frac{\mathrm{tr}(\mathbf{\Sigma})}{m}\ .$$

By Markov's inequality, we get that

$$\mathbb{P}[\|Y-\mu\|_2^2 > \delta^2] \le \frac{\mathrm{tr}(\mathbf{\Sigma})}{m\delta^2} \le \frac{1}{20}\ , \tag{F.9}$$

where the last inequality is true if we choose $m = 20\mathrm{tr}(\mathbf{\Sigma})/\delta^2$. Having Equation (F.9) at hand, the probability of success of the above estimator can be boosted to $1-\tau$ by using Claim 9.3.11. We use that claim with G being the distribution of Y, $B = G$, $\epsilon = 1/20$ and $R = \delta$. This completes the proof. □

As a corollary of the above, we obtain the estimators $\widehat{\mu}_t$ of Algorithm 8.

Lemma 9.4.11. *In the setting of Algorithm 8, there exist estimators $\widehat{\mu}_t$ such that, with probability at least $1-\tau$, for all $t \in [K]$ we have that $\|\widehat{\mu}_t - \mu_t\|_2 \le \delta/100$. Furthermore, each $\widehat{\mu}_t$ can be computed on a stream of $n = O\left(\frac{R^2}{\delta^2/\epsilon}\log(K/\tau) + \frac{d(1+\delta^2/\epsilon)}{\delta^2}\log(K/\tau)\right)$ independent samples from P_t, in time $O(nd\log(K/\tau))$ and using memory $O(d\log(K/\tau))$.*

Proof. We use the estimator of Lemma F.3.4 with τ/K in place of τ. It remains to bound $\mathrm{tr}(\mathbf{\Sigma}_t)$. We have that $d_{\mathrm{TV}}(P_t, G) = 1 - O(\epsilon)$, thus, by Fact 9.2.4 we can write $P_t = (1-\alpha)G_0 + \alpha B$, with $\alpha = O(\epsilon)$ and $G_0(x) = h(x)G(x)/(\int h(x)G(x)\mathrm{d}x)$ some weighted version of the inlier's distribution with $\mathbb{E}_{X\sim G}[h(X)] = 1-\alpha$ (same argument that we have used before in the proof of Lemma 9.2.11). We have that

$$\mathbf{\Sigma}_t = (1-\alpha)\mathbf{\Sigma}_{G_0} + \alpha\mathbf{\Sigma}_B + \alpha(1-\alpha)(\mu_{G_0} - \mu_B)(\mu_{G_0} - \mu_B)^\top .$$

Due to stability, the first term has $\mathbf{\Sigma}_{G_0} \preceq (1+\delta^2/\epsilon)\mathbf{I}_d$. For the second term we use that

$$\mathrm{tr}(\mathbf{\Sigma}_B) = \mathop{\mathbb{E}}_{X\sim B}[\mathrm{tr}((X-\mu_B)(X-\mu_B)^\top)] = \mathop{\mathbb{E}}_{X\sim B}[\|X-\mu_B\|_2^2] \le O(R^2) .$$

We also bound the trace of the last term by $O(\epsilon R^2)$. Therefore, we obtain that $\mathrm{tr}(\mathbf{\Sigma}_t) \lesssim d(1+\delta^2/\epsilon) + \epsilon R^2$.

□

F.4 Adaptive Choice of Upper Bound on Covariance

In this section, we show that a simple procedure can be used to make the algorithm adaptive to the scale of covariance (such a procedure is useful for some of our applications in Section 9.5).

As noted earlier, the definition of stability that we have used so far (Definitions 9.2.8 and 9.2.9) was designed for distributions with covariance matrix comparable to $\mathbf{I}_d$. In particular, if inliers satisfy $\mathrm{Cov}[X] \preceq \mathbf{I}_d$, then our algorithms result in error $O(\sqrt{\epsilon})$. In many practical cases, some of which are encountered in Section 9.5, the inliers are

much better concentrated, satisfying $\mathrm{Cov}[X] \preceq \sigma \mathbf{I}_d$, with σ much smaller than 1. In that case, the optimal asymptotic error is $\Theta(\sigma\sqrt{\epsilon})$). If σ is known beforehand, then a simple preprocessing step allows our algorithms to obtain the error $O(\sigma\sqrt{\epsilon})$. We now describe a procedure using Lepski's method [Lep91; Bir01a] that can adapt to the setting when σ is unknown. Concretely, we consider the task of robustly estimating the mean μ of a distribution where inliers have bounded covariance, $\mathrm{Cov}[X] \preceq \sigma^2 \mathbf{I}_d$, but σ is unknown to the algorithm.

Let $\mathrm{RobustMean}(\tilde{\sigma}, \gamma)$ be any black-box robust mean estimation algorithm, where $\tilde{\sigma}$ is a guess for an upper bound on the covariance of inliers (ideally, we would like to use $\tilde{\sigma} = \sigma$) and γ is the probability of failure. The procedure below tries different values for $\tilde{\sigma}$ in order to find a vector that is as good as the output of $\mathrm{RobustMean}$ when run with the best choice of $\tilde{\sigma} = \sigma$. The assumption made here is that σ belongs in some known interval $[A, B]$.

As a small note, a more explicit notation would be $\mathrm{RobustMean}(S, \tilde{\sigma}, \gamma)$, where S is the dataset used, but we omit S because this depends on the data-access model: If a streaming model is assumed, then S necessarily has to be different in each call of the algorithm, otherwise there is no need for using different datasets.

Algorithm 17 Adaptive search for σ

input: $A, B, \gamma, r(\cdot)$
Denote $\tilde{\sigma}_j := B/2^j$ for $j = 0, 1, \ldots, \log(B/A)$ and set $\gamma' := \gamma/\log(B/A)$.
$J \leftarrow 0$
$\widehat{\mu}^{(0)} \leftarrow \mathrm{RobustMean}(\tilde{\sigma}_0, \gamma')$
while $\tilde{\sigma}_j \geq A$ and $\|\widehat{\mu}^{(J)} - \widehat{\mu}^{(j)}\|_2 \leq r(\tilde{\sigma}_J) + r(\tilde{\sigma}_j)$ for all $j = 0, 1, \ldots, J-1$ **do**
    $J \leftarrow J + 1$.
    $\widehat{\mu}^{(J)} \leftarrow \mathrm{RobustMean}(\tilde{\sigma}_J, \gamma')$
end while
$\widehat{J} \leftarrow J - 1$
return $\widehat{\mu}^{(\widehat{J})}$

Theorem F.4.1. *Let $\mu \in \mathbb{R}^d$ $A, B > 0$, $\sigma \in [A, B]$, and a non-decreasing function $r : \mathbb{R}^+ \to \mathbb{R}^+$. Suppose that $\mathrm{RobustMean}(\tilde{\sigma}, \gamma)$ is a black-box algorithm which is guaranteed to return a vector $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 \leq r(\tilde{\sigma})$ with probability $1 - \gamma$, whenever $\tilde{\sigma} \geq \sigma$. Then, Algorithm 17 returns $\widehat{\mu}^{(\widehat{J})}$ such that, with probability at least $1 - \gamma$, we have that $\|\widehat{\mu}^{(\widehat{J})} - \mu\|_2 \leq 3r(2\sigma)$. Moreover, Algorithm 17 calls $\mathrm{RobustMean}$ $O(\log(B/A))$ times with desired failure probability set to $\gamma/\log(B/A)$ and using at most $O(d \log(B/A))$ additional memory.*

Proof. For $j = 0, 1, \ldots, \log(B/A)$, denote by $\mathcal{E}_j$ the event that $\|\widehat{\mu}^{(j)} - \mu\|_2 \le r(\tilde{\sigma}_j)$. Let J be the index corresponding to the value of the unknown parameter σ, i.e., $\tilde{\sigma}_{J+1} \le \sigma \le \tilde{\sigma}_J$. Conditioned on the event $\cap_{j=0}^{J} \mathcal{E}_j$, we have that $\|\widehat{\mu}^{(j)} - \mu\|_2 \le r(\tilde{\sigma}_j)$ for all $j = 0, 1, \ldots, J$. Using the triangle inequality, this gives that $\|\widehat{\mu}^{(J)} - \widehat{\mu}^{(j)}\|_2 \le r(\tilde{\sigma}_J) + r(\tilde{\sigma}_j)$. This means that the stopping condition of Line 5 is satisfied on round J and thus, if $\widehat{\mu}^{(\widehat{J})}$ denotes the vector returned by the algorithm, we have that

$$\|\widehat{\mu}^{(\widehat{J})} - \widehat{\mu}^{(J)}\|_2 \le r(\tilde{\sigma}_{\widehat{J}}) + r(\tilde{\sigma}_J) \le 2r(\tilde{\sigma}_J) \le 2r(2\sigma) \ ,$$

where the first inequality uses that the condition of Line 5, the second uses that r is non-decreasing and $\tilde{\sigma}_{\widehat{J}} \le \tilde{\sigma}_J$, and the last one uses that J was defined to be such that $\tilde{\sigma}_{J+1} \le \sigma \le \tilde{\sigma}_J$ so multiplying σ by 2 makes it greater than $\tilde{\sigma}_J$. Using the triangle inequality once more, we get $\|\widehat{\mu}^{(\widehat{J})} - \mu\|_2 \le 3r(2\sigma)$. Finally, by union bound on the events $\mathcal{E}_j$, the probability of error is upper bounded by $\sum_{j=0}^{J} \gamma' \le \gamma$. The additional memory requirement of this algorithm is to store $\{\widehat{\mu}_j : j \in \{0, \ldots, \log(B/A)\}\}$. □

We now state the implications that Theorem F.4.1 has for Algorithms 6 and 8, given in Sections 9.3 and 9.4:

Corollary F.4.2. *Let $A, B > 0$. In the setting of Corollary 9.4.3, let $\sigma > 0$ be such that the scaled version $S' = \{x/\sigma \ : \ x \in S\}$ of the dataset S is $(C\epsilon, \delta)$-stable with respect to μ/σ. Assuming that $\sigma \in [A, B]$, there exists an algorithm that given $S, \epsilon, \delta, \tau, A, B$ (but not σ), accesses each point of S at most $\operatorname{polylog}(d, 1/\epsilon, 1/\tau, B/A)$ times, runs in time $nd \operatorname{polylog}(d, 1/\epsilon, 1/\tau, B/A)$, uses additional memory $d \operatorname{polylog}(d, 1/\epsilon, 1/\tau, B/A)$, and outputs a vector $\widehat{\mu}$ such that, with probability at least $1 - \tau$, it holds $\|\mu - \widehat{\mu}\|_2 = O(\sigma\delta)$.*

Proof. In order to use the search method of Algorithm 17, we define the procedure $\mathrm{RobustMean}(\tilde{\sigma}, \gamma)$ to be the following:

- Let $\tilde{S} = \{x/\tilde{\sigma} \ : \ x \in S\}$.
- Let $\tilde{\mu}$ be the vector found by the estimator of Corollary 9.4.3 on $\tilde{S}$ using γ for the desired probability of failure.
- Return $\tilde{\sigma}\tilde{\mu}$.

Theorem F.4.1 with $r(\tilde{\sigma}) = C'\sigma\delta$, for a sufficiently large $C' > 0$, implies the correctness. In terms of resources used, Algorithm 17 calls the robust mean estimation algorithm at most $\log(B/A)$ times, and thus the running time gets multiplied by $\log(B/A)$. We also need to store one vector for each call, thus $d \log(B/A)$ additional memory suffices. □

Corollary F.4.3. *Let $A, B > 0$. In the setting of Theorem 9.4.2, let $\sigma > 0$ be such that the distribution D' of the points X/σ, $X \sim D$ is $(C\epsilon, \delta)$-stable with respect to μ. Assuming that $\sigma \in [A, B]$, there exists an algorithm that given*

$$n = O\left(R^2 \max\left(d, \frac{\epsilon}{\delta^2}, \frac{(1+\delta^2/\epsilon)d}{\delta^2 R^2}, \frac{\epsilon^2 d}{\delta^4}, \frac{R^2\epsilon^2}{\delta^2}, \frac{R^2\epsilon^4}{\delta^6}\right) \operatorname{polylog}\left(d, \frac{1}{\epsilon}, \frac{1}{\tau}, R, \frac{B}{A}\right)\right) \quad \text{(F.10)}$$

samples in a stream according to the model of Definition 9.1.1, and given the parameters $\epsilon, \delta, \tau, A, B$ (but not σ), runs in time $nd \operatorname{polylog}(d, 1/\epsilon, 1/\tau, R, B/A)$, uses additional memory $d \operatorname{polylog}(d, 1/\epsilon, 1/\tau, R, B/A)$, and returns a vector $\widehat{\mu}$ such that, with probability at least $1 - \tau$, it holds $\|\mu - \widehat{\mu}\|_2 = O(\sigma\delta)$.

Finally, we note that a similar search procedure can be used for designing algorithms that are adaptive to the parameter δ when σ is known. However, we will not need this generalization for our applications.

F.5 Omitted Details from Section 9.5

F.5.1 Proof Sketch of Theorem 9.5.3

We describe how Algorithm 8 can be plugged in the algorithm of [CDGW19]. We outline the analysis and describe in more detail only the parts from [CDGW19] that need to be changed. The algorithm is Algorithm 1 from [CDGW19], which remains unchanged. This uses Algorithm 2 as a subroutine, which we replace by our estimator of Algorithm 8.

Regarding the analysis, the proof in [CDGW19] uses two claims that state correctness of the black-box robust mean estimator: Lemma 3.4 and Lemma 3.5. For our case, Lemma 3.4 is replaced by our Theorem 9.4.2 specialized to bounded covariance distributions (also see part 2 of Theorem 9.1.3 which says that the sample complexity of Algorithm 8 for that case is $\tilde{O}(d^2/\epsilon)$).

Lemma 3.5 in [CDGW19] also holds when using our estimator. We restate this as a claim below and provide a proof:

Claim F.5.1. *Let D be a distribution supported on $\mathbb{R}^d$ with unknown mean μ^* and covariance $\mathbf{\Sigma}$. Let $0 < \gamma < 1$, $0 < \epsilon < \epsilon_0$ for some universal constant ϵ_0 and $\delta = O(\sqrt{\tau\epsilon} + \epsilon\log(1/\epsilon))$ for some $\tau = O(\sqrt{\epsilon})$. Suppose that D has exponentially decaying tails and $\mathbf{\Sigma}$ is close to the identity*

matrix $\|\Sigma - \mathbf{I}_d\|_2 \le \tau$. *Denote* $R := \sqrt{(d/\epsilon)(1+\delta^2/\epsilon)}$. *Algorithm 8 uses*

$$n = \tilde{O}\left(R^2 \max\left(d, \frac{\epsilon}{\delta^2}, \frac{(1+\delta^2/\epsilon)d}{\delta^2 R^2}, \frac{\epsilon^2 d}{\delta^4}, \frac{R^2\epsilon^2}{\delta^2}, \frac{R^2\epsilon^4}{\delta^6}\right)\right) \tag{F.11}$$

samples drawn from D *and outputs a hypothesis vector* $\hat{\mu}$ *such that* $\|\hat{\mu} - \mu^*\|_2 = O(\delta)$, *with probability* $1-\gamma$. *Moreover, this is done in* $nd\,\mathrm{polylog}(d, 1/\epsilon, 1/\gamma)$ *time and* $d\,\mathrm{polylog}(d, 1/\epsilon, 1/\gamma)$ *space.*

Proof. Since D has exponentially decaying tails, we know that D is stable with respect to its mean μ^* and covariance $\Sigma \preceq O(1)\mathbf{I}_d$ with parameter $\delta = O(\epsilon \log(1/\epsilon))$ (this follows from the tails of the distribution and Definition 9.2.8). That is, for any weight function $w : \mathbb{R}^d \to [0,1]$ with $\mathbb{E}_{X\sim D}[w(X)] \ge 1-\epsilon$ we have that

$$\|\mu_{w,D} - \mu\|_2 \le \delta \quad \text{and} \quad \left\|\overline{\Sigma}_{w,D} - \Sigma\right\|_2 \le \frac{\delta^2}{\epsilon}\,.$$

We claim that D is $(\epsilon, O(\sqrt{\tau\epsilon} + \epsilon\log(1/\epsilon)))$-stable in the sense of Definition 9.2.8 (the difference from what written above is that Definition 9.2.8 uses identity matrix in place of Σ). This can be seen by using triangle inequality:

$$\left\|\overline{\Sigma}_{w,D} - \mathbf{I}_d\right\|_2 \le \left\|\overline{\Sigma}_{w,D} - \Sigma\right\|_2 + \|\Sigma - \mathbf{I}_d\|_2 \le \frac{1}{\epsilon}\left(\delta + \sqrt{\epsilon\tau}\right)^2\,. \tag{F.12}$$

The proof is concluded by recalling the guarantee of Algorithm 8 for $(\epsilon, O(\sqrt{\tau\epsilon} + \epsilon\log(1/\epsilon)))$-stable distributions and using Claim 9.3.12 for the value of R. □

We also note that [CDGW19] uses a fast matrix inversion and multiplication procedure for calculating the rotated versions $Y = \hat{\Sigma}_i^{-1/2} X$ of the samples X. In our case, the run-time of our robust mean-estimation procedure exceeds that of these methods, thus we do not need to use them. We can instead use any numerically stable method that has running time up to $\tilde{O}(d^6)$ and approximates the result within error $\mathrm{poly}(\epsilon\kappa/d)$ (see, e.g., [BGKS20]). Finally, since Claim F.5.1 is essentially used for the d^2-dimensional distributions of the points $Y \otimes Y$, we get the d^4 factor in the final sample complexity, as well as the d^2 factors in the time and space complexity.

F.5.2 Omitted Proofs from Section 9.5.2

Corollary 9.5.6. *In the setting of Theorem 9.5.5, suppose that the distribution of gradients satisfies* $\mathrm{Cov}[\nabla f(\theta)] \preceq \sigma^2\mathbf{I}_d$ *with* $\sigma^2 = \alpha^2\|\theta-\theta^*\|_2^2 + \beta^2$ *for all* $\theta \in \Theta$, *where* $\alpha\sqrt{\epsilon} < \tau_\ell$. *Assume*

that the radius of the domain Θ, $r := \max_{\theta\in\Theta} \|\theta\|_2$ is finite. There exists a single-pass streaming algorithm that given $O(T(d^2/\epsilon)\log(1+\alpha r/\beta)\mathrm{polylog}(d, 1/\epsilon, T/\tau, 1+\alpha r/\beta))$ samples, runs in time $Tnd\,\mathrm{polylog}(d, 1/\epsilon, T/\tau, 1+\alpha r/\beta)$, uses memory $d\,\mathrm{polylog}(d, 1/\epsilon, T/\tau, 1+\alpha r/\beta)$, and returns a vector $\widehat{\theta} \in \mathbb{R}^d$ such that $\|\widehat{\theta} - \theta^*\|_2 = O(\sqrt{\epsilon}\beta/(1-\kappa))$ with probability at least $1-\tau$.

Proof. This follows by using the estimator of Corollary F.4.3 in place of $g(\cdot)$ in Algorithm 10. The known bounds for σ, $A \le \sigma \le B$ are $A = \beta$ and $B = 2\alpha r + \beta$, thus $B/A \le 1 + 2\alpha r/\beta$. The distribution of the scaled gradients $\frac{1}{\sigma}\nabla f(\theta)$ is $(C\epsilon, O(\sqrt{\epsilon}))$-stable. For these parameters, n from Equation (F.10) gives $n = (d^2/\epsilon)\mathrm{polylog}(d, 1/\epsilon, \tau', 1+\alpha r/\beta)$, where τ' is the desired probability of failure for each call of the estimator. Setting $\tau' = \tau/T$ ensures that the estimates of all rounds are successful with probability $1-\tau$. Successful estimates of the gradients are within $O(\sigma\delta) = O((\alpha\|\theta-\theta^*\|_2 + \beta)\sqrt{\epsilon})$ from the true one in Euclidean norm, thus in every round we have an $(\sqrt{\epsilon}\alpha, \sqrt{\epsilon}\beta)$-gradient estimation (in the sense of Definition 9.5.4). Finally, Theorem 9.5.5 requires the condition $\alpha\sqrt{\epsilon} < \tau_\ell$. Assuming that this is true, that theorem concludes the proof. □

Theorem 9.5.12 (Robust Logistic Regression; full version of Theorem 9.1.6)**.** *Consider the logistic regression model of Equation (9.26) with the domain Θ of the unknown regressor being the ball of radius r, for some universal constant $r > 0$, and suppose that Assumption 9.5.10 holds. Assume that $0 < \epsilon < \epsilon_0$ for a sufficiently small constant ϵ_0. There is a single-pass streaming algorithm that uses $n = (d^2/\epsilon)\,\mathrm{polylog}\,(d, 1/\epsilon, 1/\tau)$ samples, runs in time $nd\,\mathrm{polylog}(d, 1/\epsilon, 1/\tau)$, uses memory $d\,\mathrm{polylog}(d, 1/\epsilon, 1/\tau)$, and returns a vector $\widehat{\theta} \in \mathbb{R}^d$ such that $\|\widehat{\theta} - \theta^*\|_2 = O(\sqrt{\epsilon})$ with probability at least $1-\tau$.*

Proof. The algorithm is that of Theorem 9.5.5 using the estimator of Algorithm 8 in place of $g(\cdot)$ in Algorithm 10. The distribution of the gradients is $(C\epsilon, O(\sqrt{\epsilon}))$-stable because of Lemma 9.5.11. For these stability parameters, a sufficient number of samples is $(d^2/\epsilon)\,\mathrm{polylog}\,(d, 1/\epsilon, T/\tau)$ (see Equation (9.8) with $\delta = O(\sqrt{\epsilon})$ and $R = O(\sqrt{d})$), where T is the number of iterations over which take a union bound. It thus remains to specify the parameters τ_ℓ, τ_u, k, T.

Using Assumption 9.5.10, we can calculate bounds on the parameters τ_ℓ, τ_u. For τ_ℓ, let v be a unit vector from $\mathbb{R}^d$. Let the event $\mathcal{E}_{v,\theta} := \{(v^\top X)^2 \ge c_1 \text{ and } |\theta^\top X| \le 2rC^2/c_2\}$, where c_1, c_2, C are the constants from Assumption 9.5.10. The probability of the complement of this event is

$$\mathbb{P}[\mathcal{E}_{v,\theta}^c] \le \mathbb{P}[(v^\top X)^2 < c_1] + \mathbb{P}[|\theta^\top X| > 2rC^2/c_2] \le 1 - c_2 + c_2/2 \le 1 - c_2/2\,,$$

where the first term is bounded using the anti-concentration property and the second is bounded using the concentration property along with Markov's inequality. Thus, using the formula of Equation (9.27) for the Hessian, we have that

$$v^\top \nabla^2 \bar{f}(\theta) v \geq \mathbb{P}[\mathcal{E}_{v,\theta}] \mathop{\mathbb{E}}_{X \sim D_x} \left[\frac{e^{\theta^\top X}}{(1+e^{\theta^\top X})^2} (v^\top X)^2 \,\middle|\, \mathcal{E}_{v,\theta} \right] \geq 0.5 c_2 \frac{e^{2rC^2/c_2}}{(1+e^{2rC^2/c_2})^2} c_1 \; .$$

Regarding the upper bound τ_u, using the bounded covariance property we get that $v^\top \nabla^2 \bar{f}(\theta) v \leq C^2 \sup_{a \in \mathbb{R}} e^a/(1+e^a)^2 = C^2/4$. Therefore, we can choose the values

$$\tau_\ell = 0.5 c_2 \frac{e^{2rC^2/c_2}}{(1+e^{2rC^2/c_2})^2} c_1 \quad \text{and} \quad \tau_u = C^2/4 \; ,$$

for Algorithm 10. The guarantees of our mean estimator imply that we have an $(0, O(\sqrt{\epsilon}))$-gradient estimator (in the sense of Definition 9.5.4). Regarding the value of κ, we use Equation (9.22) with $a = 0$. Since that τ_ℓ, τ_u are positive constants, this means that κ is bounded away from 1. Therefore, we have that the factor $1/(1-\kappa)$ appearing in the final error (Equation (9.24)) is $O(1)$ and the number of iterations from Equation (9.23) are upper bounded by $T \lesssim \log_2(\|\theta_0 - \theta^*\|_2/\sqrt{\epsilon}) \lesssim \log(1/\epsilon)$, where we used that the radius of the domain Θ is $r = O(1)$. □

F.6 Bit Complexity of Algorithm 8

Until this point, we have assumed that our algorithms could save real numbers exactly in a single memory cell and perform calculations involving reals in $O(1)$ time. Thus, by saying that Algorithm 8 uses extra memory at most $d \operatorname{polylog}(d, R, 1/\epsilon, 1/\tau)$, we meant that it needs to store only that many real numbers. We now describe how the algorithm would work in the most realistic word RAM model, where finite precision numbers can be stored in registers of predetermined word size and operations like addition, subtraction and multiplication are performed in $O(1)$ time. We now show that the previous bound of $d \operatorname{polylog}(d, R, 1/\epsilon, 1/\tau)$, worsened only by another poly-logarithmic factor, holds for the total number of bits that need to be stored. We begin by clarifying how the input is given to the algorithm.

Definition F.6.1 (Single-Pass Streaming Model with Oracle Access for Real Inputs)**.** *Let S be a fixed set of points in $\mathbb{R}^d$. The elements of S are revealed one at a time to the algorithm as follows: For each point of S that is about to be revealed, the algorithm is allowed to query as*

many bits as it wants from that point with whatever order it wants. The process then continues with the next point in the stream. Each point of S is presented only once to the algorithm in the aforementioned way.

In the reminder of this section, we use the same notation as in Theorem 9.4.2. We assume $R \leq M$ and that $\|\mu\|_2 \leq M$, for some $M = (d/\epsilon)^{\mathrm{polylog}(d/\epsilon)}$ (otherwise, the estimation of the mean with extra memory of the order $d\mathrm{polylog}(d/\epsilon)$ becomes impossible). The modified algorithm for this model is the following: Every input point X is ignored if found to have norm greater than $2M$. Otherwise, it is deterministically rounded to an X' so that their difference $X - X' := \eta(X)$ has norm at most η, for some $\eta \leq O(\min\{\delta, \frac{\delta^2}{\epsilon R}, R\})$ (see below for more on this choice of η). The exact same algorithm as Algorithm 8 is run on these rounded points.

Correctness First, we note the rejection step removes less than an ϵ-fraction of the input, thus the resulting distribution has not changed more than ϵ in total variation distance from the original one. Moreover, the distribution of the rounded points has essentially the same stability property required by our theorem. Concretely, if we choose the rounding error η to be $\eta = O(\min\{\delta, \frac{\delta^2}{\epsilon M}, M\})$, then it can be seen (Lemma F.6.2 below) that the distribution of the rounded points is $(\epsilon, O(\delta))$-stable and $\mathbb{P}_{X'}[\|X' - \mu\|_2 = O(R)] \geq 1 - \epsilon$, which are the only assumptions needed for Algorithm 8 to provide an accurate estimate up to $O(\delta)$ error.

Lemma F.6.2. *Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. Let G be an (ϵ, δ)-stable distribution with respect to some vector $\mu \in \mathbb{R}^d$ and assume G is a distribution such that $\|X - \mu\|_2 \leq M$ almost surely for some $M > 0$. For any deterministic function $\eta : \mathbb{R}^d \to \mathbb{R}^d$ with $\|\eta(x)\|_2 \leq \eta$ for all x in the support of G, if G' denotes the distribution of the points $X' = X + \eta(X)$, where $X \sim G$, then G' is $(\epsilon, O(\delta + \eta + \sqrt{\epsilon\eta M}))$-stable with respect to μ.*

Proof. We check the two conditions for stability. Let a weight function $w : \mathbb{R}^d \to [0, 1]$ with $\mathbb{E}_{X\sim G}[w(X)] \geq 1 - \epsilon$ and let $\delta' = O(\delta + \eta + \sqrt{\epsilon\eta M})$. We have that

$$\begin{aligned}
\|\mu_{w,G'} - \mu\|_2 &\leq \|\mu_{w,G'} - \mu_{w,G}\|_2 + \|\mu_{w,G} - \mu\|_2 \\
&\leq \left\| \int_{\mathbb{R}^d} (x + \eta(x)) \frac{w(x)G(x)}{\mathbb{E}_{X\sim G}[w(X)]} \mathrm{d}x - \mu_{w,G} \right\|_2 + \delta \\
&\leq \left\| \int_{\mathbb{R}^d} \eta(x) \frac{w(x)G(x)}{\mathbb{E}_{X\sim G}[w(X)]} \mathrm{d}x \right\|_2 + \delta \\
&\leq \eta + \delta \leq \delta' \,,
\end{aligned}$$

where first inequality uses the triangle inequality and the second one uses the stability of G. Regarding the second stability condition, we have the following:

$$\begin{aligned}
&\left\|\overline{\mathbf{\Sigma}}_{w,G'}-\mathbf{I}_d\right\|_2 \\
&\quad\le \left\|\overline{\mathbf{\Sigma}}_{w,G'}-\overline{\mathbf{\Sigma}}_{w,G}\right\|_2+\left\|\overline{\mathbf{\Sigma}}_{w,G}-\mathbf{I}_d\right\|_2 \\
&\quad\le \left\|\int_{\mathbb{R}^d}(x-\mu+\eta(x))(x-\mu+\eta(x))^\top\frac{w(x)G(x)}{\mathbb{E}_{X\sim G}[w(X)]}\mathrm{d}x\right\|_2+\frac{\delta^2}{\epsilon} \\
&\quad\le \left\|\int_{\mathbb{R}^d}(x-\mu)\eta(x)^\top\frac{w(x)G(x)}{\mathbb{E}_{X\sim G}[w(X)]}\mathrm{d}x\right\|_2+\left\|\int_{\mathbb{R}^d}(x-\mu)^\top\eta(x)\frac{w(x)G(x)}{\mathbb{E}_{X\sim G}[w(X)]}\mathrm{d}x\right\|_2 \\
&\quad+\left\|\int_{\mathbb{R}^d}\eta(x)\eta(x)^\top\frac{w(x)G(x)}{\mathbb{E}_{X\sim G}[w(X)]}\mathrm{d}x\right\|_2+\frac{\delta^2}{\epsilon} \\
&\quad\le 2M\eta+\eta^2+\frac{\delta^2}{\epsilon}\le\frac{\delta'^2}{\epsilon}\,,
\end{aligned}$$

where we used stability of G, triangle inequality and the bounds $\|x-\mu\|_2\le M$, $\|\eta(x)\|_2\le\eta$. □

Total Bits of Memory Used In order for the differences $X-X':=\eta(X)$ to have $\|\eta(X)\|_2\le\eta$ for all X, it is sufficient to round every coordinate to absolute error $O(\eta/\sqrt{d})$. Recall that by our assumption on the a priori bound on the norm of the true mean and the way that we reject input points of large norm, we know that all points surviving will have norm at most $2M$. Thus, each coordinate of such points can be stored in a word of $O(\log(Md/\eta))$ bits after being rounded to accuracy η. Therefore, each d-dimensional point that the algorithm will need to manipulate can be stored using d registers of size $O(\log(Md/\eta))$. However, we need to show that the results of all intermediate calculations can be calculated in low memory. We show the following result to this end:

Claim F.6.3. *In the context of Theorem 9.4.2, given a stream of d-dimensional points, where each coordinate has bit complexity B, Algorithm 8 can be implemented in a word RAM machine using $d\,\mathrm{polylog}(d,1/\epsilon,1/\tau,R)$ many of registers of size $B\,\mathrm{polylog}(d,1/\epsilon,1/\tau,R)$.*

Proof Sketch. Multiplying two numbers of bit complexity B_1 and B_2 may result in bit complexity B_1+B_2. Adding k numbers of bit complexity B, may make the resulting bit complexity $B+\log(k)$. We need to check that every step of the algorithm performs calculations that cannot cause the bit complexity to grow by more than poly-log factors.

Line 8 performs only comparisons and counting. Regarding Line 26: As pointed out at the beginning of Section 9.4, the vector $v_{t,j}\leftarrow\widehat{\mathbf{M}}_t z_{t,j}$ is calculated by multiplying

$z_{t,j}$ by $\widehat{\mathbf{B}}_{t,k}$ for $k = 1, \ldots, \log d$ iteratively. Consider a single iteration, say the first one. Performing $\widehat{\mathbf{B}}_{t,k} z_{t,j}$ involves calculating $\frac{1}{n}\left(\sum_x xx^\top\right) z_{t,j}$ (see Section 9.4.3.2), which can be done as $\frac{1}{n}\sum_x x(x^\top z_{t,j})$, i.e., calculating the inner products $x^\top z_{t,j}$ first). A single inner product of that form is just a sum of d numbers of bit complexity B with appropriate signs, thus the bit complexity increases only by $O(\log d)$. Finally, multiplying by x and taking the mean over for all of the x's can add only another $O(B + \log(n))$. Since the number of iterations of such calculations is $\log d$, the final result has the claimed bit complexity.

Regarding the Downweighting filter (Algorithm 9), there are a couple of places where the weights w_t are involved in calculations. We note that Algorithm 7 stores only the counts ℓ_t, which fit in registers of size $\log(\ell_{\max}) = \mathrm{polylog}(d, 1/\epsilon, R)$. These counts are used to calculate $w_t(x)$ as $w_t(x) = \prod_{t' \le t}(1 - \tilde{\tau}_{t'}(x)/r)^{\ell_{t'}}$ whenever there is such a need. An exact calculation would require operations of the form x^y, for some $x \in [0, 1]$ and $y \in [\ell_{\max}]$, i.e., exponentiation of a real number. In fact, as we will show later on, instead of calculating $w_t(x)$ with perfect accuracy, it suffices to use an approximate value of $w_t(x) \pm \eta$ for some error $|\eta| < \mathrm{poly}(1/d, 1/R, \epsilon, \tau)^{\log d}$. This will allow us to calculate a good enough approximation in $\mathrm{polylog}(d, R, 1/\epsilon, 1/\tau)$ bits as follows: we can use exponentiation by squaring algorithm for calculating $w_t(x)$'s and round the result in each step to make it fit into our registers. We first explain this in more detail below.

Claim F.6.4. *Let $x \in [0, 1]$, $y \in \mathbb{Z}_+$, and assume that both x, y have bit complexity B. The power x^y can be calculated up to a rounding error of $2^{-\Omega(B)}$ in the word RAM model that uses registers of size $2B$. Furthermore, this can be done in $O(B)$ standard arithmetic operations.*

Proof. We can use exponentiation by squaring: This consists of writing x in binary as $b_k \cdots b_0$ for $k = \log y$ and calculating the sequence $r_{k+1}, \ldots, r_0$ as $r_{k+1} = 1$, $r_i = r_{k+1}^2 x^{b_i}$ for $i = k, \ldots, 0$. We assume every r_i gets rounded to $2B$ bits. Because of the rounding, we incur error 2^{-2B} in each round. However, the error of the previous rounds gets amplified, since the result of that round (true value plus error) gets squared. We consider one such iteration to see how that sequence of errors grows: In the t-th iteration, let res_{t-1} denote the true result (before rounding) from the previous round and η_t the rounding error of that round (i.e., $r_t = \mathrm{res}_t + \eta_t$). Then, we have that

$$\mathrm{res}_t + \eta_t := (\mathrm{res}_{t-1} + \eta_{t-1})^2 + 2^{-2B} \le \mathrm{res}_{t-1}^2 + \eta_{t-1}^2 + 2\eta_{t-1} + 2^{-2B} ,$$

where w.l.o.g. we assume that $\mathrm{res}_t \le 1$ always. Thus, the rounding error grows as $\eta_t \le \eta_{t-1}^2 + 2\eta_{t-1} + 2^{-2B} \le 3\eta_{t-1} + 2^{-2B}$. In the first round, we start with rounding error

of 2^{-2B}. Thus, after $k = \log(y) = B$ rounds, the final error is $\eta_k \leq 2^{-\Omega(B)}$. □

We continue with examining how fine approximations for w_t are needed. First, in Section 9.4.3.2, we use the estimator $\widehat{W}_t = \mathbb{E}_{X\sim\mathcal{U}(S_0)}[w_t(X)]$, which we require to be η-close to $\mathbb{E}_{X\sim P}[w_t(X)]$ (Equation (9.16)) for some $\eta > \mathrm{poly}(1/d, 1/R, \epsilon, \tau)$. Therefore, when calculating w_t, it suffices to round the intermediate results to error η. This would mean using Claim F.6.4 with $B = O(\log((1/d, 1/R, \epsilon, \tau))$.

Second, the weights w_t are also used in evaluating the stopping condition of the Downweighting filter. Line 3 of that filter is implemented using the estimator of Lemma 9.4.16. As it can be seen in Equation (9.19), it suffices to use rounded versions of w_t in $(1/n)\sum_{i=1}^N w_t(X_i)\tilde{\tau}_t(X_i)$, as long as it does not change the result by an additive factor of $c\widehat{\lambda_t}\|\mathbf{U}_t\|_F^2$, for a small constant c. Since $\tilde{\tau}_t(X_i) = O(dR^{2+4\log d})$ (Equation (9.1)) and $\widehat{\lambda_t}\|\mathbf{U}_t\|_F^2 > (\delta^2/\epsilon)^{\Theta(\log d)}$ (otherwise, the algorithm has terminated), we can again round w_t up to error $\mathrm{poly}(1/d, 1/R, \epsilon)^{\log(d)}$. This means that the results of these calculations fit into $\mathrm{polylog}(d, R, 1/\epsilon, R)$ bits.

Finally, there are two places in Algorithm 8 where we need to simulate samples from the weighted distribution P_{w_t}: (i) Line 19, whose implementation is outlined in Section 9.4.3 and (ii) Line 35. We focus on the first one since the argument for the other case is identical. To simulate P_{w_t}, we use rejection sampling, as described at the beginning of Section 9.4.3, with the only difference that we use the rounded versions of the weights w_t. We are thus essentially simulating samples from a slightly different distribution $P_{\widehat{w_t}}$. However, this is close to P_{w_t} in total variation distance, as shown below.

Claim F.6.5. *Let P be a distribution on $\mathbb{R}^d$ and let P_w denote the weighted version of P according to the function $w : \mathbb{R}^d \to [0,1]$, i.e., $P_w(x) = w(x)P(x)/\int_{\mathbb{R}^d} w(x)P(x)\mathrm{d}x$. For any $w, \widehat{w} : \mathbb{R}^d \to [0,1]$ such that $\int_{\mathbb{R}^d} w(x)P(x)\mathrm{d}x \geq 1/2$ and $\sup_{x\in\mathbb{R}^d}|\widehat{w}(x) - w(x)| \leq \xi$ with $\xi \leq 1/8$, it holds that $d_{\mathrm{TV}}(P_{\widehat{w}}, P_w) \leq 8\xi$.*

Proof. First, letting the normalization factors $\widehat{C} := \int_{\mathbb{R}^d}\widehat{w}(x)P(x)\mathrm{d}x$ and $C := \int_{\mathbb{R}^d} w(x)P(x)\mathrm{d}x$, we note that $|C - \widehat{C}| \leq \xi$. Letting $\Delta w(x) := \widehat{w}(x) - w(x)$ and $\Delta C := \widehat{C} - C$, we have that

$$\begin{aligned}
d_{\mathrm{TV}}(P_{\widehat{w}}, P_w) &= \frac{1}{2}\int_{\mathbb{R}^d}|P_{\widehat{w}}(x) - P_w(x)|\,\mathrm{d}x = \frac{1}{2}\int_{\mathbb{R}^d}\left|\frac{w(x)+\Delta w(x)}{C+\Delta C} - \frac{w(x)}{C}\right|P(x)\mathrm{d}x \\
&\leq \frac{1}{2}\int_{\mathbb{R}^d}\left|\frac{Cw(x)+C\Delta w(x) - Cw(x) - \Delta C w(x)}{C^2 + C\Delta C}\right|P(x)\mathrm{d}x \\
&\leq 4\int_{\mathbb{R}^d}|C\Delta w(x) - \Delta C w(x)|\,P(x)\mathrm{d}x \leq 8\xi\,,
\end{aligned}$$

where in the last line, we first use that $C^2 + C\Delta C \geq 1/4 - \xi \geq 1/8$ (since $1/2 \leq C \leq 1$ and $0 \leq \xi \leq 1/8$) and then we use that $|C\Delta w(x) - \Delta C w(x)| \leq |\Delta w(x)| + |\Delta C| \leq 2\xi$. □

As N (more than one) samples are drawn from P_t in the t-th iteration (see Section 9.4.3), we require the joint distribution of these N samples from P_{w_t} and $P_{\hat{w}_t}$ to be within total variation τ (the probability under which the conclusion of Lemma 9.4.13 holds true). This bound on the total variation distance implies that Lemma 9.4.13 continues to hold for $P_{\hat{w}_t}$ with an additional failure probability of τ. To do this, we use Claim F.6.5 with $\xi = \Theta(\tau/N)$, which means that these rounded weights have bit complexity $\Theta(\log \xi) = \text{polylog}(d, R, 1/\epsilon, 1/\tau)$. □

G APPENDIX TO CHAPTER 10

G.1 Additional Details from Section 10.2

We will use the following additional fact that states the dependence on failure probability for simple binary hypothesis testing:

Fact G.1.1 (Failure probability and sample complexity)**.** *Let $\psi : \cup_{n=1}^{\infty} \mathcal{X}^n \to \{p, q\}$ be the optimal likelihood ratio test for two distributions p and q. If $n \gtrsim \frac{\log(1/\delta)}{d_{\mathrm{h}}^2(p,q)}$ for $\delta \leq 0.1$, then the failure probability of the test ψ with n samples is less than δ.*

The fact above follows by the optimality of the likelihood ratio test and a boosting argument using the median.

We now state upper bounds and lower bounds on the sample complexity of Scheffe's test.

Proposition G.1.2 (Sample complexity of Scheffe's test (folklore))**.** *The sample complexity of Scheffe's test is at most $O\left(1/d_{\mathrm{h}}^4(p, q)\right)$. Furthermore, this is tight in the following sense: For any $\rho \in (0, 1)$, there exist p and q such that $n^*(p, q) = O(1/\rho)$, whereas the sample complexity of Scheffe's test is $\Omega\left(1/\rho^2\right)$.*

Proof. We begin by showing the upper bound on sample complexity. Let p and q be the two given distributions and let $\rho = d_{\mathrm{h}}^2(p, q)$. Let $\mathbf{T}$ be the channel corresponding to Scheffe's test. Since Scheffe's test preserves the total variation distance, we have $d_{\mathrm{TV}}(p, q) = d_{\mathrm{TV}}(\mathbf{T}p, \mathbf{T}q)$. By Fact 10.2.2, we have

$$d_{\mathrm{h}}(\mathbf{T}p, \mathbf{T}q) \geq d_{\mathrm{TV}}(\mathbf{T}p, \mathbf{T}q) = d_{\mathrm{TV}}(p, q) \geq 0.5 d_{\mathrm{h}}^2(p, q) \geq 0.5\rho.$$

Thus, by Fact 10.2.4, the sample complexity is at most $O(1/d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)) = O(1/\rho^2) = O(1/d_h^4(p, q))$.

We now turn our attention to the tightness of the upper bound. Without loss of generality, we consider the setting when $\rho \leq 0.01$. Consider the following two distributions on Δ_3: $p = (\rho, 1/2 - 2\rho, 1/2 + \rho)$ and $q = (0, 1/2, 1/2)$. Let $\mathbf{T}$ be the channel corresponding to Scheffe's test. Then we have $\mathbf{T}p = (1/2 + 2\rho, 1/2 - 2\rho)$ and $\mathbf{T}q = (1/2, 1/2)$. An elementary calculation shows that $d_{\mathrm{h}}^2(p, q) = \Theta(\rho)$ and $d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) = \Theta(\rho^2)$. Applying Fact 10.2.4, we obtain the desired conclusion. □

G.2 Reverse Data Processing

In this section, we prove our results from Section 10.3. Appendix G.2.1 contains the proof of the reverse data processing inequality (Theorem 10.3.2). We establish the tightness of the reverse data processing inequality for Hellinger distance (Lemma 10.3.6) in Appendix G.2.2. We state and prove the generalized version of the reverse Markov inequality in Appendix G.2.3. Finally, we establish the tightness of the reverse Markov inequality in Appendix G.2.4.

Fix the distributions p and q over $[k]$. For $0 \le l < u < \infty$, we first define the following sets[38]:

$$A_{l,u} = \left\{ i \in [k] : \frac{p_i}{q_i} \in [l, u) \right\} \text{ and}$$
$$A_{l,\infty} = \left\{ i \in [k] : \frac{p_i}{q_i} \in [l, \infty] \right\}. \tag{G.1}$$

We will use the notation from Definition 10.2.8.

G.2.1 Reverse Data Processing: Proof of Theorem 10.3.2

Theorem 10.3.2 (Reverse data processing inequality)**.** *Let I_f be a well-behaved f-divergence with $(\alpha, \kappa, C_1, C_2)$ as defined in Definition 10.3.1. Let p and q be two fixed distributions over $[k]$ such that for all $i \in [k]$, we have $q_i \ge \nu p_i$ and $p_i \ge \nu q_i$, for some $\nu \in [0, 1]$. Then for any $D \ge 2$, there exists a channel $\mathbf{T}^* \in \mathcal{T}_D^{\texttt{thresh}}$ (and thus in $\mathcal{T}_D$) such that*

$$1 \le \frac{I_f(p, q)}{I_f(\mathbf{T}^* p, \mathbf{T}^* q)} \le 4 \frac{f(\nu)}{f(1/(1+\kappa))} + \frac{52 C_2}{C_1} \max\left\{1, \frac{R}{D}\right\}, \tag{10.6}$$

where $R = \min\{k, k'\}$ and $k' = 1 + \log\left(\frac{4 C_2 \kappa^\alpha}{I_f(p,q)}\right)$. Furthermore, given f, p, and q, there is a $\mathrm{poly}(k, D)$-time algorithm that finds a $\mathbf{T}^*$ achieving the rate in inequality (10.6).

Proof. Let $\kappa > 0$ be as in Definition 10.3.1. By definition of the f-divergence, we have the following:

$$I_f(p, q) = \sum_{i \in A_{1+\kappa,\infty}} q_i f\left(\frac{p_i}{q_i}\right) + \sum_{i \in A_{1,1+\kappa}} q_i f\left(\frac{p_i}{q_i}\right)$$

[38] When $q(x) = 0$ for some x and $p(x) \neq 0$, we think of $p(x)/q(x) = \infty$. Without loss of generality, we can assume that for each $x \in [k]$, at least one of $p(x)$ or $q(x)$ is non-zero.

$$+ \sum_{i \in A_{1/(1+\kappa),1}} q_i f\left(\frac{p_i}{q_i}\right) + \sum_{i \in A_{0,1/(1+\kappa)}} q_i f\left(\frac{p_i}{q_i}\right), \tag{G.2}$$

where the sets $A_{l,u}$ are defined as in equation (G.1). Note that the sets $A_{1+\kappa,\infty}$ and $A_{0,1/(1+\kappa)}$ contain the elements that have a large ratio of probabilities under the two distributions. We now consider two cases.

Case 1: Main contribution by large ratio alphabets: We first consider the case when $\sum_{i \in A_{1+\kappa,\infty} \cup A_{0,1/(1+\kappa)}} q_i f\left(\frac{p_i}{q_i}\right) \geq \frac{I_f(p,q)}{2}$. As we will show later, this is the simple case ($D = 2$ already achieves the claim). By symmetry of the I_f-divergence for the well-behaved f-divergence (I.2 in Definition 10.3.1), it suffices to consider the case when $\sum_{i \in A_{1+\kappa,\infty}} q_i f\left(\frac{p_i}{q_i}\right) \geq \frac{I_f(p,q)}{4}$.[39]

We will show that there exists $\mathbf{T} \in \mathcal{T}_2^{\texttt{thresh}}$ such that $I_f(\mathbf{T}p, \mathbf{T}q) \geq \frac{f(1/(1+\kappa))}{4f(\nu)} I_f(p,q)$. Let $\mathbf{T}$ be the channel corresponding to the threshold $1+\kappa$, i.e., $\mathbf{T}$ corresponds to the function $i \mapsto \mathbb{I}_{A_{1+\kappa,\infty}}(i)$. Note that $p' := \mathbf{T}p$ and $q' := \mathbf{T}q$ are distributions on $\{0,1\}$ with $(\mathbf{T}p)_1 = \sum_{i \in A_{1+\kappa,\infty}} p_i$ and $(\mathbf{T}q)_1 = \sum_{i \in A_{1+\kappa,\infty}} q_i$. Furthermore, $p' \geq (1+\kappa)q'$. Using convexity and nonnegativity of f, and the fact that $f(1) = 0$ (see I.1), we have $f(x) \leq f(y)$ for $0 \leq y \leq x \leq 1$. Using the nonnegativity of f (I.1), symmetry of f (I.2), and monotonically decreasing property of f on $[0,1]$, we obtain the following:

$$\begin{aligned} I_f(\mathbf{T}p, \mathbf{T}q) &= p' f\left(\frac{q'}{p'}\right) + (1-p') f\left(\frac{1-q'}{1-p'}\right) \\ &\geq p' f\left(\frac{q'}{p'}\right) \\ &\geq p' f\left(\frac{1}{1+\kappa}\right). \end{aligned} \tag{G.3}$$

Moreover, by the assumption that $\sum_{i \in A_{1+\kappa,\infty}} p_i f\left(\frac{q_i}{p_i}\right) \geq 0.25 I_f(p,q)$ (where we use the symmetry property of f in I.2), we have

$$0.25 I_f(p,q) \leq \sum_{i \in A_{1+\kappa,\infty}} p_i f\left(\frac{q_i}{p_i}\right) \leq \sum_{i \in A_{1+\kappa,\infty}} p_i f(\nu) = p' f(\nu), \tag{G.4}$$

[39] That is, we can apply the following argument to the distributions $\tilde{p} := q$ and $\tilde{q} := p$ with $\tilde{A}$ defined as in equation (G.1) (with $\tilde{p}$ and $\tilde{q}$). There is a slight asymmetry because of the elements that have likelihood ratio exactly $1+\kappa$ or $1/(1+\kappa)$, but note that if $\sum_{i \in A_{0,1/(1+\kappa)}} q_i f\left(\frac{p_i}{q_i}\right) \geq \frac{I_f(p,q)}{4}$, we have $\sum_{i \in \tilde{A}_{1+\kappa,\infty}} \tilde{q}_i f\left(\frac{\tilde{p}_i}{\tilde{q}_i}\right) \geq \sum_{i \in A_{0,1/(1+\kappa)}} q_i f\left(\frac{p_i}{q_i}\right) \geq \frac{I_f(p,q)}{4}$, because $A_{0,1/(1+\kappa)} \subseteq \tilde{A}_{1+\kappa,\infty}$, since the interval in $A_{0,l}$ is left-open.

where we use the facts that $q_i/p_i \in [\nu, 1]$ and f is decreasing on $[\nu, 1]$. Combining inequalities (G.3) and (G.4), we obtain

$$I_f(\mathbf{T}p, \mathbf{T}q) \geq \frac{f(1/(1+\kappa))}{4f(\nu)} I_f(p, q), \tag{G.5}$$

which implies that $\frac{I_f(p,q)}{I_f(\mathbf{T}p,\mathbf{T}q)} \leq \frac{4f(\nu)}{f(1/(1+\kappa))}$, proving the desired result.

We now comment on the computational complexity of finding a $\mathbf{T}^*$ that achieves the rate (G.5). Since the channel $\mathbf{T}^*$ only depends on κ, the algorithm only needs to check whether the threshold should be $1+\kappa$ or $1/(1+\kappa)$, which requires at most $\mathrm{poly}(k)$ operations.

Case 2: Main contribution by small ratio alphabets: We now consider the case when $\sum_{i \in A_{1,1+\kappa} \cup A_{1/(1+\kappa),1}} q_i f\left(\frac{p_i}{q_i}\right) \geq \frac{I_f(p,q)}{2}$. By symmetry (I.2), it suffices to consider the case when $\sum_{i \in A_{1,1+\kappa}} q_i f\left(\frac{p_i}{q_i}\right) \geq \frac{I_f(p,q)}{4}$.[40] This requires us to handle the elements where p_i and q_i are close, and the following arguments form the main technical core of this section.

We first state a reverse Markov inequality, proved in Appendix G.2.3, whose role will become clear later in the proof:

Lemma G.2.1 (Generalized reverse Markov inequality)**.** *Let Y be a random variable over $[0, \beta)$ with expectation $\mathbb{E}[Y] > 0$. Let $k' = 1 + \log(\beta/\mathbb{E}[Y])$. Then*

$$\sup_{0 \leq \nu_1 \leq \cdots \leq \nu_D = \beta} \sum_{j=1}^{D-1} \nu_j \, \mathbb{P}\left(Y \in [\nu_j, \nu_{j+1})\right) \geq \frac{1}{13} \mathbb{E}[Y] \min\left\{1, \frac{D}{R}\right\}, \tag{G.6}$$

where $R = k' := 1 + \log(\beta/\mathbb{E}[Y])$. Furthermore, the bound (G.6) can be achieved by ν_j's such that $\nu_j = \min\{\beta, x2^j\}$ for some $x \in [0, \beta]$.

For the special case where Y is supported on k points, we may set $R = \min\{k, k'\}$, and there is a $\mathrm{poly}(k, D)$ algorithm to find ν_j's that achieve the bound (G.6).

For any $i \in A_{1,1+\kappa}$, both q_i and p_i are positive. Let $\delta_i = \frac{p_i}{q_i} - 1$, which lies in $[0, \kappa)$ by definition. Then $p_i = q_i(1 + \delta_i)$. Let X be a random variable over $[0, \kappa)$ such that for $i \in A_{1,1+\kappa}$, we define $\mathbb{P}(X = \delta_i) = q_i$ and $\mathbb{P}(X = 0) = 1 - \sum_{i \in A_{1,1+\kappa}} q_i$. We define Ω to be the support of the random variable X.

We now apply Lemma G.2.1 to the random variable $Y = X^\alpha$. Let $\beta = \kappa^\alpha$ and $R_2 = \min\{k, 1 + \log(\kappa^\alpha/\mathbb{E}[X^\alpha])\}$. Let $0 \leq \nu_1' \leq \cdots \leq \nu_D' = \beta$ be thresholds achieving the

[40] There is a slight asymmetry here as well, but similar to the previous footnote, it suffices to consider this case.

bound (G.6). Let $\nu_j = (\nu'_j)^{1/\alpha}$ for all $j \in [D]$. We thus have

$$\sum_{j=1}^{D-1} \nu_j^\alpha \, \mathbb{P}(X \in [\nu_j, \nu_{j+1})) \geq \frac{1}{13} \, \mathbb{E}[X^\alpha] \min\left\{1, \frac{D}{R_2}\right\}. \tag{G.7}$$

We now define the thresholds $\Gamma = (\gamma_1, \ldots, \gamma_{D-1})$ such that $\gamma_j = 1 + \nu_j$ for $i \in [D-1]$. We set $\gamma_0 = 0$ and $\gamma_\infty = 0$. Recall that by definition, for $j \in [D-1]$, we have $A_{\gamma_j, \gamma_{j+1}} = \{i : p_i/q_i \in [\gamma_j, \gamma_{j+1})\}$. Since $1 \leq \gamma_1 \leq \gamma_{D-1} = 1 + \nu_{D-1} \leq 1 + \kappa$, we have the following for $j \in [D-2]$:

$$A_{\gamma_j, \gamma_{j+1}} = \{i : p_i/q_i \in [\gamma_j, \gamma_{j+1})\} = \{i : \delta_i \in [\nu_j, \nu_{j+1})\}.$$

Note that for any $j \in [D-2]$ and any function g, we have

$$\begin{aligned} \sum_{i \in A_{\gamma_j, \gamma_{j+1}}} g(\delta_i) q_i &= \sum_{i \in A_{\gamma_j, \gamma_{j+1}}} g(\delta_i) \, \mathbb{P}(X = \delta_i) \\ &= \sum_{x \in \Omega \cap [\nu_j, \nu_{j+1})} g(x) \, \mathbb{P}(X = x) = \mathbb{E}\left[g(X) \mathop{\mathbb{I}}_{X \in [\nu_j, \nu_{j+1})} \right]. \end{aligned} \tag{G.8}$$

Using I.3 and the fact that $0 \leq \delta_i \leq \kappa$, we further have

$$\sum_{i \in A_{1, 1+\kappa}} q_i f\left(\frac{p_i}{q_i}\right) = \sum_{i \in A_{1, 1+\kappa}} q_i f(1 + \delta_i) \leq \sum_{i \in A_{1, 1+\kappa}} C_2 q_i \delta_i^\alpha = C_2 \, \mathbb{E}[X^\alpha], \tag{G.9}$$

where the last equality uses the same arguments as in inequality (G.8). Finally, we note that inequality (G.9) and the assumption $I_f(p, q) \leq 4 \sum_{i \in A_{1, 1+\kappa}} q_i f(p_i/q_i)$ implies that

$$\begin{aligned} I_f(p, q) &\leq 4 C_2 \, \mathbb{E}[X^\alpha], \qquad \text{and} \\ R_2 &\leq \min\{k, 1 + \log(4 C_2 \kappa^\alpha / I_f(p, q))\} = R. \end{aligned} \tag{G.10}$$

We use p' and q' to denote the probability measures $\mathbf{T}p$ and $\mathbf{T}q$, respectively, where $\mathbf{T}$ corresponds to the thresholds Γ. Thus, for $j \in [0 : D-1]$, we have $p'(j) = \sum_{i \in A_{\gamma_j, \gamma_{j+1}}} p_i$; we have an analogous expression for $q'(j)$. We now define the positive measure p'', as follows:

$$\begin{cases} p''_j = \sum_{i \in A_{\gamma_j, \gamma_{j+1}}} p_i, & \text{for } j \in [0 : D-2], \\ p''_j = \sum_{i \in A_{\gamma_{D-1}, 1+\kappa}} p_i, & \text{for } j = D-1, \end{cases}$$

and define q'' similarly. Recall that $\gamma_{D-1} = 1 + \nu_{D-1} \leq 1 + \kappa$. Equivalently, we have

$p''_j := \sum_{i \in A_{\gamma_j, \min\{\gamma_{j+1}, 1+\kappa\}}} p_i$ for each $j \in [0 : D-1]$, since $\gamma_D = \infty$. Note that p'' and q'' might not be probability measures, as their sums might be smaller than 1, but they are equal to p' and q', respectively, on all elements except the last. Moreover, we may define the "f-divergence" between p'' and q'' by mechanically applying the standard expression for f-divergence, but replacing the probability measures by p'' and q'', instead. The f-divergence between p'' and q'' thus obtained is smaller than the f-divergence between p' and q', since

$$q'_{D-1} f\left(\frac{p'_{D-1}}{q'_{D-1}}\right) \geq q''_{D-1} f\left(\frac{p''_{D-1}}{q''_{D-1}}\right),$$

which follows by noting that $q''_{D-1} \leq q'_{D-1}$, $p'_{D-1}/q'_{D-1} \geq p''_{D-1}/q''_{D-1} \geq 1$, and $f(x) \geq f(y) \geq 0$ for any $x \geq y \geq 1$.[41] We thus obtain the following relation:

$$I_f(p', q') \geq \sum_{j=0}^{D-1} q''_j f\left(\frac{p''_j}{q''_j}\right). \tag{G.11}$$

Fix $j \in [D-1]$. Using the facts that $0 \leq \frac{p''_j}{q''_j} - 1 \leq \kappa$ and $f(1+x) \geq C_1 x^\alpha$ for $x \in [0, \kappa]$ (cf. I.3), we have the following for any j such that $q''_j > 0$:

$$\begin{aligned}
q''_j f\left(\frac{p''_j}{q''_j}\right) &= q''_j f\left(1 + \frac{p''_j - q''_j}{q''_j}\right) \\
&\geq C_1 q''_j \left(\frac{p''_j - q''_j}{q''_j}\right)^\alpha \\
&= C_1 q''_j \left(\frac{\sum_{i \in A_{\gamma_j, \min\{\gamma_{j+1}, 1+\kappa\}}} q_i \delta_i}{\sum_{i \in A_{\gamma_j, \min\{\gamma_{j+1}, 1+\kappa\}}} q_i}\right)^\alpha \\
&\geq C_1 q''_j \nu_j^\alpha \qquad \left(\text{using } \delta_i \geq \nu_j \text{ for } i \in A_{\gamma_j, \gamma_{j+1}}\right) \\
&\geq C_1 \nu_j^\alpha \, \mathbb{P}(X \in [\nu_j, \nu_{j+1})).
\end{aligned} \tag{G.12}$$

We note that this inequality is also true if $q''_j = 0$, because $q''_j = \mathbb{P}\left(X \in [\nu_j, \nu_{j+1})\right)$, and if the former is zero, then the expression in inequality (G.12) is also zero, while $q''_j f\left(\frac{p''_j}{q''_j}\right)$ is nonnegative.

[41] We briefly outline how $p'_{D-1}/q'_{D-1} \geq p''_{D-1}/q''_{D-1}$: Let $p'_{D-1} = p''_{D-1} + x$ and $q'_{D-1} = q''_{D-1} + y$ for $x, y \in \mathbb{R}$. By construction, we have that $x, y \geq 0$ and $x/y \geq 1 + \kappa \geq p''_{D-1}/q''_{D-1}$. Expanding $p'_{D-1}/q'_{D-1} - p''_{D-1}/q''_{D-1}$, we get the desired conclusion.

Overall, we obtain the following series of inequalities:

$$\begin{aligned} I_f(p', q') &\geq \sum_{j=1}^{D-1} q''_j f\left(\frac{p''_j}{q''_j}\right) && \text{(using inequality (G.11) and } f \geq 0) \\ &\geq C_1 \sum_{j=1}^{D-1} \nu_j^\alpha \, \mathbb{P}(X \in [\nu_j, \nu_{j+1})) && \text{(using inequality (G.12))} \\ &\geq \frac{C_1}{13} \mathbb{E}[X^\alpha] \min\left\{1, \frac{D}{R_2}\right\} && \text{(using inequality (G.7))} \\ &\geq \frac{C_1}{52 C_2} I_f(p, q) \min\left\{1, \frac{D}{R}\right\} && \text{(using inequality (G.10))}\,. \end{aligned}$$

This shows that there exists a $\mathbf{T} \in \mathcal{T}_D^{\texttt{thresh}}$ such that

$$\frac{I_f(p, q)}{I_f(\mathbf{T}p, \mathbf{T}q)} \leq \frac{52 C_2}{C_1} \max\left\{1, \frac{R}{D}\right\}. \tag{G.13}$$

We now comment on the computational complexity of finding a $\mathbf{T}^*$ that achieves the rate (G.13). Finding the thresholds Γ is equivalent to finding $(\nu'_1, \ldots, \nu'_{D-1})$. As noted in Lemma G.2.1 and its proof, the guarantee of inequality (G.6) can be achieved by choosing ν'_j in one of the following ways:

- Setting $\nu'_j = \min\{\kappa^\alpha, x 2^j\}$ for all j and optimizing over x. As the random variable Y has support of at most k, this algorithm runs in $\mathrm{poly}(k, D)$-time.
- Choosing the top $D-1$ elements that maximize $\delta_i q_i$, and defining ν'_j appropriately.

□

G.2.2 Tightness of Reverse Data Processing Inequality: Proof of Lemma 10.3.6

Lemma 10.3.6 (Reverse data processing is tight)**.** *There exist positive constants c_1, c_2, c_3, c_4, c_5, and c_6 such that for every $\rho \in (0, c_1)$ and $D \geq 2$, there exist $k \in [c_2 \log(1/\rho), c_3 \log(1/\rho)]$ and two distributions p and q on $[k]$ such that $d^2_{\mathrm{h}}(p, q) \in [c_4\rho, c_5\rho]$ and*

$$\inf_{\mathbf{T} \in \mathcal{T}_D^{\texttt{thresh}}} \frac{d^2_{\mathrm{h}}(p, q)}{d^2_{\mathrm{h}}(\mathbf{T}p, \mathbf{T}q)} \geq c_6 \cdot \frac{R'}{D}, \tag{10.8}$$

where $R' = \max\{k, k'\}$ and $k' = \log(1/\rho)$. Thus, $R' = \Theta(k) = \Theta(\log(1/\rho))$.

Proof. We will design p and q such that $p_i/q_i \in [0.5, 1.5]$ for all i. Fix any set of thresholds $\Gamma = \{\gamma_1, \ldots, \gamma_{D-1}\}$ which, without loss of generality, lie in $[0.5, 1.5]$. Let $\mathbf{T}$ be the corresponding channel. Let p' and q' be the distributions after using the channel $\mathbf{T}$.

Note that k will depend on ρ, which will be decided later. For now, let k be even, equal to $2m$. Let $\tilde{q}$ be an arbitrary distribution on $[m]$, to be decided later. Using this $\tilde{q}$, we define a distribution q on $[k]$, as follows:

$$q_i = \begin{cases} 0.5\tilde{q}_i, & \text{if } i \in [m] \\ 0.5\tilde{q}_{i-m}, & \text{if } i \in [k] \setminus [m]. \end{cases}$$

Let $\tilde{\delta} \in [0, 0.5]^m$, also to be decided later. Using $\tilde{\delta}$, we define δ, as follows:

$$\delta_i = \begin{cases} \tilde{\delta}_i, & \text{if } i \in [m] \\ -\tilde{\delta}_{i-m}, & \text{if } i \in [k] \setminus [m]. \end{cases}$$

We now define p as follows: For $i \in [k]$, define $p_i = q_i(1 + \delta_i)$. Equivalently,

$$p_i = \begin{cases} 0.5\tilde{q}_i(1 + \tilde{\delta}_i), & \text{if } i \in [m] \\ 0.5\tilde{q}_{i-m}(1 - \tilde{\delta}_{i-m}), & \text{if } i \in [k] \setminus [m]. \end{cases}$$

Thus, p is a valid distribution if q is a valid distribution. Let $\tilde{X}$ be the random variable such that $\mathbb{P}\{\tilde{X} = \tilde{\delta}_i\} = \tilde{q}_i$. We will need the following results, whose proofs are given at the end of this section:

Claim G.2.2. *We have the following inequality:*

$$0.02\,\mathbb{E}[\tilde{X}^2] \le d_{\mathrm{h}}^2(p, q) \le \mathbb{E}[\tilde{X}^2].$$

Claim G.2.3. *Let $\mathbf{T} \in \mathcal{T}_D^{\texttt{thresh}}$ be a channel corresponding to a threshold test. Then*

$$d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \le \sup_{0<\nu_1'<\cdots<\nu_D'=1} \sum_{j=1}^{D-1} \mathbb{P}\left\{\tilde{X} \ge \nu_j'\right]\right\} \left(\mathbb{E}\left[\tilde{X} | \tilde{X} \ge \nu_j'\right]\right)^2. \tag{G.14}$$

We will now show that there exist p and q (i.e., $\tilde{q} \in \mathbb{R}^m$ and $\tilde{\delta} \in \mathbb{R}^m$) such that the desired conclusion holds. Defining $\tilde{q}$ and $\tilde{\delta}$ is equivalent to showing the existence of a random variable $\tilde{X}$ satisfying the desired properties. This is given in Claim G.2.4 below, showing that there exists a distribution $\tilde{X}$ such that the following hold: (i)

$\mathbb{E}[\tilde{X}^2] = \Theta(\rho)$; (ii) the expression on the right-hand side of inequality (G.14), for any choice of thresholds Γ, is upper-bounded by a constant multiple of $\frac{\mathbb{E}[\tilde{X}^2]D}{R'}$; and (iii) $R' = \max\{m, k'\} = \Theta(\log(1/\rho))$.

Claim G.2.4 (Tightness of reverse Markov inequality)**.** *There exist constants c_1, c_2, c_3, c_4, c_5, and c_6 such that for every $\rho \in (0, c_5)$, there exists an integer $k \in [c_3 \log(1/\rho), c_4 \log(1/\rho)]$ and a probability distribution p, supported over k points in $(0, 0.5]$, such that the following hold:*

1. $\mathbb{E}[X^2] \in [c_1\rho, c_2\rho]$*, and for every $D \le 0.1k$,*

$$\sup_{0<\delta_1<\cdots<\delta_D=1} \sum_{j=1}^{D-1} \mathbb{P}\{X \ge \delta_j\} \left(\mathbb{E}[X | X \ge \delta_j]\right)^2 \le c_6 \cdot \mathbb{E}[X^2] \frac{D}{R'}, \tag{G.15}$$

   where $R' = \max\{k, k'\}$ and $k' = \log(3/\mathbb{E}[X^2])$.

2. $\mathbb{E}[Y] = [c_1\rho, c_2\rho]$*, and*

$$\sup_{0<\delta'_1<\cdots<\delta'_D=1} \sum_{j=1}^{D-1} \delta'_j \, \mathbb{P}\left(Y \in [\delta'_j, \delta'_{j+1})\right) \le c_6 \cdot \mathbb{E}[Y] \frac{D}{R'}, \tag{G.16}$$

   where $R' = \max\{k, k'\}$ and $k' = \log(3/\mathbb{E}[Y])$. Moreover, $R' = \Theta(\log(1/\rho))$.

We provide the proof of Claim G.2.4 in Appendix G.2.4. Using Claims G.2.2 to G.2.4, we obtain the following for any threshold channel $\mathbf{T}$:

$$d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) \lesssim \mathbb{E}[\tilde{X}^2] \frac{D}{R'} \lesssim d_{\mathrm{h}}^2(p, q) \frac{D}{R'},$$

completing the proof. □

The omitted proofs of Claim G.2.2 and Claim G.2.3 are given below.

Proof. (Proof of Claim G.2.2) We have the following:

$$\begin{aligned}
d_{\mathrm{h}}^2(p, q) &= \sum_{i \in [m]} \left(\sqrt{q_i\,(1+\delta_i)} - \sqrt{q_i}\right)^2 + \sum_{i \in [k] \setminus [m]} \left(\sqrt{q_i} - \sqrt{q_i\,(1+\delta_i)}\right)^2 \\
&= 0.5 \sum_{i \in [m]} \left(\sqrt{\tilde{q}_i\left(1+\tilde{\delta}_i\right)} - \sqrt{\tilde{q}_i}\right)^2 + 0.5 \sum_{i \in [m]} \left(\sqrt{\tilde{q}_i} - \sqrt{\tilde{q}_i\left(1-\tilde{\delta}_i\right)}\right)^2 \\
&= 0.5 \sum_{i \in [m]} \tilde{q}_i \left(\left(\sqrt{1+\tilde{\delta}_i} - 1\right)^2 + \left(1 - \sqrt{1-\tilde{\delta}_i}\right)^2\right).
\end{aligned}$$

Using the fact that for $x \in [0,1]$, we have

$$\begin{aligned}\sqrt{1+x} - 1 &\geq 0.1x,\\ 1 - \sqrt{1-x} &\geq 0.1x,\\ \sqrt{1+x} &\leq 1+x,\\ 1-x &\leq \sqrt{1-x},\end{aligned}$$

we obtain

$$\mathbb{E}[\tilde{X}^2] \geq d_{\mathrm{h}}^2(p,q) \geq 0.02\,\mathbb{E}[\tilde{X}^2].$$

□

Proof. (Proof of Claim G.2.3) Suppose $\mathbf{T}$ corresponds to a threshold test with thresholds $\Gamma = \{\gamma_1, \ldots, \gamma_{D-1}\}$ such that $\gamma_j < \gamma_{j+1}$. We define $\gamma_0 = \min_i p_i/q_i$ and $\gamma_D = \max_i p_i/q_i$. It suffices to consider the case when all $\gamma_j \in [0.5, 1.5]$ for $j \in [0:D]$. Let $p' = \mathbf{T}p$ and $q' = \mathbf{T}q$. Let $j^* \in [D-1]$ be such that $\gamma_{j^*-1} < 1$ and $\gamma_{j^*} \geq 1$.

We now define the ν_j's as follows for $j \in [0:D-1]$:

$$\nu_j = \begin{cases}\gamma_j - 1, & \text{if } j \geq j^*\\ 1 - \gamma_j, & \text{otherwise.}\end{cases}$$

Thus, $\nu_j \in [0,1)$.
For $j \in [0:D-1]$, define

$$A_j := \{i \in [k] : (p_i/q_i) \in [\gamma_j, \gamma_{j+1})\} = \{i : 1 + \delta_i \in [\gamma_j, \gamma_{j+1})\}.$$

For $j \geq j^*$, we have $A_j = \{i : \delta_i \in [\nu_j, \nu_{j+1})\}$. For $j < j^*$, we have

$$A_j = \{i \in [k] : -\delta_i \in (\nu_{j+1}, \nu_j]\} = \left\{i \in [k] : \tilde{\delta}_{i-m} \in (\nu_{j+1}, \nu_j]\right\}.$$

For $j \in [0:D-1]$, we have $p'_j = \sum_{i \in A_j} p_i$ and $q'_j = \sum_{i \in A_j} q_i$.

We have the following decomposition of the squared Hellinger distance between p' and q':

$$d_{\mathrm{h}}^2(p',q') = \sum_{j<j^*} \left(\sqrt{p'_j} - \sqrt{q'_j}\right)^2 + \sum_{j \geq j^*} \left(\sqrt{p'_j} - \sqrt{q'_j}\right)^2 \tag{G.17}$$

We analyze these two terms separately:

Case 1**:** $j \geq j^*$**:** Let j be such that $q'_j > 0$. We have $p'_j \in [q'_j, 1.5q'_j]$. Using the fact that $\sqrt{1+x} - 1 \leq x$ for $x \in [0, 0.5]$, we have

$$\left(\sqrt{p'_j} - \sqrt{q'_j}\right)^2 = q'_j \left(\sqrt{1 + \frac{p'_\gamma - q'_j}{q'_j}} - 1\right)^2 \leq \frac{(p'_i - q'_j)^2}{q'_j}. \tag{G.18}$$

Since $\gamma_j \geq 1$, note that

$$q'_j = \sum_{i \in A_j} q_i = \sum_{i \in [m]: \tilde{\delta}_i \in [\nu_j, \nu_{j+1})} q_i = \sum_{i \in [m]: \tilde{\delta}_i \in [\nu_j, \nu_{j+1})} 0.5\tilde{q}_i = 0.5\,\mathbb{P}\{\tilde{X} \in [\nu_j, \nu_{j+1})\}.$$

Similarly, we have

$$p'_j - q'_j = \sum_{i \in A_j} \delta_i q_i = \sum_{i \in [m]: \tilde{\delta}_i \in [\nu_j, \nu_{j+1})} \delta_i q_i = 0.5 \sum_{i \in [m]: \tilde{\delta}_i \in [\nu_j, \nu_{j+1})} \tilde{\delta}_i \tilde{q}_i = 0.5\,\mathbb{E}\left[\tilde{X} \mathop{\mathbb{I}}_{\tilde{X} \in [\nu_j, \nu_{j+1})}\right].$$

Combining the last two displayed equations with inequality (G.18) and using the definition of conditional expectation, we then obtain

$$\begin{aligned}
\sum_{j \geq j^*} \left(\sqrt{p'_j} - \sqrt{q'_j}\right)^2 &\leq 0.5 \sum_{j \geq j^*} \mathbb{P}\left\{\tilde{X} \in [\nu_j, \nu_{j+1})\right\} \left(\mathbb{E}\left[\tilde{X} | \tilde{X} \in [\nu_j, \nu_{j+1})\right]\right)^2 \\
&\leq 0.5 \sum_{j \geq j^*} \mathbb{P}\left\{\tilde{X} \geq \nu_j\right\} \left(\mathbb{E}\left[\tilde{X} | \tilde{X} \geq \nu_j)\right]\right)^2.
\end{aligned} \tag{G.19}$$

Case $2 : j < j^*$**:** Let $j < j^*$ be such that $q'_j > 0$. We have $p'_j \in [q'_j/2, q'_j)$. Using the fact that $1 - \sqrt{1-x} \leq x$ for $x \in [0, 1]$, we have

$$\left(\sqrt{q'_j} - \sqrt{p'_j}\right)^2 = q'_j \left(1 - \sqrt{1 - \frac{q'_j - p'_i}{q'_j}}\right)^2 \leq \frac{(q'_j - p'_i)^2}{q'_j}. \tag{G.20}$$

Since $\gamma_j < 1$, we have

$$\begin{aligned}
q'_j &= \sum_{i \in A_j} q_i = \sum_{i \in [k] \setminus [m]: \tilde{\delta}_{i-m} \in (\nu_{j+1}, \nu_j]} q_i = \sum_{i \in [k] \setminus [m]: \tilde{\delta}_{i-m} \in (\nu_{j+1}, \nu_j]} 0.5\tilde{q}_{i-m} \\
&= 0.5\,\mathbb{P}\{\tilde{X} \in (\nu_{j+1}, \nu_j]\}.
\end{aligned}$$

Similarly, we have

$$q'_j - p'_i = \sum_{i \in A_j} (-\delta_i q_i) = \sum_{i \in [k] \setminus [m] : \tilde{\delta}_{i-m} \in (\nu_{j+1}, \nu_j]} \tilde{\delta}_i (0.5 \tilde{q}_{i-m}) = 0.5 \, \mathbb{E}\left[\tilde{X} \mathop{\mathbb{I}}_{\tilde{X} \in (\nu_{j+1}, \nu_j]} \right].$$

Combining the last two displayed equations with inequality (G.20) and using the definition of conditional expectation, we then obtain

$$\begin{aligned} \sum_{j < j^*} \left(\sqrt{p'_j} - \sqrt{q'_j} \right)^2 &\leq 0.5 \sum_{j < j^*} \mathbb{P}\left\{ \tilde{X} \in (\nu_{j+1}, \nu_j] \right\} \left(\mathbb{E}\left[\tilde{X} | \tilde{X} \in (\nu_{j+1}, \nu_j] \right] \right)^2 \\ &\leq 0.5 \sum_{j < j^*} \mathbb{P}\left\{ \tilde{X} > \nu_j \right] \right\} \left(\mathbb{E}\left[\tilde{X} | \tilde{X} > \nu_j \right] \right] \right)^2 . \end{aligned} \tag{G.21}$$

Combining inequalities (G.19) and (G.21), we can complete the proof by noting that $\tilde{X}$ is a discrete random variable, so the distinction between $\tilde{X} \geq \nu_j$ (cf. inequality (G.19)) and $\tilde{X} > \nu_j$ (cf. inequality (G.21)) does not matter when taking the supremum. □

G.2.3 Reverse Markov Inequality

Lemma G.2.1 (Generalized reverse Markov inequality)**.** *Let Y be a random variable over $[0, \beta)$ with expectation $\mathbb{E}[Y] > 0$. Let $k' = 1 + \log(\beta / \mathbb{E}[Y])$. Then*

$$\sup_{0 \leq \nu_1 \leq \cdots \leq \nu_D = \beta} \sum_{j=1}^{D-1} \nu_j \, \mathbb{P}\left(Y \in [\nu_j, \nu_{j+1}) \right) \geq \frac{1}{13} \, \mathbb{E}[Y] \min\left\{ 1, \frac{D}{R} \right\}, \tag{G.6}$$

where $R = k' := 1 + \log(\beta / \mathbb{E}[Y])$. Furthermore, the bound (G.6) can be achieved by ν_j's such that $\nu_j = \min\{\beta, x 2^j\}$ for some $x \in [0, \beta]$.

For the special case where Y is supported on k points, we may set $R = \min\{k, k'\}$, and there is a $\mathrm{poly}(k, D)$ algorithm to find ν_j's that achieve the bound (G.6).

Proof. We can safely assume that $D \leq R$. Under this assumption on D, we will show that the desired expression is lower-bounded by both of the following quantities (up to constants): $\frac{\mathbb{E}[Y] D}{k}$ and $\frac{\mathbb{E}[Y] D}{k'}$. We will also assume that $\beta = 1$; otherwise, it suffices to apply the following argument to $\frac{Y}{\beta}$.

Dependence on k: Suppose Y has support size k.[42] Let the support elements be $\{\delta'_i\}_{i=1}^k$, such that $\delta'_1 < \delta'_2 < \cdots < \delta'_k < 1$. Let $\{p_i\}_{i=1}^k$ be such that $\mathbb{P}(Y = \delta'_i) = p_i$ and $\sum_{i=1}^k p_i = 1$.

[42] It is easy to see that if the support size is strictly smaller than k, we have a tighter bound.

It suffices to prove that there exists a labeling $\pi : [D-1] \to [k]$ such that $\pi(1) < \pi(2) < \cdots < \pi(D-1)$ and the following bound holds:

$$\sum_{j=1}^{D-1} \delta'_{\pi(j)} p_{\pi(j)} \geq \mathbb{E}[Y] \left(\frac{D-1}{k} \right). \tag{G.22}$$

This is true because for $j \in [D-1]$, we have

$$p_{\pi(j)} = \mathbb{P}\left\{ Y \in [\delta_{\pi(j)}, \delta_{\pi(j)+1}) \right\} \leq \mathbb{P}\left\{ Y \in [\delta_{\pi(j)}, \delta_{\pi(j+1)}) \right\},$$

where we define $\delta_{k+1} := 1$ and $\pi(D) := 1$, and the desired conclusion follows by setting $\nu_j = \delta_{\pi(j)}$. In the rest of the proof, we will show that such a π exists.

Let $\sigma : [k] \to [k]$ be a permutation such that $p_{\sigma(i)} \delta_{\sigma(i)} \geq p_{\sigma(i+1)} \delta_{\sigma(i+1)}$. Then we have

$$\frac{\mathbb{E}[Y]}{k} = \frac{\sum_{i=1}^{k} p_i \delta_i}{k} = \frac{\sum_{i=1}^{k} p_{\sigma(i)} \delta_{\sigma(i)}}{k} \leq \frac{\sum_{i=1}^{D-1} p_{\sigma(i)} \delta_{\sigma(i)}}{D-1} = \frac{\sum_{i=1}^{D-1} p_{\pi'(i)} \delta_{\pi'(i)}}{D-1},$$

for some $\pi' : [D-1] \to [k]$ such that $\pi'(1) < \pi'(2) < \cdots < \pi'(D-1)$. Thus, we have established inequality (G.22). Note that the desired bound is achieved by choosing the ν_j's, as follows: Let S be the set of top $D-1$ elements among the support of Y that maximize $y\,\mathbb{P}(Y = y)$, and let the ν_j's have values in S such that they are increasing and distinct. It is clear that this assignment can be implemented in $\mathrm{poly}(k, D)$ time.

Dependence on k': We begin by noting that the desired expression can also be written as

$$\sum_{j=1}^{D-1} (\nu_j - \nu_{j-1}) \, \mathbb{P}\{Y \geq \nu_j\},$$

where $\nu_0 := 0$. We need to obtain a lower bound on the supremum of this expression over the ν_j's. In fact, we will show a stronger claim, where we fix the ν_j's in a particular way: We will take ν_j to be of the form $x 2^{j-1}$, for $j \in [D-1]$, and optimize over $x \in (0, 1)$. Note that we can allow $\nu_j \geq 1$ without loss of generality, because their contribution to the desired expression would then be 0. In the rest of the proof, we will show the following claim for $c_1 = \frac{1}{13}$:

$$\sup_{x \in (0,1)} x\,\mathbb{P}\{Y \geq x\} + \sum_{j=2}^{D-1} \left(x 2^{j-1} - x 2^{j-2} \right) \mathbb{P}\left\{ Y \geq x 2^{j-1} \right\} \geq c_1\, \mathbb{E}[Y] \frac{D}{k'}. \tag{G.23}$$

Suppose that the desired conclusion does not hold. We will now derive a contradiction. Under the assumption that inequality (G.23) is false, we have the following, for each $x \in (0,1)$:

$$c_1 \,\mathbb{E}[Y]\frac{D}{k'} > x\,\mathbb{P}\left\{Y \geq x\right\} + \sum_{j=2}^{D-1} x\left(2^{j-1} - 2^{j-2}\right)\mathbb{P}\left\{Y \geq x2^{j-1}\right\}$$
$$= x\left(\mathbb{P}\left\{Y \geq x\right\} + \sum_{j=1}^{D-2} 2^{j-1}\,\mathbb{P}\left\{2^{-j}Y \geq x\right\}\right).$$

We thus obtain the following, for all $x \in (0,1)$:

$$\mathbb{P}\left\{Y \geq x\right\} + \sum_{j=1}^{D-2} 2^{j-1}\,\mathbb{P}\left\{2^{-j}Y \geq x\right\} < c_1\,\mathbb{E}[Y] \cdot \frac{D}{k'} \cdot \frac{1}{x}. \tag{G.24}$$

Using the fact that the probabilities are bounded by 1, we also have the following bound on the expression on the left side of inequality (G.24):

$$\mathbb{P}\left\{Y \geq x\right\} + \sum_{j=1}^{D-2} 2^{j-1}\,\mathbb{P}\left\{2^{-j}Y \geq x\right\} \leq 1 + \sum_{j=1}^{D-2} 2^{j-1} = 2^{D-2}. \tag{G.25}$$

Combining the two bounds in inequalities (G.24) and (G.25), the following holds for every $x \in (0,1)$:

$$\mathbb{P}\left\{Y \geq x\right\} + \sum_{j=1}^{D-2} 2^{j-1}\,\mathbb{P}\left\{2^{-j}Y \geq x\right\} \leq \min\left\{2^{D-1}, \frac{c_1 D\,\mathbb{E}[Y]}{k'x}\right\}. \tag{G.26}$$

Using the fact that $Y \in [0,1]$, we also have the following for every $a > 1$:

$$\mathbb{E}\left[\frac{Y}{a}\right] = \int_0^{1/a} \mathbb{P}\left(\frac{Y}{a} \geq t\right) dt = \int_0^1 \mathbb{P}\left(\frac{Y}{a} \geq t\right) dt,$$

implying that

$$\int_0^1 \left(\mathbb{P}\left\{Y \geq t\right\} + \sum_{j=1}^{D-2} 2^{j-1}\,\mathbb{P}\left\{Y2^{-j} \geq t\right\}\right) dt = \mathbb{E}[Y] + \sum_{j=1}^{D-2} 2^{j-1}\,\mathbb{E}[2^{-j}Y] = \frac{D\,\mathbb{E}[Y]}{2}. \tag{G.27}$$

Combining inequalities (G.26) and (G.27), we obtain the following, for an arbitrary

$x_* \in (0,1)$:

$$\begin{aligned}\mathbb{E}[Y] &= \frac{2}{D}\int_0^1 \left(\mathbb{P}\left\{Y \geq t\right\} + \sum_{i=1}^{D-2} 2^{i-1}\,\mathbb{P}\left\{Y 2^{-i} \geq t\right\}\right) dt \\ &\leq \frac{2}{D}\int_0^1 \min\left\{2^{D-1}, \frac{c_1 D\,\mathbb{E}[Y]}{k' x}\right\} dt \\ &\leq \frac{2^D x_*}{D} + \frac{2c_1\,\mathbb{E}[Y]}{k'}\log\left(\frac{1}{x_*}\right). \end{aligned} \tag{G.28}$$

Let $x_* = 2^{-D}\left(\mathbb{E}[Y]/k'\right)$. Then

$$\log(1/x_*) = D + \log(1/\mathbb{E}[Y]) + \log(k') \leq 3k',$$

where we use the fact that $\max\{D, \log(1/\mathbb{E}[Y])\} \leq k'$. Using $k' \geq 1$ and $D \geq 2$, the expression on the right-hand side of inequality (G.28) can be further upper-bounded, to obtain the following inequality:

$$\mathbb{E}[Y] \leq \left(\frac{\mathbb{E}[Y]}{k'D}\right) + \frac{2c_1\,\mathbb{E}[Y]}{k'}(3k') \leq \frac{\mathbb{E}[Y]}{2} + 6c_1\,\mathbb{E}[Y],$$

which is a contradiction, since $\mathbb{E}[Y] > 0$ and $c_1 < \frac{1}{12}$. Thus, we conclude that inequality (G.23) is true. □

G.2.4 Tightness of Reverse Markov Inequality

Claim G.2.4 (Tightness of reverse Markov inequality)**.** *There exist constants c_1, c_2, c_3, c_4, c_5, and c_6 such that for every $\rho \in (0, c_5)$, there exists an integer $k \in [c_3 \log(1/\rho), c_4 \log(1/\rho)]$ and a probability distribution p, supported over k points in $(0, 0.5]$, such that the following hold:*

1. $\mathbb{E}[X^2] \in [c_1\rho, c_2\rho]$*, and for every $D \leq 0.1k$,*

$$\sup_{0<\delta_1<\cdots<\delta_D=1} \sum_{j=1}^{D-1} \mathbb{P}\left\{X \geq \delta_j\right\}\left(\mathbb{E}\left[X \middle| X \geq \delta_j\right]\right)^2 \leq c_6 \cdot \mathbb{E}[X^2]\frac{D}{R'}, \tag{G.15}$$

   where $R' = \max\{k, k'\}$ and $k' = \log(3/\mathbb{E}[X^2])$.

2. $\mathbb{E}[Y] = [c_1\rho, c_2\rho]$*, and*

$$\sup_{0<\delta_1'<\cdots<\delta_D'=1} \sum_{j=1}^{D-1} \delta_j' \,\mathbb{P}\left(Y \in [\delta_j', \delta_{j+1}')\right) \leq c_6 \cdot \mathbb{E}[Y]\frac{D}{R'}, \tag{G.16}$$

where $R' = \max\{k, k'\}$ *and* $k' = \log(3/\mathbb{E}[Y])$. *Moreover,* $R' = \Theta(\log(1/\rho))$.

Proof. For now, let $k \in \mathbb{N}$ be arbitrary; we will choose k so that $\mathbb{E}[X^2] \in [c_1\rho, c_2\rho]$. Consider the following discrete random variable Y supported on $\{2^{-i} : i \in [k]\}$:

$$\mathbb{P}\left\{Y = 2^{-i}\right\} = r2^i,$$

where r is chosen so that it is a valid distribution, i.e., r satisfies $1 = \sum_{i=1}^{k} r2^i = 2r(2^k - 1)$. Let $X = \sqrt{Y}$. We then have

$$\mathbb{E}[X^2] = \mathbb{E}[Y] = \sum_{i=1}^{k} 2^{-i}\left(r2^i\right) = rk. \tag{G.29}$$

Consider a $\delta_i' \in [2^{-j}, 2^{-(j-1)})$, for some $j \in [k]$, and let $\delta_i = \sqrt{\delta_i'}$. For any such choice, we obtain the following:

$$\mathbb{P}(X \geq \delta_i) = \mathbb{P}(Y \geq \delta_i') = \mathbb{P}(Y \geq 2^{-j}) = \sum_{i\in[j]} \mathbb{P}\left\{Y = 2^{-i}\right\} = \sum_{i\in[j]} r2^i = 2r\left(2^j - 1\right) \leq 2r2^j. \tag{G.30}$$

Thus, for any δ', we have $\delta' \mathbb{P}\{Y \geq \delta'\} \leq 2r$, showing that the expression in inequality (G.16) is upper-bounded by $2(D-1)r$, which is equal to $\frac{2\mathbb{E}[Y](D-1)}{k}$, by equation (G.29). It remains to show that $R \leq c_6 k/2$.

We first calculate bounds on k so that $\mathbb{E}[Y] = \Theta(\rho)$. Note that by construction, we have $r = 1/\left(2(2^k - 1)\right)$, implying that $r \in [2^{-k+1}, 2^{-k-1}]$. Since $\mathbb{E}[Y] = rk$, it suffices to choose k such that $f(k) \in [2c_1\rho, 0.5c_2\rho]$, where $f(j) := 2^{-j}j$. As $\frac{f(j+1)}{f(j)} \in (1/2, 1)$ for $j > 1$, we have $f(\lfloor \ln(1/\rho) \rfloor) \geq \rho\lceil \ln(1/\rho)\rceil$ and $f(\lceil 2\ln(1/\rho)\rceil) \leq \rho^2\lceil \log(1/\rho)\rceil$, so we know that such a k exists in $[\ln(1/\rho), 2\ln(1/\rho)]$ when $c_1 = 0.5$, $c_2 = 10$, and $c_5 = 2^{-20}$.

We now calculate the quantity R. By the definition of k', we have

$$k' = \log\left(\frac{3}{\mathbb{E}[X^2]}\right) = \log\left(\frac{3}{rk}\right) \geq \log\left(\frac{3 \cdot 2^{k-1}}{k}\right) = k\log 2 - \log k + \log(3/2).$$

As k is large enough, we have $k' \in [0.5k, 2k]$. Since $R = \max\{k, k'\}$, we have $R \in [0.5k, 2k]$. This completes the proof of the claim in inequality (G.16), with $c_6 = 4$.

We now prove the claim in inequality (G.15). We begin with the following:

$$\mathbb{E}[X \mathbb{I}_{X \geq 2^{-j/2}}] = \sum_{i\in[j]} 2^{-0.5i}\left(r2^i\right) = \sum_{i\in[j]} r2^{0.5i} = (r\sqrt{2})\left(\frac{2^{0.5j} - 1}{\sqrt{2} - 1}\right) \leq 10r2^{0.5j}.$$

For $\delta_i \in [2^{-j/2}, 2^{-(j-1)/2})$, for some j, we have the following:

$$\begin{aligned}\mathbb{P}\left\{X \geq \delta_i\right\}\left(\mathbb{E}\left[X | X \geq \delta_i\right]\right)^2 &= \mathbb{P}(X \geq 2^{-j/2})\left(\mathbb{E}[X | X \geq 2^{-j/2}]\right)^2 \\ &= \frac{\left(\mathbb{E}[X\,\mathbb{I}_{X \geq 2^{-j/2}}]\right)^2}{\mathbb{P}(X \geq 2^{-j/2})} \\ &\leq \frac{100r^2 2^j}{2r(2^j - 1)} \leq 100r.\end{aligned}$$

Thus, the supremum over any arbitrary δ_i is also upper-bounded by $100r$. Hence, we can upper-bound the expression on the left-hand side of inequality (G.15) by $100(D-1)r$, which is equal to $100 \cdot \frac{\mathbb{E}[X^2](D-1)}{k}$. Using the same calculations as in the first part of the proof, we prove inequality (G.15) with $c_6 = 200$. □

G.3 Simple Binary Hypothesis Testing

In this section, we prove results concerning binary hypothesis testing that were omitted from Section 10.4. We prove the equivalence between identical and non-identical channels for simple binary hypothesis testing (cf. Lemma 10.4.2) in Appendix G.3.1. Appendix G.3.2 focuses on robust binary hypothesis testing.

G.3.1 Equivalence between Identical and Non-identical Channels

Lemma 10.4.2 (Equivalence between identical and non-identical channels for simple hypothesis testing)**.** *Let $\mathcal{T}$ be a collection of channels from $\mathcal{X} \to \mathcal{Y}$. Let p and q be two distributions on $\mathcal{X}$. Then*

$$n^*_{\texttt{non-identical}}(p, q, \mathcal{T}) = \Theta\left(n^*_{\texttt{identical}}(p, q, \mathcal{T})\right).$$

Proof. Recall that we use $\beta_{\mathrm{h}}(p, q)$ to denote the Hellinger affinity between p and q. It suffices to consider the case where $n^*_{\texttt{identical}}(p, q, \mathbf{T})$ is larger than a fixed constant. Define the following:

$$h_* = \sup_{\mathbf{T} \in \mathcal{T}} d_{\mathrm{h}}(\mathbf{T}p, \mathbf{T}q), \quad \text{and} \quad \beta_* := \inf_{\mathbf{T} \in \mathcal{T}} \beta_{\mathrm{h}}(\mathbf{T}p, \mathbf{T}q),$$

and note that $\beta_* = 1 - 0.5h_*^2$. Let $n_* := n^*_{\texttt{identical}}(p, q, \mathbf{T})$. Let $\mathbf{T}_*$ be any channel such that $\beta_{\mathrm{h}}(\mathbf{T}_*p, \mathbf{T}_*q) \leq \beta_* + \epsilon\beta_*$, for some $\epsilon > 0$ satisfying $(1+\epsilon)^{n_*} \leq 2$. Let $p_* = \mathbf{T}_*p$ and

$q_* = \mathbf{T}_* q$.

Identical channels: Optimal $\mathbf{T}_*$: If each channel is identically $\mathbf{T}_*$, the joint distributions of n_* samples will either be $p_*^{\otimes n_*}$ or $q_*^{\otimes n_*}$.

Let $f(n) = d_{\mathrm{TV}}\left(p_*^{\otimes n}, q_*^{\otimes n}\right)$. Note that the probability of error for $p_*^{\otimes n}$ and $q_*^{\otimes n}$ is equal to $1 - f(n)$ (cf. Fact 10.2.4). Since the sample complexity of $\mathcal{B}(p_*, q_*)$ is at least n_*, the probability of error with $n_* - 1$ samples must be greater than 0.1, i.e, $f(n_* - 1) < 0.9$. Using Fact 10.2.2, we have $d_{\mathrm{h}}^2\left(p_*^{\otimes(n_*-1)}, q_*^{\otimes(n_*-1)}\right) \leq 1.8$, and consequently, $\beta_{\mathrm{h}}\left(p_*^{\otimes(n_*-1)}, q_*^{\otimes(n_*-1)}\right) \geq 0.1$. Using the tensorization of Hellinger affinity (cf. Fact 10.2.2) and the relation between β_* and $\beta_{\mathrm{h}}(p_*, q_*)$, we have

$$(\beta_*)^{n_*-1} \geq \left(\frac{\beta_{\mathrm{h}}(p^*, q^*)}{1+\epsilon}\right)^{n_*-1} \geq \frac{1}{2}\beta_{\mathrm{h}}\left(p_*^{\otimes(n_*-1)}, q_*^{\otimes(n_*-1)}\right) \geq 0.05. \tag{G.31}$$

Non-identical channels: We now show that even if n non-identical channels are allowed but $n \leq 0.01 n_*$, the probability of error is at least 0.2. For a choice of $\mathbf{T}_1, \ldots, \mathbf{T}_n$, let $P_n' := \prod_{i=1}^n \mathbf{T}_i p$ and $Q_n' := \prod_{i=1}^n \mathbf{T}_i q$ be the resulting joint probability distributions under p and q, respectively. As the probability of error of the best test is $1 - d_{\mathrm{TV}}\left(P_n', Q_n'\right)$ (cf. Fact 10.2.4), it suffices to show that if $n \leq 0.01 n_*$, then $d_{\mathrm{TV}}(P_n', Q_n') \leq 0.8$.

Using Fact 10.2.2, it suffices to show that $d_{\mathrm{h}}^2(P_n', Q_n') \leq 0.64$. Equivalently, it suffices to show that $\beta_{\mathrm{h}}(P_n', Q_n') \geq 0.68$. Using the tensorization of Hellinger affinity and optimality of β_*, we have

$$\beta_{\mathrm{h}}(P_n', Q_n') = \prod_{i=1}^n \beta_{\mathrm{h}}(\mathbf{T}_i p, \mathbf{T}_i q) \geq \beta_*^n = (\beta_*^{n_*-1})^{\frac{n}{n_*-1}} \geq \left((0.05)^{\frac{1}{50}}\right)^{\frac{50n}{n_*-1}} \geq 0.9^{\frac{50n}{n_*-1}},$$

where we use inequality (G.31). Thus, if $n \leq 0.01 n_*$, the Hellinger affinity is larger than 0.68, implying that the total variation is small and the probability of error is large. □

G.3.2 Robust Binary Hypothesis Testing

In this section, we provide the proofs of Theorem 10.4.6, Proposition 10.4.8, and the other claims that were omitted from Section 10.4.3. We begin by establishing Theorem 10.4.6.

Theorem 10.4.6 (Sample complexity of $\mathcal{B}_{\texttt{robust}}(p, q, \mathcal{T}_D)$)**.** *There exists a constant $c > 0$ such that for any $p, q \in \Delta_k$ with $\epsilon < \frac{d_{\mathrm{TV}}(p,q)}{2}$ and any $D \geq 2$, we have*

$$n^*_{\texttt{robust}}(p, q, \epsilon, \mathcal{T}_D) \leq c \cdot n^* \cdot \max\left\{1, \frac{\min\{k, \log n^*\}}{D}\right\}, \tag{10.16}$$

where $n^* := n^*_{\texttt{robust}}(p, q, \epsilon)$. *Furthermore, there is an algorithm which, given* p, q, ϵ, *and* D, *finds a channel* $\mathbf{T}^* \in \mathcal{T}_D^{\texttt{thresh}}$ *in* $\mathrm{poly}(k, D)$ *time that achieves the rate in inequality* (10.16).

Proof. Consider the setting without any communication constraints. Let $f(n, \phi, \tilde{p}, \tilde{q})$ be the probability of error by using the test ϕ on n i.i.d. samples under $\tilde{p}$ and $\tilde{q}$. Concretely, we define

$$f(n, \phi, \tilde{p}, \tilde{q}) := \mathop{\mathbb{P}}_{(x_1,\ldots,x_n)\sim\tilde{p}^{\otimes n}}(\phi(y_1, \ldots, y_n) \neq p) + \mathop{\mathbb{P}}_{(x_1,\ldots,x_n)\sim\tilde{q}^{\otimes n}}(\phi(y_1, \ldots, y_n) \neq q).$$

Huber [Hub65] showed that when the corruption is in total variation distance, there exist $p_1 \in \mathcal{P}_1$ and $q_1 \in \mathcal{P}_2$, called the least favorable distributions (LFDs), which maximize $\inf_\phi f(n, \phi, \tilde{p}, \tilde{q})$ over $\tilde{p}$ and $\tilde{q}$ in $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. Moreover, the optimal test is a likelihood ratio test, where the likelihoods are computed with respect to p_1 and q_1 (this corresponds to a clipped likelihood ratio test when likelihoods are computed with respect to p and q). Thus, the robust sample complexity is $n^* := n^*_{\texttt{robust}}(p, q, \epsilon) = \Theta(1/d^2_{\mathrm{h}}(p_1, q_1))$.

For the communication-constrained setting, Veeravalli, Basar, and Poor [VBP94] showed that p_1 and q_1 above are also LFDs for $\mathcal{B}_{\texttt{robust}}(p, q, \epsilon, \mathcal{T}_D)$. By applying Corollary 10.3.4 to p_1 and q_1, we see that there exists a threshold channel $\mathbf{T}$ such that $d^2_{\mathrm{h}}(\mathbf{T}p_1, \mathbf{T}q_1) \geq d^2_{\mathrm{h}}(p_1, q_1)$ up to a logarithmic factor. Let ϕ^* be the likelihood ratio test between $\mathbf{T}p_1$ and $\mathbf{T}q_1$. Applying Veeravalli, Basar, and Poor [VBP94, Theorem 1], we conclude that the probability of error for this test, for any $(\tilde{p}, \tilde{q}) \in \mathcal{P}_0 \times \mathcal{P}_1$, is less than the error on (p_1, q_1). Since the latter is less than 0.1 when $n \geq n_0 = \Theta(1/d^2_{\mathrm{h}}(\mathbf{T}p, \mathbf{T}q))$ by Fact 10.2.4, the sample complexity is at most n_0. As $\mathbf{T}$ is computed using Corollary 10.3.4 and $n^* = \Theta(1/d^2_{\mathrm{h}}(p_1, q_1))$, we have $n^*_{\texttt{robust}}(p, q, \epsilon, \mathcal{T}_D) \leq n^* \max\{1, \min\{k, \log n^*\}/D\}$.

Finally, we comment on the runtime of the algorithm. The LFDs can be calculated in polynomial time, as outlined in Huber [Hub65; HS73], and given these LFDs, the optimal channel $\mathbf{T}$ can again be computed in polynomial time, as mentioned in Corollary 10.3.4. □

The following result shows that the optimal channels are moderately robust, i.e., they are robust up to ϵ^2-corruption (up to logarithmic factors):

Proposition 10.4.8 (Optimal channels are moderately robust)**.** *Let* p *and* q *be two distributions over* $[k]$. *Define* $\epsilon_0 := c d^2_{\mathrm{TV}}(p, q) \cdot \min\left\{1, \frac{D}{\log(1/d_{\mathrm{TV}}(p,q))}\right\}$ *for a small enough constant* c.[43] *Let* $\mathbf{T}^*$ *be a channel that maximizes* $d^2_{\mathrm{h}}(\mathbf{T}p, \mathbf{T}q)$ *over* $\mathbf{T} \in \mathcal{T}_D$. *Let* n^*_D *be the sample complexity*

[43] This upper bound can be generalized to $\epsilon_0 := c d^2_{\mathrm{h}}(\mathbf{T}^*p, \mathbf{T}^*q)$.

of $\mathbf{T}^*$ for p and q (recall that $n_D^* = \Theta(n^*(p, q, \mathcal{T}_D))$). Let ϕ^* be the corresponding optimal test.[44] *Then there exists a test ϕ' that uses $\mathbf{T}^*$ for each user and solves $\mathcal{B}_{\texttt{robust}}(p, q, \epsilon_0, \mathcal{T}_D)$ with sample complexity $\Theta(n_D^*)$.*

Proof. Let $\tilde{p}$ and $\tilde{q}$ be arbitrary distributions satisfying $d_{\mathrm{TV}}(p, \tilde{p}) \le \epsilon_0$ and $d_{\mathrm{TV}}(q, \tilde{q}) \le \epsilon_0$. Suppose the following two conditions hold: (i) $n\epsilon_0 \le 0.01$, and (ii) $n \ge C/d_{\mathrm{h}}^2(\mathbf{T}^*p, \mathbf{T}^*q)$ for a large enough constant C. We will first demonstrate the existence of a test that works under these two conditions.

Sub-additivity of the total variation distance (cf. Fact 10.2.2) implies that $d_{\mathrm{TV}}(p^{\otimes n}, \tilde{p}^{\otimes n}) \le n\epsilon_0$ and $d_{\mathrm{TV}}(q^{\otimes n}, \tilde{q}^{\otimes n}) \le n\epsilon_0$. Under condition (i) above, the probability of each event $\mathcal{E}$ over the output of ϕ is the same under $p^{\otimes n}$ (similarly, $q^{\otimes n}$) and $\tilde{p}^{\otimes n}$ (similarly, $\tilde{q}^{\otimes n}$), up to an additive error of 0.01. Thus, if $(\phi^*, \mathcal{R})$ succeeds with probability 0.92 for p and q with at most n samples, they succeed with probability 0.9 under $\tilde{p}$ and $\tilde{q}$. The former condition holds under condition (ii), by Fact G.1.1. Thus, when both of these conditions hold simultaneously, the probability of error of $(\phi^*, \mathcal{R})$ under $\tilde{p}$ and $\tilde{q}$ is at most 0.1. These two conditions on n are satisfied if $n = n_0 := C/d_{\mathrm{h}}^2(\mathbf{T}^*, \mathbf{T}^*q)$ and $\epsilon \le 0.01/n_0$.

We now define $\phi' : \cup_{n=1}^{\infty} \mathcal{Y}^n \to \{p, q\}$, as follows: If $n < n_0$, define ϕ' arbitrarily; otherwise, discard $n - n_0$ samples[45] and define $\phi'(y_1, \ldots, y_n)$ to be $\phi^*(y_1, \ldots, y_{n_0})$. By our previous calculations, it follows that $(\phi', \mathcal{R})$ solves $\mathcal{B}_{\texttt{robust}}(p, q, \epsilon)$ as long as $\epsilon \le \epsilon_0 := 0.01/n_0 = \Theta(1/d_{\mathrm{h}}^2(\mathbf{T}^*p, \mathbf{T}^*q))$. By Corollary 10.3.4, we have $d_{\mathrm{h}}^2(\mathbf{T}^*p, \mathbf{T}^*q) \ge c d_{\mathrm{h}}^2(p, q)/\log(1/d_{\mathrm{h}}^2(p, q))$. Applying Fact 10.2.2, we obtain the desired result. □

We now provide another explicit example that shows that (i) the robust sample complexity has phase transitions with respect to the amount of corruption, and (ii) the sample complexity of Scheffe's test can be strictly suboptimal.

Example G.3.1 (Sample complexity of $\mathcal{B}_{\texttt{robust}}(p, q, \cdot)$)**.** *Let $0 < \alpha < \beta < \delta < 1$, satisfying $\delta < 2\beta - \alpha$. Let p and q be the following two distributions:*

$$p := \left(1/2 - 2\epsilon - \epsilon^{1+\alpha} + \epsilon^{1+\beta} - \epsilon^{1+\delta}, 1/2 + 2\epsilon, \epsilon^{1+\alpha} - \epsilon^{1+\beta}, \epsilon^{1+\delta}\right),$$
$$q := \left(1/2, 1/2 - \epsilon^{1+\alpha}, \epsilon^{1+\alpha}, 0\right),$$

where $\epsilon \le 0.01$. Note that $d_{\mathrm{TV}}(p, q) = \Theta(\epsilon)$. We have $n^*(p, q) = \Theta\left(1/\epsilon^{1+\delta}\right)$. For any fixed $\gamma > 0$, the sample complexity $n^*_{\texttt{robust}}(p, q, \epsilon^{1+\gamma})$ satisfies the following growth condition for ϵ

[44] The optimal test corresponds to a likelihood ratio test between $\mathbf{T}^*p$ and $\mathbf{T}^*q$.

[45] A more efficient strategy is to divide the samples into $\lfloor n/n_0 \rfloor$ buckets of size n_0 (discarding samples if necessary), apply ϕ^* on each of those buckets individually, and then output the median.

small enough, where we omit constant factors for brevity:

$$n^*_{\texttt{robust}}(p, q, \epsilon^{1+\gamma}) = \begin{cases} \frac{1}{\epsilon^{1+\delta}}, & \text{if } \gamma > \delta \\ \frac{1}{\epsilon^{1+2\beta-\alpha}}, & \text{if } \gamma \in (\beta, \delta) \\ \frac{1}{\epsilon^2}, & \text{if } \gamma \in (0, \beta). \end{cases}$$

By Theorem 10.4.6, the sample complexity under communication constraints of $D = 2$ messages satisfies the same result up to constants (note that the optimal channel may change with respect to γ). On the other hand, for all $\gamma > 0$, the sample complexity of Scheffe's test for $\mathcal{B}_{\texttt{robust}}(p, q, \epsilon^{1+\gamma})$ is $\Theta(1/\epsilon^2)$.

Proof. Note that

$$d^2_{\mathrm{h}}(p, q) = \Theta\left(\epsilon^2 + \epsilon^{2+2\beta-1-\alpha} + \epsilon^{1+\delta}\right) = \Theta\left(\epsilon^{1+2\beta-\alpha} + \epsilon^{1+\delta}\right) = \Theta\left(\epsilon^{1+\delta}\right),$$

since $\delta < 2\beta - \alpha$, which leads to the claim on $n^*(p, q)$ by Fact 10.2.4.

The lower bounds on $n^*_{\texttt{robust}}(p, q, \epsilon^{1+\gamma})$ follow by applying Fact 10.2.4 on the following choices of $\tilde{p}$ and $\tilde{q}$, which lie within $\epsilon^{1+\gamma}$ in total variation distance:

1. $\gamma > \delta$: This follows directly by choosing $\tilde{p} = p$ and $\tilde{q} = q$.
2. $\gamma \in (\beta, \delta)$: This follows by choosing $\tilde{p} = p$ and $\tilde{q} = (1/2 - \epsilon^{1+\delta}, 1/2 - \epsilon^{1+\alpha}, \epsilon^{1+\alpha}, \epsilon^{1+\delta})$.
3. $\gamma \in (0, \beta)$: This follows by choosing $\tilde{p} = p$ and

$$\tilde{q} = (1/2 - \epsilon^{1+\delta} - \epsilon^{1+\beta}, 1/2 - \epsilon^{1+\alpha}, \epsilon^{1+\alpha} - \epsilon^{1+\beta}, \epsilon^{1+\delta}).$$

We now discuss the channels that achieve the upper bound. We will choose $\mathbf{T}$ corresponding to $\mathbb{I}_A(\cdot)$, as follows:

1. $\gamma > \delta$: Take $A = \{4\}$. Then $\mathbb{E}_{\tilde{p}}(A) \geq 2\epsilon^{1+\delta}/3$ and $\mathbb{E}_{\tilde{q}}(A) \leq \epsilon^{1+\delta}/3$. As mentioned later, this can be tested with $O(1/\epsilon^{1+\delta})$ samples.
2. $\gamma \in (\beta, \delta)$: Take $A = \{3\}$. Then $\mathbb{E}_{\tilde{p}}(A) \leq \epsilon^{1+\alpha} - 2\epsilon^{1+\beta}/3$ and $\mathbb{E}_{\tilde{q}}(A) \geq \epsilon^{1+\alpha} - \epsilon^{1+\beta}/3$. As mentioned later, this can be tested with $O(1/\epsilon^{1+2\beta-\alpha})$ samples.
3. $\gamma \in (0, \beta)$: This follows by taking $A = \{2\}$ and using similar arguments as above.

Finally, we prove that the sample complexity of Scheffe's test is $\Theta(1/\epsilon^2)$. Scheffe's test transforms p and q to Bernoulli distributions, with probabilities of observing 1 equal

to $\left(1/2 + 2\epsilon + \epsilon^{1+\delta}\right)$ and $(1/2 - \epsilon^{1+\alpha})$, respectively. It is easy to see that the Hellinger distance between these two Bernoulli distributions is $\Theta(\epsilon^2)$, implying that the sample complexity of Scheffe's test is $\Theta(1/\epsilon^2)$. Its robustness to $\epsilon^{1+\gamma}$-corruption follows from similar arguments as above.

For completeness, we outline the typical concentration argument that is needed to perform the tests above. Let X be a mean of i.i.d. indicator random variables, i.e., $X = (\sum_i Y_i)/n$, where $Y_i \in \{0, 1\}$ and $\mathbb{E}[Y_i] = \mu \leq 1/2$. Then $\mathbb{E}[X] = \mu$ and $\mathbf{Var}(X) \leq \mu/n$. Chebyshev's inequality implies that with probability 0.01, we have $X \in [\mu - 10\sqrt{\mu/n}, \mu + 10\sqrt{\mu/n}]$. Thus, if $n \geq 10^4/\mu$, then with probability 0.01, we have $X \in [2\mu/3, 4\mu/3]$. By similar logic, if $n \geq 10^4/\delta^2$, then with probability 0.01, we have $X \in [\mu - \delta, \mu + \delta]$. □

Finally, we provide additional details regarding Example 10.4.5 below:

Details regarding Example 10.4.5. Consider the set $A = \{2\}$. We have $p(A) = 0.5 + 3\epsilon$ and $q(A) = 0.5$. Any valid $\tilde{p}$ and $\tilde{q}$ lying within ϵ in total variation distance of p and q, respectively, satisfy $\tilde{p}(A) \geq 0.5 + 2\epsilon$ and $\tilde{q}(A) \leq 0.5 + \epsilon$. Thus, estimating the mean of $\mathbb{I}_A(X)$ up to error $\epsilon/2$ gives a valid test, which takes $O(1/\epsilon^2)$ samples by the arguments outlined above. The lower bound follows by applying Fact 10.2.4 to $\mathcal{B}(\tilde{p}, \tilde{q})$, where $\tilde{p} = (0.5 - 3\epsilon, 0.5 + 3\epsilon, 0)$ and $\tilde{q} = (0.5, 0.5, 0)$. It can be seen that for this choice of $\tilde{p}$, we have $\mathbf{T}^*\tilde{p} = (1, 0)$ and $\mathbf{T}^*q = (1, 0)$.

G.4 Upper Bounds for M-ary Hypothesis Testing

In this section, we prove upper bounds on the M-ary hypothesis testing problem under communication constraints in various settings: Appendix G.4.1 focuses on non-identical channels (both adaptive and non-adaptive) and Appendix G.4.2 focuses on identical channels.

G.4.1 A Tournament Procedure Using a Binary Test

We prove Proposition 10.5.2 below:

Proposition 10.5.2 (Upper bounds using threshold tests)**.** *Let $\mathcal{P}$ be set of M distributions in Δ_k such that $\rho = \min_{p,q\in\mathcal{P}:p\neq q} d_{\mathrm{h}}(p, q)$. Let $k' = \log(1/\rho)$ and define the blow-up factor $R := \frac{\min\{k,\log(1/\rho)\}}{D} + 1$. Then the sample complexity of the simple M-ary hypothesis testing problem satisfies the bounds*

1. $n^*_{\texttt{non-identical}}(\mathcal{P}, \mathcal{T}_D) \lesssim \frac{M^2 \log M}{\rho^2} \cdot R$,
2. $n^*_{\texttt{adaptive}}(\mathcal{P}, \mathcal{T}_D) \lesssim \frac{M \log M}{\rho^2} \cdot R$.

Proof. Let $\mathcal{P} = \left\{p^{(1)}, \ldots, p^{(M)}\right\}$. We first prove the results for non-adaptive, non-identical channels. Denote the set $\mathcal{S} = \{\{i, j\} : i \neq j, i \in [M], j \in [M]\}$. For each $\{i, j\} \in \mathcal{S}$, let $\mathbf{T}_{\{i,j\}} \in \mathcal{T}_D^{\texttt{thresh}}$ be the channel achieving the guarantee in Corollary 10.3.4. Since $d^2_{\mathrm{h}}(p^{(i)}, p^{(j)}) \geq \rho^2$, Corollary 10.3.4 states that $d^2_{\mathrm{h}}(\mathbf{T}_{\{i,j\}} p^{(i)}, \mathbf{T}_{\{i,j\}} p^{(j)}) \geq \rho^2/R$. Let $m = (CR \log M)/\rho^2$, for a large enough constant $C > 0$ to be decided later.

Fix any ordering $\sigma(\cdot)$ of the set $\mathcal{S}$. Consider the strategy where we take a total of $0.5M(M-1)m$ users, such that r^{th} user uses the channel $\mathbf{T}_{\sigma(\lceil r/m \rceil)}$, i.e., each channel is repeated m times in a predetermined order. For any $\{i, j\} \in \mathcal{S}$, let $\mathcal{A}_{\{i,j\}}$ denote the set of samples observed by the central server after passing through $\mathbf{T}_{\{i,j\}}$.

We now describe the strategy at the central server: For any $\{i, j\} \in \mathcal{S}$, consider the optimal test $\psi_{\{i,j\}}$ between $\left\{\mathbf{T}_{\{i,j\}} p_i, \mathbf{T}_{\{i,j\}} p_j\right\}$ that uses the samples $\mathcal{A}_{\{i,j\}}$ and maps to either $\{i\}$ or $\{j\}$. We say this is a game between i and j, and call $\psi_{\{i,j\}}\left(\mathcal{A}_{\{i,j\}}\right)$ the winner of the game. The central server outputs the unique hypothesis that wins all of its games against other hypotheses, i.e., the unique element in the set $\{i : \forall j \neq i, \psi_{\{i,j\}}\left(\mathcal{A}_{\{i,j\}}\right) = i\}$.

Let $i \in [M]$ be the unknown true hypothesis. It suffices to show that i never loses a game against any other hypothesis. For any $j \neq i$, we have $d^2_{\mathrm{h}}(\mathbf{T}_{\{i,j\}} p^{(i)}, \mathbf{T}_{\{i,j\}} p^{(j)}) \geq \rho^2/R$. Thus, we have $\mathbb{P}(\psi(\mathcal{A}_{\{i,j\}}) \neq i) \leq 0.01/M^2$ by Fact G.1.1, since C is large enough. Taking a union bound over all $j \neq i$, we see that the probability of error is less than $0.01/M$. Taking the sum, we see that the sum of the probabilities of errors satisfies condition (10.3). Thus, we obtain $n^*_{\texttt{non-identical}}(\mathcal{P}, \mathcal{T}_D) \lesssim M^2 m \lesssim \frac{M^2 \log M}{\rho^2} \cdot R$.

We now turn our attention to the adaptive setting. Consider the following strategy:

1. Set $j = 2$ and $\widehat{i} = 1$.
2. While $j \leq M$:
   a) m users choose $\mathbf{T}_{\{j,\widehat{i}\}}$.
   b) Let $\mathcal{A}_{\{j,\widehat{i}\}}$ be the set of m observed samples.
   c) Assign $\widehat{i} \leftarrow \psi_{j,\widehat{i}}(\mathcal{A}_{\{j,\widehat{i}\}})$ and $j \leftarrow j + 1$.
3. Output $\widehat{i}$.

First, it is easy to see that the procedure terminates after taking Mm samples. Turning to the correctness of the algorithm, let i^* be the true unknown probability distribution.

It suffices to show that i^* never loses a game against any other j. The same arguments as above show that this does not happen. □

G.4.2 Upper Bounds for Identical Channels

This section contains the proof of Lemma 10.5.6 that was omitted from Section 10.5.

Lemma 10.5.6 (JL-sketch)**.** *There exists a constant $c > 0$ such that the following holds: Let $\left\{p^{(1)}, \ldots, p^{(M)}\right\} \subseteq \Delta_k$ be M distributions such that $\min_{i \neq j} d_{\mathrm{TV}}(p^{(i)}, p^{(j)}) > \epsilon$. Then*

$$\max_{\mathbf{T} \in \mathcal{T}_D} \min_{i \neq j} d_{\mathrm{TV}}(\mathbf{T}p^{(i)}, \mathbf{T}p^{(j)}) \geq c \cdot \frac{\epsilon}{\sqrt{k} M^{\frac{2}{D-1}} \sqrt{D \log(Dk)}}.$$

Proof. Let $D' := D - 1$. Consider a matrix $\mathbf{H} \in \mathbb{R}^{D' \times k}$ that satisfies the following constraints: $H_{i,j} \geq 0$ for all (i, j), and $\sum_{i=1}^{D'} H_{i,j} \leq 1$ for all $j \in [k]$. Let $\mathcal{H}_{D'}$ be the set of all such matrices. It is easy to see that given any matrix $\mathbf{H} \in \mathcal{H}_{D'}$, it is possible to generate a unique matrix $\mathbf{T} \in \mathcal{T}_D$ by adding an extra row to make the column sums 1. Consider such pairs $(\mathbf{H}, \mathbf{T})$ in $\mathcal{H}_{D'} \times \mathcal{T}_D$, and note that $\left\|\mathbf{H}p^{(i)} - \mathbf{H}p^{(j)}\right\|_1 \leq \left\|\mathbf{T}p^{(i)} - \mathbf{T}p^{(j)}\right\|_1 = 0.5 d_{\mathrm{TV}}\left(\mathbf{T}p^{(i)}, \mathbf{T}p^{(j)}\right)$. We will generate $\mathbf{H}$ randomly such that, with positive probability, it belongs to $\mathcal{H}_{D'}$ and $\min_{i \neq j} \left\|\mathbf{H}p^{(i)} - \mathbf{H}p^{(i)}\right\|_1$ is large.

We will show the following result:

Lemma G.4.1. *There exists a constant $c > 0$ such that, for any $A = \{a_1, \ldots, a_N\} \subseteq \mathbb{R}^k$ such that the sum of the components of each a_i is equal to 0, there is a linear map $\mathbf{H} \in \mathcal{H}_{D'}$ such that the following holds:*

$$\|\mathbf{H}a\|_2 \geq c\|a\|_2 \frac{1}{\sqrt{D' \log(D'k)}} N^{-\frac{1}{D'}}, \qquad \forall a \in A.$$

Before giving the proof of Lemma G.4.1, we show how to use it to complete the proof of Lemma 10.5.6. Let $A = \left\{p^{(i)} - p^{(j)} : 1 \leq i < j \leq M\right\}$, and observe that A satisfies the conditions of Lemma G.4.1 with $N \leq M^2$. Using the $\mathbf{H}$ in Lemma G.4.1 and the fact that for $x \in \mathbb{R}^p$, we have $\|x\|_1 \geq \|x\|_2 \geq \|x\|_1/\sqrt{p}$, we obtain

$$\begin{aligned}\left\|\mathbf{H}\left(p^{(i)} - p^{(j)}\right)\right\|_1 \geq \left\|\mathbf{H}\left(p^{(i)} - p^{(j)}\right)\right\|_2 &\geq 2c \cdot \left\|\left(p^{(i)} - p^{(i)}\right)\right\|_2 \cdot \frac{1}{\sqrt{D \log(Dk)}} M^{-\frac{2}{D'}} \\ &\geq 2c \cdot \left\|\left(p^{(i)} - p^{(i)}\right)\right\|_1 \cdot \frac{1}{\sqrt{k}} \frac{1}{\sqrt{D \log(Dk)}} M^{-\frac{2}{D'}}\end{aligned}$$

$$\geq c \cdot \frac{\epsilon}{\sqrt{k}} \frac{1}{\sqrt{D \log(Dk)}} M^{-\frac{2}{D'}},$$

where we use the fact that $d_{\mathrm{TV}}\left(p^{(i)}, p^{(j)}\right) = \frac{1}{2}\left\|p^{(i)} - p^{(j)}\right\|_1$. This completes the proof of Lemma 10.5.6. □

We now provide the proof of Lemma G.4.1 that was omitted above.

Proof. (Proof of Lemma G.4.1) Let $Q_1, Q_2 > 0$ be numbers to be determined later. Let $\mathbf{J} \in \mathbb{R}^{D' \times k}$ be the matrix of all ones, and let $\mathbf{G} \in \mathbb{R}^{D' \times k}$ be a matrix with i.i.d. $\mathcal{N}(0,1)$ entries $\{G_{i,j}\}$. We will choose $\mathbf{H}$ to be of the following form:

$$\mathbf{H} := \frac{1}{Q_1}\left(\mathbf{J} + \frac{\mathbf{G}}{Q_2}\right).$$

The following claim shows that with probability at least 0.9, we have $\mathbf{H} \in \mathcal{H}_{D'}$ (the proof is given later).

Claim G.4.2. *If $Q_2 \geq 10\sqrt{\log(kD')}$ and $Q_1 \geq D' + 10\sqrt{D' \log k}/Q_2$, then with probability at least $9/10$, we have $\mathbf{H} \in \mathcal{H}_{D'}$.*

We will now show that with high probability, $\mathbf{H}$ preserves the Euclidean norm of each $a \in A$:

Claim G.4.3. *There exists a constant $c' > 0$ such that with probability at least $9/10$, we have*

$$\|\mathbf{H}a\|_2 \geq c' \frac{\|a\|_2}{Q_1 Q_2} \sqrt{D'} N^{-\frac{1}{D'}}, \qquad \forall a \in A.$$

Given Claims G.4.2 and G.4.3, if we choose $Q_2 = 10\sqrt{\log(kD')}$ and $Q_1 = 11D'$, then with probability at least 0.8, we have $\mathbf{H} \in \mathcal{H}_{D'}$, and for all $a \in A$, we have

$$\|\mathbf{H}a\|_2 \geq c' \|a\|_2 \frac{1}{\sqrt{\log(Dk)}} \sqrt{D} N^{-\frac{1}{D'}}.$$

This completes the proof of Lemma G.4.1.

We now provide the proofs of intermediate results that we have used.

Proof. (Proof of Claim G.4.2) We need to ensure that the following two events hold simultaneously:

$$\mathcal{E}_1 := \{\forall (i,j) \in [D'] \times [k] : \ H_{i,j} \geq 0\},$$

$$\mathcal{E}_2 := \left\{ \forall j \in [k] : \sum_{i=1}^{D'} H_{i,j} \leq 1 \right\}.$$

The event $\mathcal{E}_1$ holds when for all (i,j), we have $1+G_{i,j}/Q_2 \geq 0$ if and only if $G_{i,j} \geq -Q_2$. Thus, it suffices to take $Q_2 > \sup_{i\in[D'],j\in[k]} |G_{i,j}|$. Taking $Q_2 = 10\sqrt{\log(kD')}$, standard results on maxima of Gaussian random variables [Wai19] imply that $\mathcal{E}_1$ holds with probability at least 0.95.

We now focus on $\mathcal{E}_2$. For $j \in [k]$, letting $Z_j := \sum_{i=1}^{D'} H_{i,j} = \sum_{i=1}^{D'} \frac{1}{Q_1}(1 + G_{i,j}/Q_2)$, we have $Z_j - \frac{D'}{Q_1} \sim \mathcal{N}\left(0, \frac{D'}{Q_1^2 Q_2^2}\right)$. A union bound then implies that with probability 0.95, for all $Z_j \in [k]$, we have

$$Z_j \leq \frac{D'}{Q_1} + 10\sqrt{\frac{D'}{Q_1^2Q_2^2}}\sqrt{\log(k)} = \frac{1}{Q_1}\left(D' + 10\frac{\sqrt{D'\log k}}{Q_2}\right).$$

For this to be at most 1, we need $Q_1 \geq D' + 10\frac{\sqrt{D'\log k}}{Q_2}$. By a union bound, the events $\mathcal{E}_1$ and $\mathcal{E}_2$ hold simultaneously with probability at least 0.9. This completes the proof. □

Proof. (Proof of Claim G.4.3) Without loss of generality, we will assume that $\|a\|_2 = 1$. It suffices to show that for all $a \in A$, with probability at least $1 - \frac{1}{10N}$, we have $\|\mathbf{H}a\|_2 \geq c'\frac{1}{Q_1Q_2}\sqrt{D'}N^{-\frac{1}{D'}}$. Equivalently, we will show that $\mathbb{P}\left(\|\mathbf{H}a\|_2 < c'\frac{1}{Q_1Q}\sqrt{D'}N^{-\frac{1}{D'}}\right) \leq \frac{1}{10N}$. For any $a \in A$, we note that $\mathbf{J}a$ is a zero vector, so

$$\mathbf{H}a = \frac{1}{Q_1}\left(\mathbf{J} + \frac{\mathbf{G}}{Q_2}\right)a = \frac{1}{Q_1}\left(\mathbf{J}a + \frac{\mathbf{G}a}{Q_2}\right) = \frac{1}{Q_1Q_2}\mathbf{G}a.$$

Letting $G_1, \ldots, G_{D'}$ be the rows of $\mathbf{G}$, and letting $\chi^2_{D'}$ be a chi-square random variable with D' degrees of freedom, we have

$$\|\mathbf{H}a\|_2^2 = \left(\frac{1}{Q_1Q_2}\right)^2 \sum_{i=1}^{D'} (G_i^\top a)^2 \sim \left(\frac{1}{Q_1Q_2}\right)^2 \chi^2_{D'},$$

since G_i is an isotropic multivariate Gaussian and a has unit norm. Standard approximations for $\chi^2_{D'}$[46] imply that

$$\mathbb{P}\left(\chi^2_{D'} \leq t\right) \leq \left(\frac{et}{D'}\right)^{D'/2}.$$

[46]This can be obtained by upper-bounding the pdf of the χ^2 random variable and using Stirling's approximation.

Furthermore, for $t_* = \frac{D'}{e}\left(\frac{1}{10N}\right)^{\frac{2}{D'}}$, the expression on the right-hand side is less than $0.1/N$. Then

$$\mathbb{P}\left(\|\mathbf{H}a\|_2^2 \le \left(\frac{1}{Q_1Q_2}\right)^2 t_*\right) = \mathbb{P}\left\{\chi^2_{D'} \ge t_*\right\} \le \frac{1}{10N}.$$

Thus, with probability at least $1 - \frac{1}{10N}$, we have

$$\|\mathbf{H}a\| \ge \sqrt{\frac{1}{e}}\left(\frac{1}{10}\right)^{\frac{1}{D'}} \cdot \sqrt{D'}N^{-\frac{1}{D'}} \cdot \frac{1}{Q_1Q_2},$$

completing the proof of the claim with $c' = \frac{1}{\sqrt{e}10^{1/D'}} \ge 0.001$. □

This completes the proof of Lemma G.4.1. □

G.5 Lower Bounds for M-ary Hypothesis Testing

In this section, we provide the proof of Theorem 10.5.8. We prove the two bounds in Theorem 10.5.8 separately: the $\Omega(M)$ lower bound from the strong data processing inequality is proved in Appendix G.5.1 (see Corollary G.5.4), and the $\Omega(M^{1/3})$ lower bound from the SQ lower bound is proved in Appendix G.5.2 (see Corollary G.5.9). Finally, we prove a $\Omega(\sqrt{M})$ lower bound for non-adaptive, non-identical channels from the impossibility of ℓ_1-embedding in Appendix G.5.3 (see Theorem G.5.12).

In this section, we abuse notation by using p_1, p_2, etc., and P_1, P_2, etc., to denote different probability distributions.

G.5.1 Strong Data Processing

Preliminaries: We will closely follow the terminology of Braverman, Garg, Ma, Nguyen, and Woodruff [BGMNW16], to which we refer the reader for more details. Let $\mathcal{Q} = \{Q_0, Q_1\}$ be two distributions on $\mathcal{X}$. For any $i \in [M]$, we define P_i to be the distribution over $(Z_1, \ldots, Z_M)$, where the Z_j's are independent, and $Z_j \sim Q_0$ for $j \ne i$ and $Z_i \sim Q_1$. We use P_0 to denote the distribution $Q_0^{\otimes n}$. We define $\mathcal{P}_M = \{P_0, P_1, \ldots, P_M\}$. Our goal is to perform statistical estimation using n machines. The model generation process is as follows: V is sampled uniformly from $\{0, \ldots, M\}$. Conditioned on $V = v$, for each $j \in [n]$, machine j receives an i.i.d. sample X_j from the distribution P_v. When it is clear from context, we will use X as shorthand for $(X_1, \ldots, X_n)$.

We will work in the blackboard protocol. Here, all machines simultaneously write the first iteration of their messages on a "blackboard," and the subsequent iterations of messages are on subsequent blackboards and may depend on the contents of all past blackboards. The combined content (in bits) of all blackboards is called the transcript of the protocol, which is denoted by Π. The blackboard protocol is also called the "fully adaptive" or simply "adaptive" protocol. Since it imposes the fewest constraints on permitted actions, lower bounds proved for this protocol are valid for other protocols, as well. A special case of interest is the "sequentially adaptive" protocol, where machines communicate in a fixed order, with subsequent messages allowed to depend on past messages. In the communication-constrained setting considered in this paper, we restrict the size of the transcript $|\Pi|$ to be at most $n \log D$, as each machine is permitted to send at most D messages ($\log D$ bits).

The estimator $\widehat{v}$ maps each transcript Π to an element of $\mathcal{P}_M$. The failure probability is then defined as $R(\Pi, \hat{v}, \mathcal{P}_M) := \max_{v \in \{0,1,\ldots,M\}} \Pr[\hat{v}(\Pi) \neq v | V = v]$. We use $T(n, \mathcal{P}_M)$ to denote the task of hypothesis testing among the distributions in $\mathcal{P}_M$ with n machines, and we say that $(\Pi, \widehat{v})$ solves $T(n, \mathcal{P}_M)$ if the protocol works on n machines and $R(\Pi, \hat{v}, \mathcal{P}_M) \leq 0.1$. We will use the definitions of IC(Π) (the information cost of Π) and min-IC(Π) (the minimum information cost of Π) from Braverman, Garg, Ma, Nguyen, and Woodruff [BGMNW16].

Lemma G.5.1 (Direct-sum for multiple hypothesis testing [BGMNW16])**.** *Let $M \geq 1$, and let $\mathcal{Q}$ and $\mathcal{P}_M$ be defined as above. If there exists a protocol estimator pair $(\Pi, \hat{v})$ that solves the detection task $T(m, \mathcal{P}_M)$ with information cost I, then there exists a protocol estimator pair $(\Pi', \hat{v}')$ that solves the detection task $T(m, \mathcal{Q})$ with minimum information cost I' satisfying $I' \lesssim \frac{I}{M}$.*

For the set of two distributions, we use the following hardness result:

Lemma G.5.2. *There exists a constant $c \geq 1$ such that for every $\beta \in (0, 1)$, there exist two distributions $\mathcal{Q} = \{Q_0, Q_1\}$ such that any $(\Pi, \widehat{v})$ that solves $T(m, \mathcal{Q})$ with failure probability at most $1/4$ for any m satisfies* min-IC$(\Pi) \geq \frac{c}{\beta}$*. Moreover, $d_{\mathrm{TV}}(Q_0, Q_1) = \Theta(\sqrt{\beta})$.*

Proof. For two distributions p and q, we use $\beta(p, q)$ to denote the SDPI constant, as defined in Braverman, Garg, Ma, Nguyen, and Woodruff [BGMNW16]. We will use the following result, which shows that if p and q have bounded likelihood ratios, then the SDPI constant is small:

Lemma G.5.3 ([DJWZ14])**.** *Let $\beta \in (0,1)$. If two Bernoulli distributions with parameters p and q satisfy*

$$e^{-\sqrt{\beta}}p \le q \le e^{\sqrt{\beta}}p,$$
$$e^{-\sqrt{\beta}}(1-p) \le 1-q \le e^{\sqrt{\beta}}(1-p),$$

then $\beta(p,q) \le (2e^2)\beta$.

Let Q_0 and Q_1 be binary distributions with probabilities of observing 1 equal to q_0 and q_1, respectively. Set $q_0 = 1/2$ and $q_1 = e^{-\sqrt{\beta}}/2$. Then

$$\frac{q_0}{q_1} = e^{\sqrt{\beta}} \quad \text{and} \quad \frac{1-q_0}{1-q_1} \le \frac{1}{2-e^{-\sqrt{\beta}}},$$

and both ratios lie between $e^{-\sqrt{\beta}}$ and $e^{\sqrt{\beta}}$. Thus, we have $\beta(\mu_0,\mu_1) \lesssim \beta$ from Lemma G.5.3.

Fix a constant $c' \le 0.1$. Let the protocol be Π. Since the protocol is successful, using Facts 10.2.2 and 10.2.4, we have $d_{\mathrm{h}}^2(\Pi|_{V=0}, \Pi|_{V=1}) \ge c'$, for a constant c'. Applying Braverman, Garg, Ma, Nguyen, and Woodruff [BGMNW16, Theorem 1.1], we obtain

$$d_{\mathrm{h}}^2(\Pi|_{V=0}, \Pi|_{V=1}) \le c\beta \cdot \text{min-IC}(\Pi),$$

which yields the desired conclusion. Finally, the bound on total variation follows from direct calculation and the fact that $\beta \in (0,1)$. □

Combining Lemmata G.5.1 and G.5.2, we obtain the following result:

Corollary G.5.4. *For every $\epsilon \in (0,1)$, there exist $M+1$ distributions $\{P_0, \dots, P_M\}$ such that (i) $d_{\mathrm{TV}}(p,q) \ge \epsilon$ for all $p \ne q$ in $\mathcal{P}_M$, and (ii) $n^*_{\texttt{adaptive}}(\mathcal{P}_M, \mathcal{T}_D) \gtrsim \frac{M}{\epsilon^2 \log D}$.*

Proof. Let $\mathcal{Q} = \{Q_0, Q_1\}$ be the two distributions from Lemma G.5.2 such that $d_{\mathrm{TV}}(Q_0,Q_1) \ge \epsilon$ and every successful protocol Π for $T(n,\mathcal{Q})$ satisfies min-IC$(\Pi) \ge \frac{c}{\epsilon^2}$. Construct $\mathcal{P}_M$ as defined above using $\mathcal{Q}$. It can be seen that $d_{\mathrm{TV}}(p,q) \ge \epsilon$ for any distinct p and q in $\mathcal{P}_M$. Suppose there exists a successful protocol $\widehat{\Pi}$ for $T(n^*, \mathcal{P}_M)$ with each machine sending at most $\log D$ bits. Then we have

$$I = \sup_{v \in [M+1]} I_v\left(\widehat{\Pi}; X | R_{\text{pub}}\right) \le \sup_{v \in [M+1]} h(\widehat{\Pi}) \le n^* \log D,$$

where $h(\widehat{\Pi})$ denotes the entropy of the transcript $\widehat{\Pi}$. Thus, Lemma G.5.1 implies that there exists a successful protocol $\widehat{\Pi}'$ for $T(n^*, \mathcal{Q})$ such that min-IC$(\widehat{\Pi}') \lesssim \frac{n^* \log D}{M}$. However, we have min-IC$(\widehat{\Pi}') \gtrsim 1/\epsilon^2$. Thus, we obtain $\frac{1}{\epsilon^2} \lesssim \frac{n^* \log D}{M}$, or equivalently, $n^* \gtrsim \frac{M}{\epsilon^2 \log D}$. This completes the proof of Corollary G.5.4 and the proof of the first claim in Theorem 10.5.8. □

G.5.2 SQ Lower Bounds

Preliminaries: We will use the standard notations from the statistical query (SQ) complexity literature [FGRVX17; Fel17]. In particular, we will use the following oracles: STAT(τ), VSTAT(t), and 1-MSTAT(D).

For two square-integrable functions $f, g : \mathcal{X} \to \mathbb{R}$ and a distribution P on $\mathcal{X}$, we define $\langle f, g \rangle_P := \mathbb{E}_P[f(X)g(X)]$. For a distribution P, we will abuse notation by using P to refer to both the distribution and its pmf. For two distributions P_1 and P_2, their pairwise correlation with respect to the base measure P is defined as

$$\chi_P(P_1, P_2) := \left| \left\langle \frac{P_1}{P} - 1, \frac{P_2}{P} - 1 \right\rangle_P \right| = \left| \left\langle \frac{P_1}{P}, \frac{P_2}{P} \right\rangle_P - 1 \right|.$$

The *average correlation* of a set of distributions $\mathcal{P}'$ relative to a distribution P is denoted by $\rho(\mathcal{P}', P)$ and defined as $\rho(\mathcal{P}', P) := \frac{1}{|\mathcal{P}'|^2} \sum_{P_1, P_2 \in \mathcal{P}'} \chi_P(P_1, P_2)$.

Definition G.5.5 (Decision problem)**.** *Let P be a fixed distribution and $\mathcal{P}$ a set of distributions which does not contain P. Given access to the input distribution Q, which either equals P or belongs to $\mathcal{P}$, the goal is to identify whether $Q = P$ or $Q \in \mathcal{P}$. We refer to this problem as $\mathcal{B}_S(\mathcal{P}, P)$.*

We will use $\mathrm{SDA}(\mathcal{B}_S(\mathcal{P}, P), \bar{\gamma})$ to denote the average statistical dimension with average $\bar{\gamma}$ of the decision problem $\mathcal{B}_S(\mathcal{P}, P)$ [FGRVX17, Definition 3.6]. Although our main focus will be the STAT oracle and blackboard protocol, we also mention hardness results for VSTAT and 1-MSTAT oracles, which follow from SDA.

Theorem G.5.6 ([FGRVX17, Theorem 3.7], [FPV18, Theorem 7.3])**.** *Let P be a distribution and $\mathcal{P}$ be a set of distributions over a domain X, such that $\mathrm{SDA}(\mathcal{B}_S(\mathcal{P}, P), \bar{\gamma}) = d$ for some $\bar{\gamma}$. Any (randomized) SQ algorithm that solves $\mathcal{B}_S(\mathcal{P}, P)$ with success probability $9/10$ must satisfy at least one of the following conditions:*

(i) performs $0.8d$ queries,

(*ii*) *requires a single query to VSTAT*$(1/3\bar{\gamma})$*, or*

(*iii*) *requires a single query to STAT*$(\sqrt{3\bar{\gamma}})$*.*

In particular, for any L*, any (randomized) SQ algorithm that solves* $\mathcal{B}_S(\mathcal{P}, P)$ *with success probability* $9/10$ *requires at least* m *calls to 1-MSTAT*(L)*, where* $m = \Omega\left(\frac{1}{L}\min\left\{d, \frac{1}{\bar{\gamma}}\right\}\right)$*.*

Steinhardt, Valiant, and Wager [SVW16] show that an SQ lower bound also implies a lower bound for blackboard protocols:

Theorem G.5.7 (Lower bounds for blackboard communication using SQ algorithms [SVW16, Proposition 3], [Fel17, Section B.1])**.** *Let* $\mathcal{B}_S(\mathcal{P}, P)$ *be a decision problem that can be solved with probability* 0.95 *by a communication-efficient algorithm that extracts at most* b *bits from each of* m *machines. Then* $\mathcal{B}_S(\mathcal{P}, P)$ *can be solved by an SQ algorithm, with probability at least* 0.9*, which uses at most* $2bm$ *queries of STAT with tolerance* $\tau = O\left(\frac{1}{2^b m}\right)$*. In particular, for some* $\bar{\gamma}$*, let* $d = \mathrm{SDA}(\mathcal{B}_S(\mathcal{P}, P), \bar{\gamma})$*. Then either* $\frac{1}{2^b m} \lesssim \sqrt{\bar{\gamma}}$ *or* $d \leq 2bm$*.*

We now describe a distribution family that has a large statistical dimension, on average.

Lemma G.5.8 (A decision problem with large SQ dimension)**.** *Let* $r \in \mathbb{N}$ *and fix an* $\epsilon \in (0, 1)$*. For any* $M = 2^r$*, there exist distributions* $\mathcal{P}_M := \{P_1, \ldots, P_M\} \subseteq \Delta_{M+1}$ *and* $P \in \Delta_M$ *such that* $\chi_P(P_i, P_j) = \mathbb{I}_{i=j}$ *for all* (i, j)*. In particular,* $\mathrm{SDA}\left(\mathcal{B}_S(\mathcal{P}_M, P), \bar{\gamma}\right) \geq \frac{M\bar{\gamma}}{\epsilon^2}$ *for any* $\bar{\gamma} \leq \epsilon^2$*. Moreover, for any two distinct* $p, q \in \mathcal{P}_M \cup \{P\}$*, we have* $d_{\mathrm{TV}}(p, q) \geq 0.01\epsilon$*.*

Proof. Let $k = M + 1$. Let $V = [v_1, \ldots, v_k] \in \mathbb{R}^{k\times k}$ be the Walsh-Hadamard matrix. We have $V = V^\top$ and $\langle v_i, v_j\rangle = k\,\mathbb{I}_{i=j}$. Furthermore, we have $v_i \in \{-1, 1\}^k$, where v_1 has all entries 1, and for $i > 1$, v_i has half positive entries and half negative entries. Define u to be the uniform distribution in Δ_k, and define e_i to be the distribution that places all its mass on the i^{th} element. We also write $v_i = \sum_{j=1}^k v_{i,j}e_j$. Moreover, for $i \neq j$, we have $\|v_i - v_j\|_1 \geq 0.1k$ [Hor07].

Define $P = u$, and for $m \in [M]$, define $P_m = u + \epsilon(v_{m+1}/k)$. Note that the P_m's are valid distributions. For notational purposes, we will also use $P_m(i)$ to denote the probability of element $i \in [k]$ under P_m, i.e., $P_m(i) = \frac{1}{k}(1 + \epsilon v_{m,i})$. Using the lower bound on $\|v_i - v_j\|_1$, we have $d_{\mathrm{TV}}(P, P_i) = \epsilon/2$ and $d_{\mathrm{TV}}(P_i, P_j) = 0.5\epsilon\|v_i - v_j\|_1/k \geq 0.01\epsilon$.

We now calculate $\chi_P(P_u, P_v)$: For $a \neq b$, we have

$$\chi_P(P_a, P_b) = \left|\sum_{i=1}^k kP_a(i)P_b(i) - 1\right|$$

$$
\begin{aligned}
&= \left| \sum_{i=1}^{k} k \frac{1}{k} \left(1 + \epsilon v_{a,i}\right) \frac{1}{k} \left(1 + \epsilon v_{b,i}\right) - 1 \right| \\
&= \left| \sum_{i=1}^{k} \frac{1}{k} \left(1 + \epsilon v_{a,i} + \epsilon v_{b,i} + \epsilon^2 v_{a,i} v_{b,i}\right) - 1 \right| \\
&= 0,
\end{aligned}
$$

where use the facts that $\sum_i v_{m,i} = 0$ for all m, and $\sum_i v_{a,i} v_{b,i} = 0$ for all $a \neq b$. We now consider the setting where $P_a = P_b$:

$$
\begin{aligned}
\chi_P(P_a, P_a) &= \sum_{i=1}^{k} k P_a^2(i) - 1 = \sum_{i=1}^{k} k \frac{1}{k^2} \left(1 + \epsilon v_{a,i}\right)^2 - 1 \\
&= \sum_{i=1}^{k} \frac{1}{k} \left(1 + 2\epsilon v_{a,i} + \epsilon^2 v_{a,i}^2\right) - 1 = \epsilon^2,
\end{aligned}
$$

where use the facts that $\sum_i v_{a,i} = 0$ and $|v_{a,j}| = 1$ for all a and j. Overall, we obtain the following bound on the average correlation for any subset $\mathcal{P}' \subseteq \mathcal{P}_M$:

$$
\rho(\mathcal{P}', P) = \frac{1}{|\mathcal{P}'|^2} \sum_{P_1, P_2 \in \mathcal{P}'} \chi_P(P_1, P_2) = \frac{\epsilon^2}{|\mathcal{P}'|}.
$$

Thus, we have $\mathrm{SDA}(\mathcal{P}_M, P, \gamma) \geq M\bar{\gamma}/\epsilon^2$ for any $\bar{\gamma} \leq \epsilon^2$. □

Corollary G.5.9. *Consider any $\epsilon \in (0, 1)$, $D \in \mathbb{N}$, and $M \in \mathbb{N}$ such that $M \gtrsim \frac{\log D}{\epsilon D}$. Let $\mathcal{P}_M$ and P be as defined in Lemma G.5.8. Then the following hold:*

1. $n^*(\mathcal{P}_M \cup \{P\}) \lesssim \frac{\log M}{\epsilon^2}$.
2. *(Blackboard communication model.) Consider the blackboard communication model with m machines, each with an i.i.d. sample from Q (belonging to $\mathcal{P}_M$ or P) and $\log D$ bits. Any (randomized) algorithm that solves $\mathcal{B}(\mathcal{P}_M, P)$ with success probability $9/10$ requires*

$$
m \gtrsim \frac{M^{1/3}}{\epsilon^{2/3} D^{2/3} (\log D)^{1/3}}.
$$

Proof. The bound on $n^*(\mathcal{P}_M \cup P)$ follows from Fact 10.2.4 and the fact that the distributions are separated in total variation distance.

We now turn our attention to the lower bound. Fix any $\bar{\gamma}$ such that $\bar{\gamma} \leq \epsilon^2$. Lemma G.5.8 implies that the SDA of this decision problem, denoted by d, is at least $\frac{M\bar{\gamma}}{\epsilon^2}$. Thus, Theorem G.5.7 states that $m \gtrsim \min\left\{\frac{1}{D\sqrt{\bar{\gamma}}}, \frac{d}{\log D}\right\} \gtrsim \min\left\{\frac{1}{D\sqrt{\bar{\gamma}}}, \frac{M\bar{\gamma}}{\epsilon^2 \log D}\right\}$. Taking

$\bar{\gamma} = \left(\frac{\epsilon^2 \log D}{DM}\right)^{2/3}$, which satisfies $\bar{\gamma} \le \epsilon^2$, we have $m \gtrsim \frac{M^{1/3}}{\epsilon^{2/3} D^{2/3} (\log D)^{1/3}}$. This completes the proof of Corollary G.5.9 and the second claim in Theorem 10.5.8. □

Remark G.5.10. *Note that Lemma G.5.8 also implies a lower bound of $\Omega\left(\frac{\sqrt{M}}{\epsilon D}\right)$ for the special case of sequentially-adaptive algorithms by using the lower bound for the 1-MSTAT(D) oracle in Theorem G.5.6.*

G.5.3 Lower Bounds from Impossibility of ℓ_1-embedding

The main result of this section is Theorem G.5.12. Before that, we first state the following technical lemma, adapted from Lee, Mendel, and Naor [LMN05, Lemma 3.1] (also see Charikar and Sahai [CS02]):

Lemma G.5.11. *Let $r \in \mathbb{N}$. For any $M = 2^r$ and $\epsilon \in (0, 1)$, there exists a set of distributions $\mathcal{P} = \{P, P_1, \ldots, P_M\} \subseteq \Delta_M$ such that for any $D \in \mathbb{N}$ and $\mathbf{T} \in \mathcal{T}_D$, we have*

$$\frac{1}{M} \sum_{i=1}^{M} d_{\mathrm{TV}}(\mathbf{T}P_i, \mathbf{T}P) \le \frac{\epsilon\sqrt{D}}{\sqrt{M}},$$

and for any distinct $p, q \in \mathcal{P}$, we have $d_{\mathrm{TV}}(p, q) \gtrsim \epsilon$.

Proof. Let $P, P_1, \ldots, P_M$ be the distributions from Lemma G.5.8. We follow the proof strategy in Lee, Mendel, and Naor [LMN05, Lemma 3.1]. We begin by writing

$$\begin{aligned}
\sum_{i=1}^{M} \|\mathbf{T}(P_i - P)\|_2^2 &= \epsilon^2 \sum_{i=1}^{M} \left\|\mathbf{T}\frac{v_i}{M}\right\|_2^2 && \text{(by definition of the } P_i\text{'s)} \\
&= \frac{\epsilon^2}{M^2} \sum_{i=1}^{M} \left\|\sum_{j=1}^{M} v_{i,j} \mathbf{T} e_j\right\|_2^2 \\
&= \frac{\epsilon^2}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{l=1}^{M} \langle v_{i,j} \mathbf{T} e_j, v_{i,l} \mathbf{T} e_l \rangle \\
&= \frac{\epsilon^2}{M^2} \sum_{j=1}^{M} \sum_{l=1}^{M} \langle \mathbf{T} e_j, \mathbf{T} e_l \rangle \left(\sum_{i=1}^{M} v_{i,j} v_{i,l}\right) \\
&= \frac{\epsilon^2}{M^2} \sum_{j=1}^{M} \sum_{l=1}^{M} \langle \mathbf{T} e_j, \mathbf{T} e_l \rangle \langle v_j, v_l \rangle && \text{(using symmetry of } V\text{)} \\
&= \frac{\epsilon^2}{M^2} \sum_{j=1}^{M} \|\mathbf{T} e_j\|_2^2 \|v_j\|_2^2 \\
&= \frac{\epsilon^2}{M} \sum_{j=1}^{M} \|\mathbf{T} e_j\|_2^2 && \text{(using } V^\top V = kI \text{ and } k = M\text{)}
\end{aligned}$$

$$\leq \frac{\epsilon^2}{M} \sum_{j=1}^{M} \|\mathbf{T}e_j\|_1^2 \qquad \text{(using } \|x\|_2 \leq \|x\|_1\text{)}$$
$$= \epsilon^2,$$

where the last equality uses the fact that $\mathbf{T}e_j \in \Delta_D$. Applying Cauchy-Schwarz and using the fact that $\|x\|_1 \leq \sqrt{D}\|x\|_2$ for $x \in \Delta_D$, we obtain

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M} d_{\mathrm{TV}}(\mathbf{T}P_i, \mathbf{T}P) &\leq \sqrt{\frac{1}{M}\sum_{i=1}^{M}(d_{\mathrm{TV}}(\mathbf{T}P_i, \mathbf{T}P))^2} \\
&\leq \sqrt{\frac{1}{M}\sum_{i=1}^{M}\|\mathbf{T}(P_i - P)\|_1^2} \\
&\leq \sqrt{\frac{D}{M}\sum_{i=1}^{M}\|\mathbf{T}(P_i - P)\|_2^2} \\
&\leq \frac{\epsilon\sqrt{D}}{\sqrt{M}}.
\end{aligned}$$

□

We are now ready to prove the $\Omega(\sqrt{M})$ lower bound for non-adaptive, non-identical channels:

Theorem G.5.12. *There exists a set $\mathcal{P} = \{P, P_1, \ldots, P_M\} \subseteq \Delta_M$ such that the following hold:*

1. $n^*(\mathcal{P}) \lesssim \frac{\log M}{\epsilon^2}$, *and*
2. $n^*_{\texttt{non-identical}}(\mathcal{P}, \mathcal{T}_D) \gtrsim \frac{\sqrt{M/D}}{\epsilon}$.

Proof. We will assume that $M = 2^r$ for some $r \in \mathbb{N}$. Let $\mathcal{P} = \{P, P_1, \ldots, P_M\}$ be the set of distributions from Lemma G.5.11. The upper bound on $n^*(\mathcal{P})$ follows by the lower bound on the pairwise total variation distance and Fact 10.2.4. We now turn our attention to the lower bound.

Fix any arbitrary choice of different $\{\mathbf{T}_1, \ldots, \mathbf{T}_N\}$, where $\mathbf{T}_i$ is the channel used by the i^{th} user. We use the following series of inequalities to upper-bound the minimum separation in total variation distance, for any choice of $\mathbf{T}_1, \ldots, \mathbf{T}_N$:

$$\begin{aligned}
&\max_{\mathbf{T}_l : l \in [N]} \min_{p \neq q \in \mathcal{P}} d_{\mathrm{TV}}\left(\prod_{l=1}^{N}\mathbf{T}_l p, \prod_{l=1}^{N}\mathbf{T}_l q\right) \\
&\leq \max_{\mathbf{T}_l : l \in [N]} \min_{i \in [M]} d_{\mathrm{TV}}\left(\prod_{l=1}^{N}\mathbf{T}_l P_i, \prod_{l=1}^{N}\mathbf{T}_l P\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \max_{\mathbf{T}_l : l \in [N]} \sum_{i=1}^{M} \frac{1}{M} d_{\mathrm{TV}} \left(\prod_{l=1}^{N} \mathbf{T}_l P_i, \prod_{l=1}^{N} \mathbf{T}_l P \right) && \text{(minimum is less than the average)} \\
&\leq \max_{\mathbf{T}_l : l \in [N]} \sum_{i=1}^{M} \frac{1}{M} \sum_{r=1}^{N} d_{\mathrm{TV}} \left(\mathbf{T}_l P_i, \mathbf{T}_l P \right) && \text{(subadditivity of } d_{\mathrm{TV}} \text{ (Fact 10.2.2))} \\
&= \max_{\mathbf{T}_l : l \in [N]} \sum_{r=1}^{N} \sum_{i=1}^{M} \frac{1}{M} d_{\mathrm{TV}} \left(\mathbf{T}_l P_i, \mathbf{T}_l P \right) && \text{(exchanging the sum)} \\
&= N \max_{\mathbf{T} \in \mathcal{T}_D} \sum_{i=2}^{M} \frac{1}{M} d_{\mathrm{TV}} \left(\mathbf{T} P_i, \mathbf{T} P \right) && (N \text{ independent optimization problems}) \\
&\leq \frac{N \epsilon \sqrt{D}}{\sqrt{M}} && \text{(using Lemma G.5.11)}.
\end{aligned}
$$

Thus, if $N \lesssim \frac{1}{\epsilon}\sqrt{\frac{M}{D}}$, for any choice of N channels, there exist $P', P \in \mathcal{P}$ such that the total variation distance between the resulting product distributions is at most 0.001. Consequently, Fact 10.2.4 implies that there exists no test with probability of success in distinguishing between P and P' more than 0.95 (say). Hence, one must have $N \gtrsim \frac{1}{\epsilon}\sqrt{\frac{M}{D}}$ for any successful test. Since this holds for an arbitrary choice of channels, we have the desired lower bound on $n^*_{\texttt{non-identical}}(\mathcal{P}, \mathcal{T}_D)$. □

G.6 Auxiliary Details

We first mention that the class of well-behaved f-divergences included various well-known f-divergences:

Claim G.6.1 (Examples of well-behaved f-divergences)**.** *The following are examples of well-behaved f-divergences (cf. Definition 10.3.1):*

1. *(Hellinger distance) $f(x) = (\sqrt{x} - 1)^2$ with $\kappa = 1$, $C_1 = 2^{-3.5}$, $C_2 = 1$, and $\alpha = 2$.*
2. *(Total variation distance) $f(x) = 0.5|x - 1|$ with $\kappa > 0$, $C_1 = 0.5$, $C_2 = 0.5$, and $\alpha = 1$.*
3. *(Symmetrized KL-divergence) $f(x) = x \log x - \log x$ with $\kappa = 1$, $C_1 = 0.5$, $C_2 = 1$, and $\alpha = 2$.*
4. *(Triangular discrimination) $f(x) = \frac{(x-1)^2}{1+x}$ with $\kappa = 1$, $C_1 = 1/3$, $C_2 = 1/2$, and $\alpha = 2$.*
5. *(Symmetrized χ^s-divergence) For $s \geq 1$, $f(x) = |x - 1|^s + x^{1-s}|x - 1|^s$ with $\kappa = 1$, $C_1 = 1$, $C_2 = 3$, and $\alpha = s$.*[47]

[47] The usual χ^s-divergence corresponds to $f(x) = |x - 1|^s$, for $s \geq 1$ [Sas18]. We consider the symmetrized version with $\tilde{f}(x) = f(x) + xf(1/x)$.

Proof. It is easy to see that these functions are nonnegative, convex, and satisfy the symmetry property of Definition 10.3.1. In the remainder of the proof, we outline how they satisfy the property I.3.

1. We will show that we can take $\kappa = 1$, $C_1 = 2^{-3.5}$, $C_2 = 1$, and $\alpha = 2$. The upper bound $f(1+x) = (\sqrt{1+x} - 1)^2 \le x^2$ follows by noting that $\sqrt{1+x} \le 1 + x$ for any $x \ge 0$. For the lower bound, we define $g(x) := f(1+x) - C_1 x^2$. Note that $g(0) = 0$, $g'(x) = 1 - (1+x)^{-0.5} - 2C_1 x$, and $g''(x) = 0.5(1+x)^{-1.5} - 2C_1$. We note that $g'(0) = 0$ and $g''(x) \ge g''(1) = 2^{-2.5} - 2C_1$ for all $x \in [0,1]$. Thus, $g''(x) \ge 0$ for $x \in [0,1]$, so $g(x)$ is also nonnegative on $x \in [0,1]$.

2. The result follows by noting that for $x \ge 0$, we have $f(1+x) = x$.

3. We have $f(1+x) = x\log(1+x)$. We use the fact that $\frac{x}{1+x} \le \log(1+x) \le x$ for $x \ge 0$. This directly gives us $f(1+x) = x\log(1+x) \le x^2$. The lower bound follows by noting that $\log(1+x) \ge \frac{x}{2}$ for $x \in [0,1]$, so $f(1+x) \ge x^2/2$.

4. We have $f(1+x) = x^2/(2+x)$, which lies between $x^2/3$ and $x^2/2$ for $x \in [0,1]$.

5. We have $f(1+x) = |x|^s(1 + (1+x)^{1-s})$, which is larger than x^s and less than $3x^s$ for $x \in [0,1]$.

□

Finally, we mention the following approximation for the Hellinger distance between two Bernoulli distributions that was used earlier:

Claim G.6.2 (Approximation for Hellinger distance)**.** *For $0 \le p \le p' \le 1/2$, we have the following:*

$$\sqrt{p'} - \sqrt{p} \le \sqrt{\left(\sqrt{p} - \sqrt{p'}\right)^2 + \left(\sqrt{1-p} - \sqrt{1-p'}\right)^2} \le \sqrt{2}\left(\sqrt{p'} - \sqrt{p}\right).$$

Proof. The first inequality follows by the nonnegativity of the term $\left(\sqrt{1-p} - \sqrt{1-p'}\right)^2$. To prove the second inequality, it suffices to show that $\left(\sqrt{1-p} - \sqrt{1-p'}\right)^2 \le \left(\sqrt{p'} - \sqrt{p}\right)^2$, which is equivalent to showing that $\sqrt{1-p} - \sqrt{1-p'} \le \sqrt{p'} - \sqrt{p}$. For $z \in [0, 0.5]$, define $f(z) := \sqrt{z} + \sqrt{1-z}$. The desired inequality is then equivalent to showing that $f(p) \le f(p')$, which follows if $f'(z) \ge 0$ for $z \in [0, 0.5]$. Calculating the

derivative, we obtain

$$f'(z) = \frac{1}{2\sqrt{z}} - \frac{1}{2\sqrt{1-z}} = \frac{1}{2}\frac{\sqrt{1-z}-\sqrt{z}}{\sqrt{z(1-z)}} \geq 0,$$

since $z \in [0, 0.5]$. □

H APPENDIX TO CHAPTER 11

H.1 Randomized Response in Low-Privacy Regime

In this section, we prove Lemmata 11.3.5 and 11.3.6, which were used to prove Theorem 11.1.13 in Section 11.3. Lemma 11.3.5 is proved in Appendix H.1.1 and Lemma 11.3.6 is proved in Appendix H.1.2.

H.1.1 Proof of Lemma 11.3.5

Recall the definitions of A and A' from equation Equation (11.17).

Lemma 11.3.5 (Randomized response preserves contribution of comparable elements)**.** *Let p and q be two distributions on $[\ell]$. Suppose $\sum_{i\in A\bigcup A'}(\sqrt{q_i}-\sqrt{p_i})^2 \geq \tau$. Then $\mathbf{T}^{\epsilon,\ell}_{\mathrm{RR}}$, for $\ell \leq e^{\epsilon}$, satisfies*

$$d^2_{\mathrm{h}}(\mathbf{T}^{\epsilon,\ell}_{\mathrm{RR}}p, \mathbf{T}^{\epsilon,\ell}_{\mathrm{RR}}q) \gtrsim \min\left(1, e^{\epsilon}\frac{\tau}{\ell}\right)\cdot\tau\,.$$

Thus, when $e^{\epsilon} \gtrsim \frac{\ell}{\tau}$, the randomized response preserves the original contribution of comparable elements.

Proof. Without loss of generality, we will assume that $\sum_{i\in A}(\sqrt{q_i}-\sqrt{p_i})^2 \geq \frac{\tau}{2}$. Let $p' = \mathbf{T}^{\epsilon,\ell}_{\mathrm{RR}}p$ and $q' = \mathbf{T}^{\epsilon,\ell}_{\mathrm{RR}}q$. By the definition of the randomized response, each probability x is mapped to $(1+x(e^{\epsilon}-1))/(k-1+e^{\epsilon})$. Thus, p' and q' are given by

$$p'_i = \frac{1+p_i(e^{\epsilon}-1)}{(\ell-1)+e^{\epsilon}}, \quad \text{and} \quad q'_i = \frac{1+q_i(e^{\epsilon}-1)}{(\ell-1)+e^{\epsilon}}, \quad \forall i\in\ell. \tag{H.1}$$

Recall that $\delta_i = (p_i-q_i)/q_i \in [0,1]$. For each $i\in\ell$, we now define $\delta'_i := (p'_i-q'_i)/q'_i$, which has the following expression in terms of δ_i and q_i:

$$\delta'_i = \frac{p'_i-q'_i}{q'_i} = \frac{(e^{\epsilon}-1)(p_i-q_i)}{1+q_i(e^{\epsilon}-1)} = \frac{(e^{\epsilon}-1)q_i}{1+q_i(e^{\epsilon}-1)}\cdot\delta_i. \tag{H.2}$$

Let $r = 0.01\min\left(e^{-\epsilon}, \frac{\tau}{\ell}\right)$. We define the following subsets of the domain:

$$\mathcal{E} = \{i : \delta_i \in (0,1] \text{ and } q_i \geq e^{-\epsilon}\}\,, \tag{H.3}$$

$$\mathcal{E}' = \{i : \delta_i \in (0,1] \text{ and } q_i \in (r, e^{-\epsilon})\}\,. \tag{H.4}$$

Observe that $\mathcal{E} \cup \mathcal{E}' \subseteq A$.

Since $e^\epsilon \geq \ell$, equation Equation (H.1) implies that $q'_i \geq \frac{1}{4}(e^{-\epsilon} + q_i)$. In particular, on $i \in \mathcal{E}'$, we have $q'_i \geq 0.25e^{-\epsilon}$, and on $i \in \mathcal{E}$, we have $q'_i \geq 0.25q_i$.

We now apply these approximations to equation Equation (H.2): we lower-bound the numerator by $0.5e^\epsilon q_i \delta_i$ and upper-bound the denominator based on whether $i \in \mathcal{E}$ or $i \in \mathcal{E}'$. On $\mathcal{E}'$, the denominator in equation Equation (H.2) is upper-bounded by 2, and on $\mathcal{E}$, the denominator is upper-bounded by $2q_i e^\epsilon$. This is summarized as follows: for $i \in \mathcal{E} \cup \mathcal{E}'$, we have

$$\delta'_i \geq \begin{cases} 0.1\delta_i q_i e^\epsilon, & i \in \mathcal{E}' \\ 0.1\delta_i, & i \in \mathcal{E}, \end{cases} \qquad q'_i \geq \begin{cases} 0.25e^{-\epsilon}, & i \in \mathcal{E}' \\ 0.25q_i, & i \in \mathcal{E} \end{cases}.$$

By definition of δ', it follows that $\delta'_i \in (0, 1]$ on $i \in \mathcal{E} \cup \mathcal{E}'$. Thus, the contribution from the i^{th} element to $d^2_{\text{h}}(p', q')$ is at least a constant times $q'_i(\delta'_i)^2$; see Claim 11.3.3. Applying this element-wise, we obtain the following:

$$\begin{aligned} d^2_{\text{h}}(p', q') &\gtrsim \sum_{i \in \mathcal{E}'} q'_i(\delta'_i)^2 + \sum_{i \in \mathcal{E}} q'_i(\delta'_i)^2 \\ &\gtrsim \sum_{i \in \mathcal{E}'} e^{-\epsilon} \left(0.1\delta_i q_i e^\epsilon\right)^2 + \sum_{i \in \mathcal{E}} q_i \left(0.1\delta_i\right)^2 \\ &\gtrsim e^\epsilon r \sum_{i \in \mathcal{E}'} q_i \delta_i^2 + \sum_{i \in \mathcal{E}} q_i \delta_i^2. \end{aligned} \tag{H.5}$$

Now consider the set $\mathcal{A} = \{i : i \in A \text{ and } q_i \geq r\}$, which is equal to $\mathcal{E} \cup \mathcal{E}'$. The set $\mathcal{A}$ preserves the contribution to Hellinger divergence from comparable elements, as shown below:

$$\sum_{i \in \mathcal{A}} (\sqrt{q_i} - \sqrt{p_i})^2 = \sum_{i \in A} (\sqrt{q_i} - \sqrt{p_i})^2 - \sum_{i : i \in A, q_i \leq r} (\sqrt{q_i} - \sqrt{p_i})^2 \geq \frac{\tau}{2} - 2\ell r \geq \frac{\tau}{4},$$

since $r \leq \frac{\tau}{10\ell}$.

Since $\mathcal{A} = \mathcal{E}_1 \cup \mathcal{E}_2$, one of the two terms $\sum_{i \in \mathcal{E}'} (\sqrt{q_i} - \sqrt{p_i})^2$ or $\sum_{i \in \mathcal{E}} (\sqrt{q_i} - \sqrt{p_i})^2$ must be at least $\frac{\tau}{8}$.

Now consider the following two cases:

Case 1: $\sum_{i \in \mathcal{E}} (\sqrt{q_i} - \sqrt{p_i})^2 \gtrsim \tau$**.** In this case, we are done by inequality Equation (H.5). That is,

$$d^2_{\text{h}}(p', q') \gtrsim \sum_{i \in \mathcal{E}} (\sqrt{q'_i} - \sqrt{p'_i})^2 \gtrsim \sum_{i \in \mathcal{E}} q_i \delta_i^2 \gtrsim \sum_{i \in \mathcal{E}} (\sqrt{q_i} - \sqrt{p_i})^2 \gtrsim \tau,$$

where we use Claim 11.3.3 element-wise.

Case 2: $\sum_{i\in\mathcal{E}'}(\sqrt{q_i}-\sqrt{p_i})^2 \gtrsim \tau$**.** By inequality Equation (H.5), we have

$$d_{\mathrm{h}}^2(p',q') \gtrsim e^{\epsilon}\cdot r\sum_{i\in\mathcal{E}'} q_i\delta_i^2 \gtrsim e^{\epsilon}\cdot r\tau \gtrsim \min\left(1, e^{\epsilon}\frac{\tau}{\ell}\right)\tau,$$

where we use the definition of r.

Thus, we obtain the desired lower bound in both of the cases. □

H.1.2 Proof of Lemma 11.3.6

Lemma 11.3.6 (Reduction to base case)**.** *Let p and q be two distributions on $[k]$. Then there is a channel $\mathbf{T}$, which can be computed in time polynomial in k, that maps $[k]$ to $[\ell]$ (for ℓ to be decided below) such that for $p'=\mathbf{T}p$ and $q'=\mathbf{T}q$, at least one of the following holds:*

1. *For any $\ell>2$ and $\ell\le\min\left(k, 1+\log\left(1/d_{\mathrm{h}}^2(p,q)\right)\right)$, we have*

$$\sum_{i\in B\bigcup B'}\left(\sqrt{q_i'}-\sqrt{p_i'}\right)^2 \gtrsim d_{\mathrm{h}}^2(p,q)\cdot\frac{\ell}{\min\left(k, 1+\log\left(1/d_{\mathrm{h}}^2(p,q)\right)\right)},$$

   where B and B' are defined analogously to A and A' in equation Equation (11.17), but with respect to distributions p' and q'.

2. *$\ell=2$ and $d_{\mathrm{h}}^2(p',q')\gtrsim d_{\mathrm{h}}^2(p,q)$.*

Proof. Let us begin by considering the case when $\sum_{i\in A\bigcup A'}\left(\sqrt{q_i}-\sqrt{p_i}\right)^2 \le \frac{d_{\mathrm{h}}^2(p,q)}{2}$. Following Pensia, Jog, and Loh [PJL22, Theorem 2 (Case 1 in the proof)], there exists a binary channel that preserves the Hellinger divergence up to constants. This completes the case for $\ell=2$ above.

Suppose for now that $\sum_{i\in A\bigcup A'}\left(\sqrt{q_i}-\sqrt{p_i}\right)^2 \ge \frac{d_{\mathrm{h}}^2(p,q)}{2}$, i.e., the comparable elements constitute at least half the Hellinger divergence. Consider the channel $\mathbf{T}'$ that maps the comparable elements of p and q to distinct elements, and maps the remaining elements to a single super-element. Let α be the contribution to the Hellinger divergence from the comparable elements in $\mathbf{T}'p$ and $\mathbf{T}'q$ (defined analogously to equation Equation (11.17)). It can be seen that $\alpha\ge\frac{d_{\mathrm{h}}^2(p,q)}{2}$. Let $\ell\ge 3$ be as in the statement. Now consider the channel $\mathbf{T}''$ that compresses $\mathbf{T}'p$ and $\mathbf{T}'q$ into ℓ-ary distributions that preserve the Hellinger divergence, from Pensia, Jog, and Loh [PJL22, Theorem 3.2 (Case 2 in the proof)]. Let β_ℓ be the contribution to the Hellinger divergence from the comparable elements in $\mathbf{T}''\mathbf{T}'p$

and $\mathbf{T}''\mathbf{T}'q$. Then the result in Pensia, Jog, and Loh [PJL22, Theorem 3.2] implies that $\beta_l \gtrsim \alpha\left(\ell/\min(k, 1+\log(1/d_{\mathrm{h}}^2(p,q)))\right)$. This completes the proof in this setting. □

H.2 Properties of Private Channels

Recall the definition of the set of channels $\mathcal{J}_{\ell,k}^{\gamma,\nu}$ from Definition 11.5.1 below:

Definition 11.5.1 (LP family of channels)**.** *For any $\ell \in \mathbb{N}$, let $\nu = (\nu_1, \nu_2, \ldots, \nu_\ell)$ and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_\ell)$ be two nonnegative vectors in $\mathbb{R}_+^\ell$. For $k \in \mathbb{N}$, define the set of linear programming (LP) channels $\mathcal{J}_{\ell,k}^{\gamma,\nu}$, a subset of $\mathcal{T}_{\ell,k}$, to be the (convex) set of all channels from $[k]$ to $[\ell]$ that satisfy the following constraints:*

$$\text{For each row } j \in [\ell], \text{ and for each } i, i' \in [k], \text{ we have } \mathbf{T}(j,i) \le \gamma_j \mathbf{T}(j,i') + \nu_j. \tag{11.21}$$

We begin by an equivalent characterization of the constraints above. For a channel $\mathbf{T}$ from $[k]$ to $[\ell]$, let $\{m_1, \ldots, m_\ell\}$ and $\{M_1, \ldots, M_\ell\}$ be the minimum and maximum entries of each row, respectively. Then the channel $\mathbf{T}$ satisfies the conditions Equation (11.21) if and only if for each $j \in [\ell]$, we have

$$M_j \le \gamma_j m_j + \nu_j. \tag{H.6}$$

We first show that $\mathcal{J}_{\ell,k}^{\gamma,\nu}$ satisfies Condition 11.5.3. For the special case of LDP channels, the following claim was also proved in Holohan, Leith, and Mason [HLM17]:

Claim H.2.1. *$\mathcal{J}_{\ell,k}^{\gamma,\nu}$ satisfies Condition 11.5.3.*

Proof. Let $\mathbf{T}$ be any extreme point of $\mathcal{J}_{\ell,k}^{\gamma,\nu}$. Let $\{m_1, \ldots, m_\ell\}$ and $\{M_1, \ldots, M_\ell\}$ be as defined above. Suppose that there exists $c \in [k]$, such that there exist distinct $r, r' \in [\ell]$ with $\mathbf{T}(r,c) \in (m_r, M_r)$ and $\mathbf{T}(r',c) \in (m_{r'}, M_{r'})$. In particular, both $\mathbf{T}(r,c)$ and $\mathbf{T}(r',c)$ are strictly positive and less than 1.

We will now show that $\mathbf{T}$ is not an extreme point of $\mathcal{J}_{\ell,k}^{\gamma,\nu}$. For an $\epsilon > 0$ to be decided later, consider the channel $\mathbf{T}'$ that is equal to $\mathbf{T}$ on all but two entries:

- On (r,c), $\mathbf{T}'$ assigns probability $\mathbf{T}(r,c) + \epsilon$.
- On (r',c), $\mathbf{T}'$ assigns probability $\mathbf{T}(r',c) - \epsilon$.

Now define $\mathbf{T}''$ similarly, with the difference being that on (r,c), $\mathbf{T}''$ assigns probability $\mathbf{T}(r,c) - \epsilon$, and on (r',c), $\mathbf{T}''$ assigns probability $\mathbf{T}(r',c) + \epsilon$. Both $\mathbf{T}'$ and $\mathbf{T}''$ are thus

valid channels for ϵ small enough. Let us show that $\mathbf{T}'$ and $\mathbf{T}''$ belong to $\mathcal{C}$. If we choose $\epsilon > 0$ small enough, the row-wise maximum and minimum entries of $\mathbf{T}'$ and $\mathbf{T}''$ are equal to those of $\mathbf{T}$. Here, we critically use the fact that the entries that were modified were "free." By inequality Equation (H.6), both $\mathbf{T}'$ and $\mathbf{T}''$ belong to $\mathcal{J}_{\ell,k}^{\gamma,\nu}$. Since $\mathbf{T}$ is the average of $\mathbf{T}'$ and $\mathbf{T}''$, it is not an extreme point of $\mathcal{J}_{\ell,k}^{\gamma,\nu}$. □

We now show that $\mathcal{J}_{\ell,k}^{\gamma,\nu}$ satisfies Condition 11.5.7.

Claim H.2.2. *$\mathcal{J}_{\ell,k}^{\gamma,\nu}$ satisfies Condition 11.5.7.*

Proof. We follow the notation from Condition 11.5.7. Let $\mathbf{T}$ be an extreme point of $\mathcal{J}_{\ell,k}^{\gamma,\nu}$, and let r and r' be the corresponding rows. We show that $\mathbf{T}'$ (defined in the condition) belongs to $\mathcal{J}_{\ell,k}^{\gamma,\nu}$ by showing that entries of $\mathbf{T}'$ satisfy the constraints of the r^{th} row and the r'^{th} row (since the other rows are unchanged). In fact, we establish these arguments only for the r^{th} row, and the analogous arguments hold for the r'^{th} row.

Let m_r and M_r be the row-wise minimum and maximum entry of this row in $\mathbf{T}$. Let us first consider the case when $M_r < \gamma_r m_r + \nu_r$. Then there exist positive ϵ' and δ' such that $M_r + \delta < \gamma_r(m_r - \epsilon) + \nu_r$. By inequality Equation (H.6), as long as the r^{th} row of a channel contains entries in $[m_r - \epsilon, M_r + \delta]$, the constraints of this particular row will be satisfied. Since the entries in the r^{th} row of $\mathbf{T}'$ belong to this interval, the constraints of the r^{th} row are satisfied by $\mathbf{T}'$.

Let us now consider the alternate case where $M_r = \gamma_r m_r + \nu_r$. Since m and M do not correspond to the min-tight and max-tight entries, we have $m_r < M$ and $m < M_r$. Consequently, even after perturbations by $\epsilon > 0$ and $\delta > 0$ small enough, the entries of $\mathbf{T}'$ lie in $[m_r, M_r]$. Thus, inequality Equation (H.6) implies that the constraints of the r^{th} row in $\mathbf{T}'$ are satisfied. □

Claim 11.5.9 (Closure under pre-processing)**.** *The set $\mathcal{J}_{\ell,k}^{\gamma,\nu}$ satisfies the following closure property under pre-processing:*

$$\mathcal{J}_{\ell,k}^{\gamma,\nu} = \bigcup_{\ell'=1}^{k} \left\{ \mathbf{T}_2 \times \mathbf{T}_1 : \mathbf{T}_2 \in \mathcal{J}_{\ell,\ell'}^{\gamma,\nu} \text{ and } \mathbf{T}_1 \in \mathcal{T}_{\ell',k} \right\}. \tag{11.23}$$

Proof. We first show the simple direction that

$$\mathcal{J}_{\ell,k}^{\gamma,\nu} \subseteq \bigcup_{\ell'=1}^{k} \left\{ \mathbf{T}_2 \times \mathbf{T}_1 : \mathbf{T}_2 \in \mathcal{J}_{\ell,\ell'}^{\gamma,\nu} \text{and } \mathbf{T}_1 \in \mathcal{T}_{\ell',k} \right\}.$$

Let $\mathbf{I}_k$ correspond to the identity channel on $[k]$. Then every channel $\mathbf{T} \in \mathcal{J}_{\ell,k}^{\gamma,\nu}$, can be written as $\mathbf{T} \times \mathbf{I}$. Thus, $\mathcal{J}_{\ell,k}^{\gamma,\nu} \subseteq \left\{\mathbf{T}_2 \times \mathbf{I}_k : \mathbf{T}_2 \in \mathcal{J}_{\ell,\ell'}^{\gamma,\nu}\right\}$, and the desired conclusion follows.

We now show that every channel in the right-hand side belongs to $\mathcal{J}_{\ell,k}^{\gamma,\nu}$. For an arbitrary $\ell' \in [k]$, let $\mathbf{T}_2 \in \mathcal{J}_{\ell,\ell'}^{\gamma,\nu}$. Define $\{m_1, \ldots, m_\ell\}$ and $\{M_1, \ldots, M_\ell\}$ to be the minimum and maximum entries of each row in $\mathbf{T}_2$, respectively. By inequality Equation (H.6), for each $j \in [\ell]$, we have $M_j \leq \gamma_j m_j + \nu_j$. Let $\mathbf{T}_1 \in \mathcal{T}_{\ell',k}$ be an arbitrary channel.

Let $\mathbf{T} = \mathbf{T}_2 \times \mathbf{T}_1$ be in $\mathcal{T}_{\ell,k}$, and let $\{m_1, \ldots, m_\ell\}$ and $\{M'_1, \ldots, M'_\ell\}$ be the minimum and maximum entries of each row in $\mathbf{T}$, respectively. In order to show that $\mathbf{T} \in \mathcal{J}_{\ell,k}^{\gamma,\nu}$, we need to show that for each $j \in [\ell]$, we have $M'_j \leq \gamma_j m'_j + \nu_j$. Since it already holds that $M_j \leq \gamma_j m_j + \nu_j$ for all j, it suffices to show that $M'_j \leq M_j$ and $m'_j \geq m_j$ for all j. Observe that for any $c \in [k]$ and $r \in [\ell]$, the (r, c)-entry of $\mathbf{T}$ is a convex combination of the r^{th} row in $\mathbf{T}_2$, where the weights in the convex combination correspond to the c^{th} column in $\mathbf{T}_1$. Since the maximum of a collection of items is always as large as any convex combination of these items, we have $M'_j \leq M_j$ for all j. Similarly, we have $m'_j \geq m_j$. This completes the proof. □

H.3 Other Notions of Privacy

We provide the proof of the following result, omitted from Section 11.6:

Claim 11.6.4 (Sample complexity of approximate LDP)**.** *For all $\delta \in (0, 1)$, we have*

$$n^*(p, q, (\epsilon, \delta)) \lesssim \min\left(n^*(p, q, \epsilon) \cdot \frac{1}{1-\delta}\ , n^*(p, q) \cdot \frac{1}{\delta}\right).$$

Moreover, this is tight (up to constant factors) when both p and q are binary distributions.

Proof. Let $\mathbf{T}$ be an ϵ-LDP channel that maximizes $d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q)$ among all ϵ-LDP channels. Let $\mathbf{T}'$ be the following channel that maps from $[k]$ to $[2k]$: for any element $i \in [k]$, use the channel $\mathbf{T}$, and with probability δ, map i to $k+i$. It can be seen that $\mathbf{T}'$ satisfies (ϵ, δ)-LDP. Let p' and q' be the corresponding distributions after transforming p and q using $\mathbf{T}'$. It can be seen that p' is a distribution over $[2k]$ such that the first k elements are equal to $(1-\delta)\mathbf{T}p$ coordinate-wise, and the bottom k elements are equal to δp coordinate-wise. A similar conclusion holds for q', as well. Thus, we have

$$d_{\mathrm{h}}^2(\mathbf{T}'p, \mathbf{T}'q) = (1-\delta) \cdot d_{\mathrm{h}}^2(\mathbf{T}p, \mathbf{T}q) + \delta \cdot d_{\mathrm{h}}^2(p, q)$$

$$\asymp \max\left((1-\delta)\cdot d_{\mathrm{h}}^2(\mathbf{T}p,\mathbf{T}q), \delta\cdot d_{\mathrm{h}}^2(p,q)\right)$$
$$\asymp \max\left((1-\delta)\cdot \frac{1}{n^*(p,q,\epsilon)}, \delta\cdot\frac{1}{n^*(p,q)}\right).$$

By Fact 11.2.7, the sample complexity $n^*(p,q,(\epsilon,\delta))$ is at most $1/d_{\mathrm{h}}^2(\mathbf{T}'p,\mathbf{T}'q)$, which gives the upper bound on $n^*(p,q,(\epsilon,\delta))$.

The tightness follows from the result of Kairouz, Oh, and Viswanath [KOV16, Theorem 18], which implies that $\mathbf{T}'$ defined above is an optimal channel for binary distributions. □

H.4 Auxiliary Lemmas

H.4.1 Degenerate Conditions for Joint Range

We show in this section that we can safely rule out certain degenerate conditions for p and q for our results. Let p and q be two distributions on $[k]$. In particular, we would like to assume the following:

- Consider the likelihood ratio p_i/q_i, defined to be ∞ if $q_i = 0$ and $p_i \neq 0$, and undefined if both p_i and q_i are 0. Assume that all the likelihood ratios are well-defined and unique.

If these conditions do not hold, define p' and q' to be distributions over $[k']$ for some $k' \leq k$, constructed as follows: start by removing elements that have zero probability mass under both p and q, then merge the elements with the same likelihood ratios into super-elements. Let $\mathbf{T}^* \in \mathcal{T}_{k',k}$ be the corresponding deterministic map, which satisfies $p' = \mathbf{T}^*p$ and $q' = \mathbf{T}^*q$. We make the following claim:

Claim H.4.1. *With the notation above, for any $\ell \in \mathbb{N}$ and $\mathbf{T} \in \mathcal{T}_{\ell,k}$, there exists $\mathbf{T}' \in \mathcal{T}_{\ell,k'}$ such that $(\mathbf{T}p,\mathbf{T}q) = (\mathbf{T}'p',\mathbf{T}'q')$. In particular, $\{(\mathbf{T}p,\mathbf{T}q) : \mathbf{T} \in \mathcal{C}\} = \{(\mathbf{T}p',\mathbf{T}q') : \mathbf{T} \in \mathcal{C}'\}$ for two choices of $\mathcal{C}$ and $\mathcal{C}'$: (i) $(\mathcal{C},\mathcal{C}') = (\mathcal{T}_{\ell,k},\mathcal{T}_{\ell,k'})$ and (ii) $(\mathcal{C},\mathcal{C}') = (\mathcal{P}^{\epsilon}_{\ell,k},\mathcal{P}^{\epsilon}_{\ell,k'})$.*

Claim H.4.1 ensures that the joint ranges of (p,q) and (p',q') are identical, so our structural and algorithmic results continue to hold when applied to p' and q'. We will now prove Claim H.4.1.

Proof of Claim H.4.1. Let $\{\mathcal{I}_0,\mathcal{I}_1,\ldots,\mathcal{I}_{k'}\}$ be the smallest partition of $[k]$ such that $\mathcal{I}_0$ contains elements where both p_i and q_i are zero, and for each $i \in [k']$, the likelihood ratio of

elements in $\mathcal{I}_i$ are identical. Then the channel $\mathbf{T}^*$ mentioned above has the following form: $\mathbf{T}^*(x) = i$ if $x \in \mathcal{I}_i$ and $i > 0$, and $\mathbf{T}^*(x) = 1$ if $x \in \mathcal{I}_0$. Observe that for each $i \in [k']$, we have $p'_i = \sum_{j \in \mathcal{I}_i} p_j$, $q'_i = \sum_{j \in \mathcal{I}_i} q_j$, and at most one of them is zero.

Now consider a channel $\mathbf{T} \in \mathcal{T}_{\ell,k}$, and let $\{v_1, \ldots, v_k\}$ be the columns of $\mathbf{T}$. It is easy to see that columns belonging to indices in $\mathcal{I}_0$ do not affect $(\mathbf{T}p, \mathbf{T}q)$. For $i \in [k']$, define $\theta'_i := p'_i/q'_i$ to be the likelihood ratio of the transformed distributions. Define $\mathbf{T}'$ to be the channel with columns $v'_1, \ldots, v'_k$ such that

$$v'_i = \begin{cases} \frac{\sum_{j \in \mathcal{I}_i} v_j p_j}{p'_i} & \text{if } p'_i > 0, \\ \frac{\sum_{j \in \mathcal{I}_i} v_j q_j}{q'_i} & \text{otherwise.} \end{cases}$$

First consider the case when for all $i \in [k']$, we have $0 < \theta'_i < \infty$. Then for all $i \in [k']$, we have $p'_i = \theta'_i q'_i$ and $v'_i = \frac{\sum_{j \in \mathcal{I}_i} v_j p_j}{p'_i} = \frac{\sum_{j \in \mathcal{I}_i} v_j q_j}{q'_i}$. Thus, we have

$$\begin{aligned} (\mathbf{T}p, \mathbf{T}q) &= \left(\sum_{i \in [k']} \sum_{j \in \mathcal{I}_i} v_j p_j, \sum_{i \in [k']} \sum_{j \in \mathcal{I}_i} v_j q_j \right) \\ &= \left(\sum_{i \in [k']} p'_i \cdot \left(\frac{\sum_{j \in \mathcal{I}_i} v_j p_j}{p'_i} \right), \sum_{i \in [k']} q'_i \cdot \left(\frac{\sum_{j \in \mathcal{I}_i} v_j q_j}{q'_i} \right) \right) \\ &= \left(\sum_{i \in [k']} p'_i v'_i, \sum_{i \in [k']} q'_i v'_i \right) = (\mathbf{T}'p', \mathbf{T}'q') \,. \end{aligned}$$

We now consider the case when there is an index $a \in [k']$ such that $p'_a = 0$ and an index $b \in [k']$ such that $q'_b = 0$. Then it must be that $\theta'_a = 0$ and $\theta'_b = \infty$. Then $v'_a = \frac{\sum_{j \in \mathcal{I}_i} v_j q_j}{q'_i}$ and $v'_b = \frac{\sum_{j \in \mathcal{I}_i} v_j p_j}{p'_i}$. Following the calculations above, we obtain $\sum_{j \in \mathcal{I}_i} v_j p_j = v'_i p'_i$ for each $i \in [k] \setminus \{a\}$. In fact, the same result is true for $i = a$, since both sides are 0. The same conclusion holds for q and q', as well. This completes the proof of the first claim.

We now turn to the final claim, regarding the joint range under the channel constraints of $\mathcal{C}$. The case $\mathcal{C} = \mathcal{T}_{\ell,k}$ is immediate from the preceding discussion. Let $\mathbf{T}_1 \in \mathcal{T}_{k,k'}$ be such that $(p, q) = (\mathbf{T}_1 p', \mathbf{T}_1 q')$ and $\mathbf{T}_2 \in \mathcal{T}_{k',k}$ be such that $(p', q') = (\mathbf{T}_2 p, \mathbf{T}_2 q)$. For $\mathcal{C} = \mathcal{P}^{\epsilon}_{\ell,k}$ and $\mathcal{C}' = \mathcal{P}^{\epsilon}_{\ell,k'}$, we only need to show that (i) if $\mathbf{T}' \in \mathcal{C}'$, then $\mathbf{T}'\mathbf{T}_2 \in \mathcal{C}$; and (ii) if $\mathbf{T} \in \mathcal{C}$, then $\mathbf{T}\mathbf{T}_1 \in \mathcal{C}'$. Both of these conditions hold because privacy is closed under pre-processing. $\square$

H.4.2 Valid Choice of Parameters in Theorem 11.1.7

We now give the details that were omitted in the proof of Theorem 11.1.7 in Section 11.3.2.

We first reparametrize the problem by setting $x = \gamma$ and $y = \gamma^{1+\delta}$. The constraint $\delta > 0$ is equivalent to $y < x$. Then $d_{\mathrm{TV}}(p, q) = x + y$, and

$$d_{\mathrm{h}}^2(p, q) = 2y + \left(\sqrt{1/2 + x - y} - \sqrt{1/2}\right)^2 + \left(\sqrt{1/2 - x - y} - \sqrt{1/2}\right)^2.$$

We begin by setting $\nu = x + y$, which is possible since $0 \le y < x < 0.25$ and $\nu \in (0, 0.5)$. Then $x = \nu - y$, where $y \in (0, \nu/2)$ and $\nu \in (0, 0.5)$. Our goal is now to show that there exists a valid choice of y such that $d_{\mathrm{h}}^2(p, q) = \rho$, as long as $2\nu^2 \le \rho \le \nu$.

Define $g(y)$ to be the Hellinger divergence between p and q given y, i.e.,

$$g(y) = 2y + \left(\sqrt{1/2 + \nu - 2y} - \sqrt{1/2}\right)^2 + \left(\sqrt{1/2 - \nu} - \sqrt{1/2}\right)^2.$$

Since g is a continuous function, it suffices to show that $g(0) < 2\nu^2$ and $g(\nu/2) > \nu$, which would imply that there is a choice of $y \in (0, \nu/2)$ such that $g(y) = \rho$. We have

$$g(0) = \left(\sqrt{1/2 + \nu} - \sqrt{1/2}\right)^2 + \left(\sqrt{1/2 - \nu} - \sqrt{1/2}\right)^2 \le 3\nu^2/2,$$

where we use the fact that $\left|\sqrt{1/2 + a} - \sqrt{1/2}\right| \le a$ for all $a \ge 0$, and is less than $|a|/2$ for $a \le 0$. On the other hand, $g(\nu/2) > \nu$, since $\nu < 1/2$. Thus, there is a choice of $y \in (0, \nu/2)$ such that $d_{\mathrm{h}}^2(p, q) = \rho$. Given these choices of x and y, we can infer the choice of $\gamma \in (0, 0.25)$ and $\delta > 0$.

H.4.3 Taylor Approximation to Hellinger Divergence

Claim 11.3.3 (Additive approximation for $\sqrt{\cdot}$)**.** *There exist constants $0 < c_1 \le c_2$ such that for $0 < y \le x$, we have $c_1 \cdot \frac{y^2}{x} \le (\sqrt{x} - \sqrt{x - y})^2 \le c_2 \cdot \frac{y^2}{x}$.*

Proof. It suffices to prove that for $\delta \in (0, 1]$, we have $1 - \sqrt{1 - \delta} \asymp \delta$. We first start with the upper bound: since $1 - \delta \le \sqrt{1 - \delta}$, we have $1 - \sqrt{1 - \delta} \le \delta$. We now show the lower bound and claim that $1 - \sqrt{1 - \delta} \ge 0.5\delta$ for all $\delta \in [0, 1]$. This inequality is equivalent to showing $1 - 0.5\delta \ge \sqrt{1 - \delta}$, which is equivalent to showing that $1 + 0.25\delta^2 - \delta \ge 1 - \delta$, which holds since $\delta^2 \ge 0$. □

Claim 11.3.2 (Approximation for Hellinger divergence of binary distributions)**.** *Let* $p, q \in [0,1]$*. Let* $\mathrm{Ber}(p)$ *and* $\mathrm{Ber}(q)$ *be the corresponding Bernoulli distributions with* $\min(p,q) \le 1/2$*. Then*

$$d_{\mathrm{h}}^2\left(\mathrm{Ber}(p), \mathrm{Ber}(q)\right) \asymp \frac{d_{\mathrm{TV}}^2(\mathrm{Ber}(p), \mathrm{Ber}(q))}{\max(p,q)}.$$

Proof. Let q be the larger of the two quantities, so p satisfies $p \le \frac{1}{2}$. The total variation distance is thus $q - p$. Let $\delta = (q-p)/q \in (0,1]$. Observe that $p = q - q\delta$ and the total variation distance is δq.

We begin by noting that Claim 11.3.3 implies that

$$\left(\sqrt{q} - \sqrt{p}\right)^2 = \left(\sqrt{q} - \sqrt{q - \delta q}\right)^2 \asymp \frac{\delta^2 q^2}{q} \asymp \frac{d_{\mathrm{TV}}^2\left(\mathrm{Ber}(p), \mathrm{Ber}(q)\right)}{q}. \tag{H.7}$$

We now split the analysis into two cases:

Case 1: $q \le 1/2$**.** Then Claim G.6.2 implies that $d_{\mathrm{h}}^2(\mathrm{Ber}(p), \mathrm{Ber}(q)) \asymp (\sqrt{q} - \sqrt{p})^2$. Thus, Equation (H.7) implies the result.

Case 2: $q \ge 1/2$**.** Applying Claim 11.3.3 again to the second term, we obtain

$$\begin{aligned}\left(\sqrt{1-p} - \sqrt{1-q}\right)^2 = \left(\sqrt{1-p} - \sqrt{1-p-q\delta}\right)^2 &\asymp \frac{q^2\delta^2}{1-p} \asymp \frac{q^2\delta^2}{q} \\ &\asymp \frac{d_{\mathrm{TV}}^2(\mathrm{Ber}(p), \mathrm{Ber}(q))}{q},\end{aligned} \tag{H.8}$$

where we use the fact that $1 - p \asymp q$, since $p, q \in [0.5, 1]$. The desired conclusion follows from equations Equation (H.7) and Equation (H.8). □