High-Dimensional Robust Mean Estimation in Nearly-Linear Time

Yu Cheng* Ilias Diakonikolas† Rong Ge‡

Abstract

We study the fundamental problem of high-dimensional mean estimation in a robust model where a constant fraction of the samples are adversarially corrupted. Recent work gave the first polynomial time algorithms for this problem with dimension-independent error guarantees for several families of structured distributions.

In this work, we give the first nearly-linear time algorithms for high-dimensional robust mean estimation. Specifically, we focus on distributions with (i) known covariance and sub-gaussian tails, and (ii) unknown bounded covariance. Given N samples on $\mathbb{R}^d$, an ϵ-fraction of which may be arbitrarily corrupted, our algorithms run in time $\widetilde{O}(Nd)/\mathrm{poly}(\epsilon)$ and approximate the true mean within the information-theoretically optimal error, up to constant factors. Previous robust algorithms with comparable error guarantees have running times $\widetilde{\Omega}(Nd^2)$, for $\epsilon = \Omega(1)$.

Our algorithms rely on a natural family of SDPs parameterized by our current guess ν for the unknown mean $\mu^\star$. We give a win-win analysis establishing the following: either a near-optimal solution to the primal SDP yields a good candidate for $\mu^\star$ — independent of our current guess ν — or a near-optimal solution to the dual SDP yields a new guess ν' whose distance from $\mu^\star$ is smaller by a constant factor. We exploit the special structure of the corresponding SDPs to show that they are approximately solvable in nearly-linear time. Our approach is quite general, and we believe it can also be applied to obtain nearly-linear time algorithms for other high-dimensional robust learning problems.

1 Introduction

1.1 Background Consider the following statistical task: Given N independent samples from an unknown mean and identity covariance Gaussian distribution $\mathcal{N}(\mu^\star, I)$ on $\mathbb{R}^d$, estimate its mean vector $\mu^\star$ within small ℓ_2-norm. It is straightforward to see that the empirical mean — the average of the samples — has ℓ_2-error at most $O(\sqrt{d/N})$ from $\mu^\star$ with high probability. Moreover, this error upper bound is best possible, within constant factors, among all N-sample estimators. That is, in the aforementioned basic setting, there is a sample-optimal mean estimator that runs in linear time.

In this paper, we study the robust (or agnostic) setting when a constant $\epsilon < 1/2$ fraction of our samples can be adversarially corrupted. We consider the following model of robust estimation (see, e.g., [9]) that generalizes other existing models, including Huber's contamination model [20]:

DEFINITION 1.1. *Given $0 < \epsilon < 1/2$ and a family of distributions $\mathcal{D}$ on $\mathbb{R}^d$, the* adversary *operates as follows: The algorithm specifies some number of samples N, and N samples $X_1, X_2, \ldots, X_N$ are drawn from some (unknown) $D \in \mathcal{D}$. The adversary is allowed to inspect the samples, removes ϵN of them, and replaces them with arbitrary points. This set of N points is then given to the algorithm. We say that a set of samples is* ϵ-corrupted *if it is generated by the above process.*

In the context of robust mean estimation studied in this paper, the goal is to output a hypothesis vector $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu^\star\|_2$ is as small as possible. How do we estimate $\mu^\star$ in this regime? A moment's thought reveals that the empirical mean inherently fails in the robust setting: even a single corrupted sample can arbitrarily compromise its performance. However, one can construct more sophisticated estimators that are provably robust. The information-theoretically optimal error for robustly estimating the mean of $\mathcal{N}(\mu^\star, I)$ is $\Theta(\epsilon + \sqrt{d/N})$ [33, 17, 6]. That is, when there are enough samples ($N = \Omega(d/\epsilon^2)$) one can estimate the mean to accuracy $\Theta(\epsilon)$. [1] However, the standard robust estimators (e.g., Tukey's median [33]) require exponential time in the dimension d to compute. On the other hand, a number of natural approaches (e.g., naive outlier removal, coordinate-wise median, geometric median, etc.) can only guarantee error $\Omega(\epsilon\sqrt{d})$ (see, e.g., [9, 26]), even in the infinite sample regime. That is, the performance of these estimators degrades polynomially with the dimension d, which is clearly unacceptable in high dimensions.

Recent work [9, 26] gave the first polynomial time robust estimators for a range of high-dimensional statistical tasks, including mean and covariance estimation.

*Duke University. yucheng@cs.duke.edu
†University of Southern California. diakonik@usc.edu
‡Duke University. rongge@cs.duke.edu

[1] Under different assumptions on the distribution of the good data, the optimal error guarantee may be different as well (see Section 1.2).

Specifically, [9] obtained the first robust estimators for the mean with *dimension-independent* error guarantees, i.e., whose error only depends on the fraction of corrupted samples ϵ but *not* on the dimensionality of the data. Since the dissemination of [9, 26], there has been a substantial number of subsequent works obtaining robust learning algorithms for a variety of unsupervised and supervised high-dimensional models. (See Section 1.3 for a summary of related work.)

Although the aforementioned works gave polynomial time robust learning algorithms for several fundamental learning tasks, these algorithms are at least a factor d slower than their non-robust counterparts (e.g., the sample average for the case of mean estimation), hence are significantly slower in high dimensions. It is an important goal to design robust learning algorithms with near-optimal sample complexity that are also nearly as efficient as their non-robust counterparts. In particular, we propose the following broad question:

> *Can we design (nearly-)sample optimal robust learning algorithms — with dimension independent error guarantees — that run in* **nearly-linear time***?*

Here by *nearly-linear time*, we mean that the runtime is proportional to the size of the input, within poly-logarithmic in the input size and $\mathrm{poly}(1/\epsilon)$ factors. In addition to its potential practical implications, we believe that understanding the above question is of fundamental theoretical interest as it can elucidate the effect of the robustness requirement on the computational complexity of high-dimensional statistical learning/estimation.

For example, for the prototypical problem of robustly estimating the mean of a high-dimensional distribution, previous robust algorithms [9, 26, 32] have runtime at least $\Omega(Nd^2)$ for constant ϵ. Since the input size is $\Theta(Nd)$, we would like to obtain algorithms that run in time $\widetilde{O}(Nd)/\mathrm{poly}(\epsilon)$, where the $\widetilde{O}(\cdot)$ notation hides logarithmic factors in its argument. As the main contribution of this paper, we obtain such algorithms under different assumptions about the distribution of the good data. Our algorithms have optimal sample complexity, provide the information-theoretically optimal accuracy, and — importantly — run in time $\widetilde{O}(Nd)/\mathrm{poly}(\epsilon)$.

1.2 Our Results Our first algorithmic result handles the setting where the good data distribution is sub-gaussian with known covariance. Recall that a distribution D on $\mathbb{R}^d$ with mean $\mu^\star$ is sub-gaussian if for any unit vector $v \in \mathbb{R}^d$ we have that $\Pr_{X\sim D}[|\langle v, X-\mu^\star\rangle| \geq t] \leq \exp(-t^2/2)$. For this case, we show[2]:

Theorem 1.1 (Robust Mean Estimation for Sub-Gaussian Distributions) *Let D be a sub-gaussian distribution on $\mathbb{R}^d$ with unknown mean $\mu^\star$ and identity covariance. Let $0 < \epsilon < 1/3$ and $\delta = O(\epsilon \log 1/\epsilon)$. Given an ϵ-corrupted set of $N = \Omega(d/\delta^2)$ samples drawn from D, there is an algorithm that runs in time $\widetilde{O}(Nd)/\mathrm{poly}(\epsilon)$ and outputs a hypothesis vector $\widehat{\mu}$ such that with probability at least $9/10$ it holds $\|\widehat{\mu} - \mu^\star\|_2 \leq O(\delta) = O(\epsilon \log 1/\epsilon)$.*

It is well-known (see, e.g., [10]) that the optimal error guarantee under the assumptions of Theorem 1.1 is $\Omega(\epsilon\sqrt{\log 1/\epsilon})$, even in the infinite sample regime. Moreover, the sample complexity of the learning problem is known to be $\Omega(d/\delta^2)$ even without corruptions. Thus, our algorithm has best possible error guarantee and sample complexity, up to constant factors. Prior work [9, 10] gave algorithms with the same error and sample complexity guarantees, but with runtime $\Omega(Nd^2)$, even for constant ϵ. We note that for the very special case that $D = \mathcal{N}(\mu^\star, I)$, an error of $O(\epsilon)$ is information-theoretically possible. However, as shown in [13], any Statistical Query algorithm that runs in time $\mathrm{poly}(N)$ needs to have error $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$. Our algorithm achieves this accuracy guarantee in nearly-linear time. See Section 1.3 for a detailed summary of previous work.

Theorem 1.1 handles the case that the covariance matrix of the good data distribution is known a priori. This is a somewhat limiting assumption. In our second main algorithmic result, we obtain a similarly robust algorithm under the much weaker assumption that the covariance matrix is unknown and bounded from above. Specifically, we show:

Theorem 1.2 (Robust Mean Estimation for Bounded Covariance Distributions) *Let D be a distribution on $\mathbb{R}^d$ with unknown mean $\mu^\star$ and unknown covariance matrix Σ such that $\Sigma \preceq \sigma^2 I$. Let $0 < \epsilon < 1/3$ and $\delta = O(\sqrt{\epsilon})$. Given an ϵ-corrupted set of $N = \Omega((d \log d)/\epsilon)$ samples drawn from D, there is an algorithm that runs in time $\widetilde{O}(Nd)/\mathrm{poly}(\epsilon)$ and outputs a hypothesis vector $\widehat{\mu}$ such that with probability at least $9/10$ it holds $\|\widehat{\mu} - \mu^\star\|_2 \leq O(\sigma\delta) = O(\sigma\sqrt{\epsilon})$.*

Similarly, the sample complexity of our algorithm is best possible within a logarithmic factor, even without corruptions; the $O(\sigma\sqrt{\epsilon})$ error guarantee is known to

[2] To avoid clutter in the relevant expressions, all algorithms in this paper have high constant success probability. By standard techniques, the success probability can be boosted to $1-\tau$, for any $\tau > 0$, at the cost of a $\log(1/\tau)$ increase in the sample complexity.

be information-theoretically optimal, up to constants, even in the infinite sample regime. Previous algorithms [10, 32] gave the same sample complexity and error guarantees, but again with significantly higher time complexities in high dimensions. Specifically, the iterative spectral algorithm of [10] has runtime $\Omega(N^2 \cdot d) = \Omega(d^3/\epsilon^2)$. See Section 1.3 for more detailed comparisons.

We note that an efficient algorithm for robust mean estimation under bounded covariance assumptions has been recently used as a subroutine [31, 12] to obtain robust learners for a wide range of supervised learning problems that can be phrased as stochastic convex programs. This includes linear and logistic regression, generalized linear models, SVMs (learning linear separators under hinge loss), and many others. The algorithm of Theorem 1.2 provides a faster implementation of such a subroutine, hence yields faster robust algorithms for all these problems.

1.3 Related and Prior Work Learning in the presence of outliers is an important goal in statistics and has been studied in the robust statistics community since the 1960s [20]. After several decades of work, a number of sample-efficient and robust estimators have been discovered (see [21, 18] for book-length introductions). For example, the Tukey median [33] is a sample-efficient robust mean estimator for various symmetric distributions [17, 6]. However, it is NP-hard to compute in general [23, 2] and the many heuristics for computing it degrade in the quality of their approximation as the dimension scales [8, 4, 29].

Until recently, all known computationally efficient high-dimensional estimators could only tolerate a negligible fraction of outliers, even for the simplest statistical task of mean estimation. Recent work in the theoretical computer science community [9, 26] gave the first efficient robust estimators for basic high-dimensional unsupervised tasks, including mean and covariance estimation. Since the dissemination of [9, 26], there has been a flurry of research activity on robust learning algorithms in both supervised and unsupervised settings [3, 5, 10, 13, 11, 32, 15, 14, 19, 25, 31, 12, 24, 16, 28, 7].

For the specific task of robust mean estimation, [9] designs two related algorithmic techniques with similar sample complexities and error guarantees: a convex programming method and an iterative spectral outlier removal method (filtering). The former method inherently relies on the ellipsoid algorithm (leading to polynomial, yet impractical, runtimes), while the latter only requires repeated applications of power iteration to compute the highest eigenvalue-eigenvector of a covariance-like matrix. The total number of power iteration calls can be as large as $\Omega(d)$, for constant ϵ, leading to runtimes of the form $\widetilde{\Omega}(Nd^2)$. We note that the filter-based robust mean estimation algorithm, as presented in [9], applies to the sub-gaussian case (as in Theorem 1.1). A slight variant of the method [10] applies under second moment assumptions (as in Theorem 1.2).

The work [26] gives a recursive dimension-halving technique with near-optimal accuracy, up to a logarithmic factor in the dimension. The aforementioned method requires computing the SVD of a second moment matrix $\Omega(\log d)$ times. Consequently, each iteration incurs runtime $\Omega(d^3)$. Similarly, the robust mean estimation algorithm under bounded second moments in [32] requires computing the SVD of a matrix multiple times, leading to $\Omega(d^3)$ runtime.

1.4 Our Approach and Techniques In this section, we provide a detailed outline of our algorithmic approach in tandem with a brief comparison to the most technically relevant prior work. To robustly estimate the unknown mean $\mu^\star$, we proceed as follows: Starting with an initial guess ν, in a sequence of iterations we either certify that the current guess is close to the true mean $\mu^\star$ or refine our current guess with a new one that is provably closer to $\mu^\star$.

Let Σ be the covariance of the good samples. Then we know that the second order moment $\mathbb{E}_{X\sim D}\left[(X-\nu)(X-\nu)^\top\right]$ is equal to Σ when $\nu = \mu^\star$, and is equal to $\Sigma + (\nu-\mu^\star)(\nu-\mu^\star)^\top$ in general. Therefore, the second order moment is minimized when $\nu = \mu^\star$. We use this property to distinguish whether our guess ν is close to $\mu^\star$. Of course, the input contains both good samples and bad (corrupted) samples, and the bad samples can change the first two moments significantly. To get around this problem, we try to reweight the samples: let $\Delta_{N,\epsilon}$ denote the following set

$$\left\{ w \in \mathbb{R}^N : \sum_{i=1}^N w_i = 1 \text{ and } 0 \le w_i \le \tfrac{1}{(1-\epsilon)N} \text{ for all } i \right\}.$$

Our approach will try to minimize the second order moment $\sum_{i=1}^N w_i (X_i-\nu)(X_i-\nu)^\top$ for all $w \in \Delta_{N,\epsilon}$, with the intended solution being assigning $1/|G|$ weight to all the good samples. This can be formalized as an SDP:

(1.1)
$$\begin{aligned} \text{minimize} \quad & \lambda_{\max}\left(\textstyle\sum_{i=1}^N w_i (X_i-\nu)(X_i-\nu)^\top\right) \\ \text{subject to} \quad & w \in \Delta_{N,\epsilon} \end{aligned}$$

This SDP is similar to the convex program used in [9] but has some important conceptual differences that allow us to get a faster algorithm. The convex program

in [9] is essentially this SDP with $\nu = \mu^\star$. However, of course one cannot solve it directly as we do not know $\mu^\star$. To overcome this difficulty, [9] designs a separation oracle, which roughly corresponds to finding a direction of large variance. The whole convex programming algorithm in [9] then relies on ellipsoid algorithm and is therefore slow in high dimensions.

In contrast, we fix a guess ν for the true mean in the SDP. Even though this ν may not be correct, we will show a win-win situation: either ν is a good guess in which case we get a good set of weights, or ν is far from $\mu^\star$ and we can get a ν' that is constant factor closer to $\mu^\star$.

More precisely, we will show that for any guess ν that is sufficiently close to the actual mean $\mu^\star$, the optimal value of the SDP is small. In this case, the weights $\{w_i\}$'s found by the SDP can be used to produce an accurate estimate of the mean: $\widehat{\mu}_w = \sum_{i=1}^n w_i X_i$ (see Lemma 3.2). Note that in this case the estimate $\widehat{\mu}_w$ can be more accurate than the current guess ν. When the guess ν is far from $\mu^\star$, the optimal value of the SDP is large, and the dual solution will give a certificate on why the second order moment $\sum_{i=1}^N w_i(X_i-\nu)(X_i-\nu)^\top$ cannot be small no matter how we reweight the samples using $w \in \Delta_{N,\epsilon}$. Intuitively, the reason that the second moment matrix cannot have small spectral norm is because of the extra component $(\nu-\mu^\star)(\nu-\mu^\star)^\top$ in the expected second moment matrix, so the dual solution gives us information about $\nu - \mu^\star$ (Lemma 3.3).

To get a fast algorithm, we need to solve the SDP and its dual in nearly-linear time. This is done by reducing them to a covering/packing SDPs and use the solver in [1, 30]. The main technical challenge here is that the approximate solutions to the reduced SDPs may violate some of the original constraints (specifically, the resulting w may not be in $\Delta_{N,\epsilon}$). We show that our main arguments are robust enough to handle these mild violations.

A perhaps surprising byproduct of our results is that a natural family of SDPs leads to asymptotically faster algorithms for robust mean estimation than the previous fastest spectral algorithm [9] for the most interesting parameter regime (corresponding to large dimension d so that $d \gg \mathrm{poly}(1/\epsilon)$). We view this as an interesting conceptual implication of our results: in our setting, principled SDP formulations can lead to faster runtimes compared to spectral algorithms, by exploiting the additional structure of these SDPs. This phenomenon illustrates the value of obtaining a deeper understanding of such convex formulations.

1.5 Structure of This Paper In Section 3, we describe our algorithmic approach for robust mean estimation and use it to obtain our algorithm for sub-gaussian distributions (thus establishing Theorem 1.1). In Section 4, we show that the corresponding SDPs can be solved in nearly-linear time. In Section 5, we adapt our approach from Section 3 to obtain our algorithm for robust mean estimation under bounded covariance assumptions (thus establishing Theorem 1.2). For the clarity of the presentation, some proofs have been deferred to an appendix.

2 Preliminaries

We use $[n]$ to denote the set $\{1, \ldots, n\}$. We use e_i for the i-th standard basis vector, and I for the identity matrix. For a vector x, we use $\|x\|_1$ and $\|x\|_2$ to denote the ℓ_1 and ℓ_2 norm of x respectively. We use $\langle x, y\rangle$ to denote the inner product of two vectors x and y: $\langle x, y\rangle = x^\top y = \sum_i x_i y_i$.

For a matrix A, we use $\|A\|_2$ to denote the spectral norm of A, and $\lambda_{\max}$ to denote the maximum eigenvalue of A. We use $\mathrm{tr}(A)$ to denote the trace of a square matrix A, and $\langle A, B\rangle$ or $A \bullet B$ for the entry-wise inner product of A and B: $\langle A, B\rangle = A \bullet B = \mathrm{tr}(A^\top B)$. A symmetric $n \times n$ matrix A is said to be positive semidefinite (PSD) if for all vectors $x \in \mathbb{R}^n$, $x^\top A x \geq 0$. For two symmetric matrices A and B, we write $A \preceq B$ when $B - A$ is positive semidefinite.

Throughout this paper, we use d for the dimension of the distribution in question, N for the number of samples, and ϵ for the fraction of corrupted samples. We use $\mu^\star$ to denote the (unknown) true mean of the distribution in question, and ν to be our current guess for $\mu^\star$. We write X_i for the i-th sample. Both $\mu^\star$, ν, and the X_i's are $d \times 1$ column vectors.

For a vector $w \in \mathbb{R}^N$, we use $\widehat{\mu}_w = \sum_{i\in[N]} w_i X_i$ to denote the empirical mean weighted by w. We use G to denote the set of good samples, and B to denote the set of bad samples corrupted by the adversary. For any vector $w \in \mathbb{R}^N$, we define $w_G = \sum_{i\in G} w_i$ and $w_B = \sum_{i\in B} w_i$.

We call a vector $w \in \mathbb{R}^N$ a uniform distribution over a set $S \subseteq [N]$ if $w_i = \frac{1}{|S|}$ for all $i \in S$ and $w_i = 0$ otherwise. Let $\Delta_{N,\epsilon}$ denote the convex hull of all uniform distributions over subsets $S \subseteq [N]$ of size $|S| = (1-\epsilon)N$. Formally, $\Delta_{N,\epsilon} = \{w \in \mathbb{R}^N : \sum_i w_i = 1 \text{ and } 0 \leq w_i \leq \frac{1}{(1-\epsilon)N} \text{ for all } i\}$.

3 Robust Mean Estimation for Known Covariance Sub-Gaussian Distributions

In this section, we will describe our algorithmic technique and give an algorithm establishing Theorem 1.1.

As we described in Section 1.4, our algorithm is going to make a guess ν on the actual mean $\mu^\star$, and

try to certify its correctness by an SDP. In Section 3.1, we give the SDP formulation and describe the entire algorithm. In Section 3.2, we show that the optimal value of the primal/dual SDPs are closely related to the distance $\|\nu - \mu^\star\|_2$. When the current guess ν is close, we show (Section 3.3) that the solution to the primal SDP is going to give a good estimate of $\mu^\star$. When the current guess ν is far, in Section 3.4 we analyze the dual solution and show how to find a ν' that is closer to $\mu^\star$. Finally, we combine these techniques and prove the main theorem in Section 3.5.

3.1 SDP Formulation and Algorithm Description As we mentioned in Section 1.4, we will use an SDP to try to certify that our current guess ν is close to the true mean $\mu^\star$. To achieve that, we assign weights w_i to the samples while making sure that $w \in \Delta_{N,\epsilon}$. More precisely, the primal SDP with parameter $\nu \in \mathbb{R}^d$ and $\epsilon > 0$ is defined below:

(3.2)
$$\begin{aligned} \text{minimize} \quad & \lambda_{\max}\left(\sum_{i=1}^N w_i (X_i - \nu)(X_i - \nu)^\top\right) \\ \text{subject to} \quad & w \in \Delta_{N,\epsilon} \end{aligned}$$

Intuitively, this SDP tries to re-weight the samples to minimize the second moment matrix $\sum_{i=1}^N w_i (X_i - \nu)(X_i - \nu)^\top$. The intended solution to this SDP is to assign weight $1/|G|$ on each of the good samples. This solution will have a small objective value whenever ν is close to $\mu^\star$.

When ν is far from $\mu^\star$, we need to consider the dual of (3.2). We will first derive the dual of (3.2). The primal SDP is equivalent to

$$\min_{w \in \Delta_{N,\epsilon}} \max_{M \succeq 0, \mathrm{tr}(M)=1} \langle M, \sum_i w_i (X_i - \nu)(X_i - \nu)^\top \rangle$$

Strong duality holds because the primal SDP admits a strictly feasible solution. The dual SDP is

$$\max_{M \succeq 0, \mathrm{tr}(M)=1} \min_{w \in \Delta_{N,\epsilon}} \langle M, \sum_i w_i (X_i - \nu)(X_i - \nu)^\top \rangle$$

Observe that once we fix a dual solution M, it is easy to minimize the objective function over w: we will assign maximum possible weight $w_i = \frac{1}{(1-\epsilon)N}$ to the smallest $(1-\epsilon)N$ inner products. Therefore, the dual SDP can be stated as:

(3.3)
$$\begin{aligned} \text{maximize} \quad & \text{Mean of the smallest } (1-\epsilon)\text{-fraction of} \\ & \qquad \left((X_i - \nu)^\top M (X_i - \nu)\right)_{i=1}^N \\ \text{subject to} \quad & M \succeq 0, \mathrm{tr}(M) \le 1 \end{aligned}$$

The dual SDP (3.3) certifies that there are no good weights that can make the spectral norm small. The intended solution for the dual is $M = yy^\top$, where $y = \frac{\nu - \mu^\star}{\|\nu - \mu^\star\|_2}$ is the direction between ν and $\mu^\star$. Note that when $M = yy^\top$, the value $(X_i - \nu)^\top M (X_i - \nu)$ is exactly the squared norm of the projection in the direction y. Intuitively, if we project the samples onto the direction of y, the mean of the good samples is going to be at distance $\|\nu - \mu^\star\|_2$, so even after removing the farthest ϵ-fraction of the projected samples one cannot make the remaining values of $(X_i - \nu)^\top M (X_i - \nu)$ small. Of course, in general, the dual solution can be of rank higher than 1, but we will show that any near-optimal dual solution must be close to rank 1 later in Section 3.4. The SDPs are parameterized by $\epsilon > 0$ and $\nu \in \mathbb{R}^d$, which is our current guess of the true mean $\mu^\star$. We will solve both SDPs multiple times for different values of $\nu \in \mathbb{R}^d$, and we will update ν iteratively based on the solutions to previous SDPs. Eventually, we will obtain some ν that is close enough to $\mu^\star$, so that the primal SDP is going to provide a good set of weights w, and we can output the weighted empirical mean $\widehat{\mu}_w = \sum_i w_i X_i$.

To avoid dealing with the randomness of the good samples, we require the following deterministic conditions on the good samples (which hold with probability $1 - \tau$) drawn from the sub-gaussian distribution. For all $w \in \Delta_{N,3\epsilon}$, we require the following conditions to hold for $\delta = c_1(\epsilon\sqrt{\log 1/\epsilon})$ and $\delta_2 = c_1(\epsilon \log 1/\epsilon)$ for some universal constant c_1:

$$\begin{aligned} \left\| \sum_{i \in G} w_i (X_i - \mu^\star) \right\|_2 &\le \delta \,, \\ \left\| \sum_{i \in G} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top - I \right\|_2 &\le \delta_2 \,, \end{aligned} \tag{3.4}$$

$$\forall i \in G,\ \|X_i - \mu^\star\|_2 \le O(\sqrt{d \log(N/\tau)}) \,, \tag{3.5}$$

Intuitively, Equations (3.4) show that removing samples will not distort the mean and the covariance by too much. Equation (3.5) says that the good samples are not too far from the true mean.

We note that the above deterministic conditions are identical to the ones used in the convex programming technique of [9] to robustly learn the mean of $\mathcal{N}(\mu^\star, I)$. As noted in [27], the proof of these concentration inequalities does not require the Gaussian assumption, and it directly applies to sub-Gaussian distributions with identity covariance. As shown in Section 2.1.3 of [27], after $N = \Omega(\delta^{-2}(d + \log(1/\tau)))$ samples, these conditions hold with probability at least $1 - \tau$ on the set of good samples.

Throughout the rest of this section, we will assume that the above conditions are satisfied where we set the

parameter τ to be a sufficiently small universal constant; selecting $\tau = 1/30$ suffices for all our arguments.

We are now ready to present our algorithm (Algorithm 1) for robust mean estimation. In this section, we will use $c_1, \ldots, c_7$ to denote universal constants that are independent of N, d, and ϵ. We will give a detailed description on how to set these constants in Appendix A.

Algorithm 1: Robust Mean Estimation for Known Covariance Sub-Gaussian

Input : An ϵ-corrupted set of N samples $\{X_i\}_{i=1}^N$ on $\mathbb{R}^d$ with $N = \widetilde{\Omega}(d/\epsilon^2)$ and $\epsilon < 1/3$.

Output: A vector $\widehat{\mu} \in \mathbb{R}^d$ such that, with probability 9/10, $\|\widehat{\mu} - \mu^\star\|_2 \le O(\epsilon\sqrt{\log(1/\epsilon)})$.

Let $\nu \in \mathbb{R}^d$ be the coordinate-wise median of $\{X_i\}_{i=1}^N$;

for $i = 1$ **to** $O(\log d)$ **do**

  Use Proposition 4.1 to compute either *(i)* A good solution $w \in \mathbb{R}^N$ for the primal SDP (3.2) with parameters ν and 2ϵ; or *(ii)* A good solution $M \in \mathbb{R}^{d\times d}$ for the dual SDP (3.3) with parameters ν and ϵ;

  if *the objective value of w in SDP (3.2) is at most* $1 + c_4(\epsilon \ln(1/\epsilon))$ **then**

    return *the weighted empirical mean* $\widehat{\mu}_w = \sum_{i=1}^N w_i X_i$ (Lemma 3.2);

  else

    Move ν closer to $\mu^\star$ using the top eigenvector of M (Lemma 3.3).

3.2 Optimal Value of the SDPs In this subsection, we will give upper and lower bounds on the optimal value of the SDPs (3.2) and (3.3). Recall that our high-level idea is to use the dual SDP to improve our guess ν, until it is close enough to the true mean $\mu^\star$, and then solve the primal SDP to get a good set of weights. However, we cannot write an if statement based on $r = \|\nu - \mu^\star\|_2$ because we do not know $\mu^\star$.

Lemma 3.1 allows us to estimate r from the optimal value of the SDPs. We will bound the optimal value of the SDPs from both sides using feasible primal and dual solutions. Let $\mathrm{OPT}_{\nu,\epsilon}$ denote the optimal value of the SDPs (3.2), (3.3) with parameters ν and ϵ. The following lemma shows that when ϵ is small and ν is far away from $\mu^\star$, then both the optimal values $\mathrm{OPT}_{\nu,\epsilon}$ and $\mathrm{OPT}_{\nu,2\epsilon}$ are close to $1 + \|\mu^\star - \nu\|_2^2$.

Lemma 3.1 (Optimal Value of the SDPs) *Fix $0 < \epsilon < 1/3$ and $\nu \in \mathbb{R}^d$. Let $\delta = c_1\epsilon\sqrt{\ln(1/\epsilon)}$, $\delta_2 = c_1\epsilon\ln(1/\epsilon)$ and $\beta = \sqrt{\epsilon\ln(1/\epsilon)}$. Let $\{X_i\}_{i=1}^N$ be an ϵ-corrupted set of $N = \Omega(d/\epsilon^2)$ samples drawn from a sub-gaussian distribution with identity covariance. Let $\mathrm{OPT}_{\nu,\epsilon}$ denote the optimal value of the SDPs (3.2), (3.3) with parameters ν and ϵ. Let $r = \|\nu - \mu^\star\|_2$. Then, we have:*

$$(1-\delta_2) + r^2 - 2\delta r \le \mathrm{OPT}_{\nu,2\epsilon} \le \mathrm{OPT}_{\nu,\epsilon} \le (1+\delta_2) + r^2 + 2\delta r \ .$$

In particular, when $r \ge c_2\beta$, we can simplify the above as

$$1 + 0.9r^2 \le \mathrm{OPT}_{\nu,2\epsilon} \le \mathrm{OPT}_{\nu,\epsilon} \le 1 + 1.1r^2 \ .$$

Proof. We first prove the argument for $\mathrm{OPT} = \mathrm{OPT}_{\nu,\epsilon}$.

One feasible primal solution is to set $w_i = \frac{1}{|G|}$ for all $i \in G$ (and $w_i = 0$ for all $i \in B$). Therefore,

$$\begin{aligned}
\mathrm{OPT} &\le \lambda_{\max}\left(\sum_{i=1}^N w_i (X_i - \nu)(X_i - \nu)^\top\right) \\
&= \max_{y\in\mathbb{R}^d, \|y\|_2 = 1} \sum_{i\in G} w_i \langle X_i - \nu, y\rangle^2 \\
&= \max_y \Bigg(\sum_{i\in G} w_i \langle X_i - \mu^\star, y\rangle^2 + \langle \mu^\star - \nu, y\rangle^2 \\
&\qquad + 2\langle \sum_{i\in G} w_i (X_i - \mu^\star), y\rangle \langle \mu^\star - \nu, y\rangle\Bigg) \\
&\le \max_y \left((1+\delta_2) + \langle \mu^\star - \nu, y\rangle^2 + 2\delta\langle \mu^\star - \nu, y\rangle\right) \\
&= (1+\delta_2) + \|\mu^\star - \nu\|_2^2 + 2\delta\|\mu^\star - \nu\|_2 \ .
\end{aligned}$$

The second to last step uses Condition (3.4). [3]

One feasible dual solution is $M = yy^\top$ where $y = \frac{\mu^\star - \nu}{\|\mu^\star - \nu\|_2}$. The dual objective value is the mean of the smallest $(1-\epsilon)$-fraction of $\left((X_i - \nu)^\top M (X_i - \nu)\right)_{i=1}^N$, which is at least

$$\frac{1}{(1-\epsilon)N} \min_{S\subset G, |S| = (1-2\epsilon)N} \sum_{i\in S} (X_i - \nu)^\top M (X_i - \nu) \ .$$

This is because $|G| = (1-\epsilon)N$, the smallest $(1-\epsilon)N$ entries must include S, where S is the smallest $(1-2\epsilon)N$ entries in G. Let $w_i' = \frac{1}{|S|}$ for all $i \in S$ and $w_i' = 0$ otherwise. Note that S is a subset of G so $w_G' = 1$. Also since $|S| = (1-2\epsilon)N$ we know $w' \in \Delta_{N,2\epsilon}$, therefore we

[3] Formally, in order to apply Condition (3.4), $w \in \mathbb{R}^N$ should be supported on the set of all N good samples (before the adversary changed ϵN of them). If a good sample is corrupted, its weight will always be 0 in this proof. We focus on the weights w on the set G, the *remaining* good samples.

have

$$\begin{aligned}\mathrm{OPT} &\geq \sum_{i\in S}\frac{1}{|S|}(X_i-\nu)^\top M(X_i-\nu)\\ &= \sum_{i\in G} w_i'\langle X_i-\nu, y\rangle^2\\ &= \sum_{i\in G} w_i'\langle X_i-\mu^\star, y\rangle^2 + w_G'\left\|\mu^\star-\nu\right\|_2^2\\ &\quad + 2\sum_{i\in G} w_i'\langle X_i-\mu^\star, y\rangle\left\|\mu^\star-\nu\right\|_2\\ &\geq (1-\delta_2)+\left\|\mu^\star-\nu\right\|_2^2 - 2\delta\left\|\mu^\star-\nu\right\|_2 .\end{aligned}$$

Now we consider $\mathrm{OPT}_{\nu,2\epsilon}$. Intuitively, $\mathrm{OPT}_{\nu,2\epsilon} \approx \mathrm{OPT}_{\nu,\epsilon}$ because both SDPs can throw away the bad samples first, and whether we allow them to throw away another ϵ-fraction of good samples should not affect the moments too much. It is easy to see that $\mathrm{OPT}_{\nu,2\epsilon} \leq \mathrm{OPT}_{\nu,\epsilon}$, because the feasible region with parameter 2ϵ is strictly larger ($\Delta_{N,2\epsilon} \supset \Delta_{N,\epsilon}$) for the primal SDP.

It remains to show that the same lower bound holds for $\mathrm{OPT}_{\nu,2\epsilon}$. For the dual SDP with parameter 2ϵ, the objective is the mean of the smallest $(1-2\epsilon)$-fraction of the entries, so we pick S to be the smallest $(1-3\epsilon)N$ entries in G and $w_i' = \frac{1}{(1-3\epsilon)N}$ for all $i \in S$ instead. Note that Condition (3.4) holds for all $w \in \Delta_{N,3\epsilon}$, and the rest of the proof is identical.

To obtain the simpler upper and lower bounds when $r \geq c_2\beta$, we note that the error term $\delta_2 + 2\delta r = \Theta(\epsilon\log(1/\epsilon)) = \Theta(r^2)$, so by increasing c_2 we can get $1+0.9r^2 \leq \mathrm{OPT} \leq 1+1.1r^2$.

3.3 When Primal SDP Has Good Solutions In this section, we show that a good primal solution for any guess ν will give the correct weighted empirical mean. Lemma 3.2 proves the contrapositive statement: if the empirical mean $\widehat{\mu}_w$ under a set of weights w is far away from the true mean $\mu^\star$, then no matter what our current guess ν is, w can never be a good solution. More specifically, we show that the objective value of w is at least $1+\Omega(\delta^2/\epsilon)$. Roughly speaking, we get 1 from the good samples and $\Omega(\delta^2/\epsilon)$ from the bad samples.

We briefly explain why the bad samples contribute $\Omega(\delta^2/\epsilon)$. The empirical mean of the good samples is off by at most δ by Condition (3.4). Now if $\widehat{\mu}_w$ is far away from $\mu^\star$, the bad samples must shift the mean by more than $\Omega(\delta)$. Intuitively, if an ϵ-fraction of the samples distort the mean by δ, on average each of these sample contributes an error of δ/ϵ, which introduces a total error of $\epsilon(\delta/\epsilon)^2 = \delta^2/\epsilon$ in the second moment matrix.

We use $\beta = \sqrt{\epsilon\ln(1/\epsilon)} = \Theta(\sqrt{\delta^2/\epsilon})$ to denote (asymptotically) the distance between ν and $\mu^\star$ at the end of our algorithm. This threshold appears naturally because if $\left\|\nu-\mu^\star\right\|_2 \gg \beta$, then Lemma 3.1 tells us that $\mathrm{OPT}-1 \gg \beta^2 = \delta^2/\epsilon$. This error subsumes the potential error we could get due to the bad samples shifting the mean by more than $\Omega(\delta)$, so we must guess some ν that is $O(\beta)$ from $\mu^\star$ to detect the bad samples. Note that given some ν that has distance $O(\beta)$ to $\mu^\star$, the solution to the primal SDP can give a much better estimate $\widehat{\mu}$ that is $O(\delta) \ll O(\beta)$ away from $\mu^\star$.

Lemma 3.2 (Good Primal Solutions ⇒ Correct Mean) *Fix $0 < \epsilon < 1/3$. Let $\delta = c_1\epsilon\sqrt{\ln(1/\epsilon)}$, $\delta_2 = c_1\epsilon\ln(1/\epsilon)$ and $\beta = \sqrt{\epsilon\ln(1/\epsilon)}$. Let $\{X_i\}_{i=1}^N$ be a set of ϵ-corrupted samples drawn from a sub-gaussian distribution with identity covariance, where $N = \Omega(d/\delta^2)$. For all $w \in \Delta_{N,2\epsilon}$, if $\left\|\widehat{\mu}_w-\mu^\star\right\|_2 \geq c_3\delta$ where $\widehat{\mu}_w = \sum_{i=1}^N w_iX_i$, then for all $\nu \in \mathbb{R}^d$,*

$$\lambda_{\max}\left(\sum_{i=1}^N w_i(X_i-\nu)(X_i-\nu)^\top\right) \geq 1+c_4\beta^2 .$$

Proof. Fix any $w \in \Delta_{N,2\epsilon}$. If $\left\|\mu^\star-\nu\right\|_2 \geq c_5\beta$, then because w is feasible and by Lemma 3.1,

$$\begin{aligned}&\lambda_{\max}\left(\sum_{i=1}^N w_i(X_i-\nu)(X_i-\nu)^\top\right) \geq \mathrm{OPT}_{\nu,2\epsilon}\\ &\geq 1+0.9\left\|\mu^\star-\nu\right\|_2^2 \geq 1+0.9c_5^2\beta^2 \geq 1+c_4\beta^2 .\end{aligned}$$

Therefore, for the rest of this proof, we can assume $\left\|\mu^\star-\nu\right\|_2 < c_5\beta$.

We project the samples along the direction of $(\widehat{\mu}_w-\mu^\star)$. Consider the unit vector $y = (\widehat{\mu}_w-\mu^\star)/\left\|\widehat{\mu}_w-\mu^\star\right\|_2$. To bound from below the maximum eigenvalue, it is sufficient to show that

$$\begin{aligned}y^\top\left(\sum_{i=1}^N w_i(X_i-\nu)(X_i-\nu)^\top\right)y &= \sum_{i=1}^N w_i\langle X_i-\nu, y\rangle^2\\ &\geq 1+\Omega(\delta^2/\epsilon) .\end{aligned}$$

We first bound from below the contribution of the bad

samples by $\Omega(\delta^2/\epsilon)$. By triangle inequality,

$$\begin{aligned}
&\left|\sum_{i\in B} w_i\langle X_i-\nu, y\rangle\right| \\
&\geq \left|\sum_{i\in B} w_i\langle X_i-\mu^\star, y\rangle\right| - w_B\left|\langle \mu^\star-\nu, y\rangle\right| \\
&\geq \left|\sum_{i=1}^{N} w_i\langle X_i-\mu^\star, y\rangle\right| - \left|\sum_{i\in G} w_i\langle X_i-\mu^\star, y\rangle\right| \\
&\quad - 2\epsilon\left\|\mu^\star-\nu\right\|_2 \\
&\geq \left\|\widehat{\mu}_w-\mu^\star\right\|_2 - \delta - 2\epsilon c_5\beta \\
&\geq (c_3 - 1 - 2c_5\frac{\sqrt{\epsilon}}{c_1})\delta \geq c_6\delta\ .
\end{aligned}$$

The last line follows from our choice of y, and the good samples satisfy Condition (3.4). By Cauchy-Schwarz,

$$\begin{aligned}
&\left(\sum_{i\in B} w_i\langle X_i-\nu, y\rangle^2\right)\left(\sum_{i\in B} w_i\right) \\
&\geq \left(\sum_{i\in B} w_i\langle X_i-\nu, y\rangle\right)^2 \geq c_6^2\delta^2.
\end{aligned}$$

Since $w_B \leq 2\epsilon$, we have $\sum_{i\in B} w_i\langle X_i-\nu, y\rangle^2 \geq \frac{c_6^2}{2}(\delta^2/\epsilon)$.

We continue to lower bound the contribution of the good samples to the quadratic form by $1-O(\delta_2) = 1-O(\delta^2/\epsilon)$. This is because the true covariance matrix is I. By Condition (3.4),

$$\begin{aligned}
&\sum_{i\in G} w_i\langle X_i-\nu, y\rangle^2 \\
&= \sum_{i\in G} w_i\left(\langle X_i-\mu^\star, y\rangle^2 + \langle \mu^\star-\nu, y\rangle^2\right. \\
&\qquad \left. +2\langle X_i-\mu^\star, y\rangle\langle \mu^\star-\nu, y\rangle\right) \\
&\geq \sum_{i\in G} w_i\langle X_i-\mu^\star, y\rangle^2 \\
&\qquad + 2\langle \mu^\star-\nu, y\rangle\langle \sum_{i\in G} w_i(X_i-\mu^\star), y\rangle \\
&\geq (1-\delta_2) - 2\delta\left\|\mu^\star-\nu\right\|_2 \\
&\geq 1 - (1+\frac{2c_5\beta}{\sqrt{\ln(1/\epsilon)}})\delta_2 \geq 1-c_7\delta_2\ .
\end{aligned}$$

Putting the good and bad samples together, we have $\sum_{i=1}^N w_i\langle X_i-\nu, y\rangle^2 \geq 1-c_7\delta_2+\frac{c_6^2}{2}(\delta^2/\epsilon) = 1+(\frac{c_1^2c_6^2}{2} - c_1c_7)\beta^2 \geq 1+c_4\beta^2$ as needed.

The constants in the proof are given in Appendix A.

Lemma 3.2 guarantees that any good solution to the primal SDP gives a good set of weights. In other words, whenever we have a solution to the primal SDP whose objective value is at most $1+O(\beta^2)$, we are done because the weighted empirical mean must be close to the true mean.

3.4 When Primal SDP Has No Good Solutions

We now deal with the other possibility: the primal SDP has no good solution. We will show that, in this case, we can move ν closer to $\mu^\star$ by solving the dual SDP (3.3), decreasing $\left\|\nu-\mu^\star\right\|_2$ by a constant factor.

Lemma 3.1 states that $\text{OPT} \approx 1+\left\|\nu-\mu^\star\right\|_2^2$. Intuitively, if the dual SDP throws away all the bad samples, then we know that $\text{OPT} \approx \frac{1}{|G|}\sum_{i\in G}(X_i-\nu)^\top M(X_i-\nu)$. If this quantity also concentrates around its expectation, then

$$\begin{aligned}
1+\left\|\nu-\mu^\star\right\|_2^2 &\approx \text{OPT} \\
&\approx \mathbb{E}_{X\sim\mathcal{N}(\mu^\star, I)}\left[(X-\nu)^\top M(X-\nu)\right] \\
&= \langle M, I+(\nu-\mu^\star)(\nu-\mu^\star)^\top\rangle\ .
\end{aligned}$$

Because $\text{tr}(M)=1$, we can remove 1 from both sides and get $\langle M, (\nu-\mu^\star)(\nu-\mu^\star)^\top\rangle \approx \left\|\nu-\mu^\star\right\|_2^2$. This condition implies that the top eigenvector of M aligns approximately with $(\nu-\mu^\star)$, which provides a good direction for us to move ν.

The following lemma formalizes this intuition. Specifically, Lemma 3.3 shows that despite the error from solving the SDP approximately and the errors in the concentration inequalities, we can still use the top eigenvector of M to move ν closer to $\mu^\star$.

Lemma 3.3 (Good Dual Solutions $\Rightarrow$ Better ν) *Fix $0<\epsilon<1/3$ and $\nu\in\mathbb{R}^d$. Let $\beta=\sqrt{\epsilon\ln(1/\epsilon)}$. Assume we have a solution $M\in\mathbb{R}^{d\times d}$ to the dual SDP (3.3) with parameters ν and ϵ, and the objective value of M is at least $\max(1+0.9c_4\beta^2, (1-\frac{\epsilon}{10})\text{OPT}_{\nu,2\epsilon})$. Then, we can find a vector $\nu'\in\mathbb{R}^d$, such that $\left\|\nu'-\mu^\star\right\|_2 \leq \frac{3}{4}\left\|\nu-\mu^\star\right\|_2$.*

Proof. Because M is a feasible solution to the dual SDP (3.3) with parameters ν and ϵ, we know that $\text{OPT}_{\nu,\epsilon} \geq 1+0.9c_4\beta^2$. When $\text{OPT}_{\nu,\epsilon} \geq 1+0.9c_4\beta^2$, Lemma 3.1 implies that $\left\|\mu^\star-\nu\right\|_2 \geq c_2\beta$ and $(1-\frac{\epsilon}{10})\text{OPT}_{\nu,2\epsilon} \geq 1+0.85\left\|\mu^\star-\nu\right\|_2^2$. Since the objective value is the average of the smallest $(1-\epsilon)N$ entries of $(X_i-\nu)^\top M(X_i-\nu)$, and one way to choose $(1-\epsilon)N$ entries is to focus on the good samples,

$$\begin{aligned}
&1+0.85\left\|\mu^\star-\nu\right\|_2^2 \leq \left(1-\frac{\epsilon}{10}\right)\text{OPT}_{\nu,2\epsilon} \\
&\leq \left(1-\frac{\epsilon}{10}\right)\text{OPT}_{\nu,\epsilon} \leq \frac{1}{|G|}\sum_{i\in G}(X_i-\nu)^\top M(X_i-\nu)\ .
\end{aligned}$$

We know $M \succeq 0$ and $\mathrm{tr}(M) = 1$. Without loss of generality, we can assume M is symmetric. Using Condition (3.4), we can prove that $\langle M, (\mu^\star - \nu)(\mu^\star - \nu)^\top \rangle \geq \frac{3}{4} \|\mu^\star - \nu\|_2^2$:

$$\begin{aligned}
&1 + 0.85 \|\mu^\star - \nu\|_2^2 \\
&\leq \frac{1}{|G|} \sum_{i \in G} (X_i - \nu)^\top M (X_i - \nu) \\
&= \frac{1}{|G|} \sum_{i \in G} \langle M, (X_i - \mu^\star)(X_i - \mu^\star)^\top \\
&\qquad + 2(X_i - \mu^\star)(\mu^\star - \nu) + (\mu^\star - \nu)(\mu^\star - \nu)^\top \rangle \\
&\leq 1 + \delta_2 + 2\delta \|\mu^\star - \nu\|_2 + \langle M, (\mu^\star - \nu)(\mu^\star - \nu)^\top \rangle \\
&\leq 1 + 0.1 \|\mu^\star - \nu\|_2^2 + \langle M, (\mu^\star - \nu)(\mu^\star - \nu)^\top \rangle \ .
\end{aligned}$$

We will continue to show that the top eigenvector of M aligns with $(\nu - \mu^\star)$. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ denote the eigenvalues of M, and let $v_1, \ldots, v_d$ denote the corresponding eigenvectors. The conditions on M implies that $\sum_{i=1}^d \lambda_d = 1$. We decompose $(\mu^\star - \nu)$ and write it as $\mu^\star - \nu = \sum_{i=1}^d \alpha_i v_i$ where $\sum_{i=1}^d \alpha_i^2 = \|\mu^\star - \nu\|_2^2$. Using these decompositions, we can rewrite $\langle M, (\mu^\star - \nu)(\mu^\star - \nu)^\top \rangle = \sum_{i=1}^d \lambda_i \alpha_i^2$.

First observe that $\lambda_1 \geq \frac{3}{4}$, because $\lambda_1 \sum_i \alpha_i^2 \geq \sum_i \lambda_i \alpha_i^2 \geq \frac{3}{4} \|\mu^\star - \nu\|_2^2 = \frac{3}{4} \sum_i \alpha_i^2$. Moreover, because $\frac{3}{4} \sum_i \alpha_i^2 \leq \sum_i \lambda_i \alpha_i^2 \leq \lambda_1 \alpha_1^2 + (1 - \lambda_1)(1 - \alpha_1^2) \leq \frac{3}{4} \alpha_1^2 + \frac{1}{4} \sum_i \alpha_i^2$, we know that $\langle v_1 v_1^\top, (\mu^\star - \nu)(\mu^\star - \nu)^\top \rangle = \alpha_1^2 \geq \frac{2}{3} \sum_i \alpha_i^2$. Thus, we have a unit vector $v_1 \in \mathbb{R}^d$ with $\langle v_1, \mu^\star - \nu \rangle = \alpha_1 \geq \sqrt{2/3} \|\mu^\star - \nu\|_2$, so the angle between v_1 and $\mu^\star - \nu$ is at most $\theta \leq \cos^{-1}(\sqrt{2/3})$.

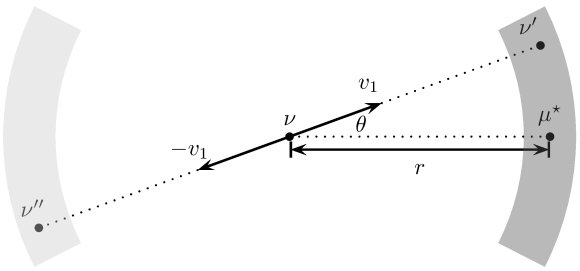

Figure 1: An illustration of the final part of the proof of Lemma 3.3. Assume we can find a unit vector v_1 that approximately aligns with $(\mu^\star - \nu)$, and we can estimate $r \approx \|\mu^\star - \nu\|_2$. Then, the point $\nu' = \nu + r' v_1$ lies in the highlighted region on the right, which is closer to $\mu^\star$. Moreover, if only know $\pm v_1$, we can distinguish ν' and ν'' by comparing the optimal value of their SDPs.

Finally, if we know the exact value of $r = \|\mu^\star - \nu\|_2$, we can update ν to $\nu' = \nu + r v_1$. This corresponds to moving ν to a point that is on a circle of radius r centered at ν (see Figure 1). The distance between ν' and $\mu^\star$ is maximized when θ is the largest, and this distance is at most $2r \sin(\theta/2) < \frac{2}{3} r$. However, in reality, we do not know $r = \|\mu^\star - \nu\|_2$, and we can only estimate it from the value of $\mathrm{OPT}_{\nu, 2\epsilon}$. Because we are solving the SDPs to precision $(1 \pm \frac{\epsilon}{10})$, by Lemma 3.1, we can estimate r' such that $0.85 r^2 \leq (r')^2 \leq 1.15 r^2$. By triangle inequality, the point $\nu' = \nu + r' v_1$ is at most $\frac{2}{3} r + |r' - r| < \frac{3}{4} r$ away from $\mu^\star$.

One technical issue is that the top eigenvector of M can be $\pm v_1$, so we have two possible directions that are opposite of each other. Let $\nu' = \nu + r' v_1$ be the point closer to $\mu^\star$, and $\nu'' = \nu - r' v_1$ be the point farther from $\mu^\star$. We can distinguish ν' and ν'' by solving the SDP (3.2) with parameters ν' and ν'' respectively, and the point with smaller optimal value is ν'. This is because ν'' moves at least $r' \geq \sqrt{0.9} r$ in the reverse direction, so the distance between ν'' and $\mu^\star$ is at least $\sqrt{(r + r' \cos\theta)^2 + (r' \sin\theta)^2} > 1.8r > c_2 \beta$. By Lemma 3.1, $\mathrm{OPT}_{\nu'', 2\epsilon} \geq 1 + 0.9 \cdot (1.8r)^2 > 1 + 2r^2$, and $\mathrm{OPT}_{\nu', 2\epsilon} \leq 1 + 1.1 r^2$. Again because $r \geq c_2 \beta$, this gap is large enough for separating them if we approximate both $\mathrm{OPT}_{\nu', 2\epsilon}$ and $\mathrm{OPT}_{\nu'', 2\epsilon}$ to a factor of $(1 \pm \frac{\epsilon}{10})$.

The constants in the proof are given in Appendix A.

3.5 Proof of Theorem 1.1 We are now ready to prove Theorem 1.1. This is mostly done by applying Lemmas 3.2 and 3.3 in appropriate scenarios. Because of the geometric improvement in Lemma 3.3, we will only apply it logarithmic number of times, and then the algorithm can finish in the case of Lemma 3.2.

Proof. [Proof of Theorem 1.1 (Correctness and Runtime of Algorithm 1)] Let $\tau = 1/30$. When $N = \Omega(d/\epsilon^2)$, Condition (3.4) holds for the good samples with probability at least $1 - \tau$, which is required in the proofs of Lemmas 3.1, 3.2, and 3.3.

We will use the empirical coordinate-wise median as our initial guess ν. It is folklore that with high probability, the coordinate-wise median is within $O(\epsilon\sqrt{d})$ of the true mean $\mu^\star$. In Algorithm 1, whenever we update ν by Lemma 3.3, we must move it closer to $\mu^\star$. Therefore, throughout the algorithm, the condition $\|\nu - \mu^\star\|_2 \leq O(\epsilon\sqrt{d})$ always holds, which is required by Proposition 4.1.

The correctness of Algorithm (1) follows immediately from Lemmas 3.2, 3.3, and Proposition 4.1. In each iteration, the algorithm either finds a good solution $w \in \mathbb{R}^N$ to the primal SDP (3.2) and terminates, in which case Lemma 3.2 guarantees that the weighted empirical mean $\widehat{\mu}_w$ is close to $\mu^\star$; or the algorithm finds a good solution $M \in \mathbb{R}^{d \times d}$ to the dual SDP (3.3), and it will use the top eigenvector of M to move the current guess ν closer to $\mu^\star$ by a constant fac-

tor, as in Lemma 3.3. The failing probability is at most $3\tau = 1/10$ by a union bound over three bad events: *(i)* the good samples do not satisfy Condition (3.4), *(ii)* the coordinate-wise median is too far away from $\mu^\star$, and *(iii)* the SDP solver is not able to produce an approximate solution at some point.

We now analyze the running time of Algorithm 1. The naïve pruning takes time $O(Nd)$. Whenever the primal SDP (3.2) has a good solution, the algorithm terminates. The initial choice of ν satisfies that $\|\nu - \mu^\star\|_2 \leq O(\epsilon\sqrt{d})$, and Lemma 3.1 implies that we have a good primal solution if $\|\mu^\star - \nu\|_2 \leq O(\beta)$. Because every time we move ν as in Lemma 3.3, the distance between ν and $\mu^\star$ decreases by a constant factor, we can move ν at most $O(\log(\epsilon\sqrt{d}/\beta)) = O(\log d)$ times. For each guess $\nu \in \mathbb{R}^d$, we invoke Proposition 4.1 to either obtain a good primal or a good dual solution. We repeat every use of Proposition 4.1 $O(\log\log d)$ times, so that the failing probability is at most $\tau = 1/30$ by a union bound over the iterations. We will use power method to compute the top eigenvector of M, which takes time $O(\log d \cdot Nd\log^2 N/\epsilon^3)$. [4] Thus, every loop of Algorithm 1 takes time $O(\log\log d)\cdot(\widetilde{O}(Nd/\epsilon^6) + \widetilde{O}(Nd/\epsilon^3)) = \widetilde{O}(Nd/\epsilon^6)$. Therefore, the overall running time is

$$O(Nd) + O(\log d)\cdot\widetilde{O}(Nd/\epsilon^6) = \widetilde{O}(Nd/\epsilon^6)\ .$$

4 Solving Primal/Dual SDPs in Nearly-Linear Time

By combining Lemmas 3.2 and 3.3 from Section 3, we know that we can make progress by either finding any solution to the primal SDP (3.2) with objective value at most $1 + c_4\beta^2$, or by finding an approximately optimal solution to the dual SDP (3.3) whose objective value is at least $1 + \frac{9}{10}c_4\beta^2$. This section is dedicated to proving Proposition 4.1, which shows that this can be done in time $\widetilde{O}(Nd)/\operatorname{poly}(\epsilon)$.

PROPOSITION 4.1. *Fix $0 < \epsilon < 1/3$, and $\nu \in \mathbb{R}^d$ with $\|\nu - \mu^\star\|_2 \leq O(\epsilon\sqrt{d})$. Let $\beta = \sqrt{\epsilon\ln(1/\epsilon)}$. We can compute in time $\widetilde{O}(Nd/\epsilon^6)$, with probability at least $9/10$, either*

1. *A solution $w \in \mathbb{R}^n$ for primal SDP (3.2) with parameters $(\nu, 2\epsilon)$, such that the objective value of w is at most $1 + c_4\beta^2$; or*
2. *A solution $M \in \mathbb{R}^{d\times d}$ for dual SDP (3.3) with parameter (ν, ϵ), such that the objective value of M is at least $\max(1 + \frac{9}{10}c_4\beta^2, (1 - \frac{\epsilon}{10})\mathrm{OPT}_{\nu,2\epsilon})$.*

Previously, nearly linear time SDP solvers were developed for packing/covering SDPs [1, 30]. At a high level, we first relate SDPs (3.2), (3.3) with a pair of packing/covering SDPs (4.6), (4.7), where we switch the objective function with some constraint and introduce an additional parameter $\rho > 0$. Next, we show that to prove Proposition 4.1, it is sufficient to solve SDPs (4.6), (4.7) approximately for the correct value of ρ, and moreover, we can run binary search to find a suitable ρ. Finally, in Section 4.1, we show that our packing/covering SDPs (4.6), (4.7) can be solved in time $\widetilde{O}(Nd/\epsilon^6)$. Note that these running times (specifically, the dependence on ϵ) can be improved if better packing/covering SDP solvers are discovered. For example, [1] mentioned the possibility of achieving a bound of $\widetilde{O}(Nd/\epsilon^5)$ by combining their approach and the techniques from [34].

Consider the following packing SDP (4.6) and its dual covering SDP (4.7) with parameters (ν, ϵ, ρ):

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^\top w \\ \text{subject to} & w_i \geq 0, \sum_{i=1}^N w_i A_i \preceq I \end{array} \tag{4.6}$$

$$\begin{array}{ll} \text{minimize} & \mathrm{tr}(M') + \|y'\|_1 \\ \text{subject to} & \rho X_i^\top M' X_i + (1-\epsilon)Ny_i \geq 1 \\ & M' \succeq 0, y' \geq 0\ , \end{array} \tag{4.7}$$

where each $A_i \in \mathbb{R}^{(d+N)\times(d+N)}$ is a PSD matrix given by[5]

$$A_i = \begin{bmatrix} \rho(X_i - \nu)(X_i - \nu)^\top & 0 \\ 0 & (1-\epsilon)N\cdot e_i e_i^\top \end{bmatrix}\ .$$

We will first show that the solutions of (4.6) and (4.7) are closely related to solutions of our original SDPs (3.2) and (3.3). Formally, the following lemma shows that if we (approximately) solve the packing/covering SDPs (4.6), (4.7) for some value of $\rho > 0$ and the resulting objective values are close to 1, then we can translate these solutions back to obtain solutions for SDPs (3.2) (3.3) with objective value roughly $1/\rho$.

[4] The number of iterations of power method is $O(\log d/\epsilon')$ if we want to compute a $(1-\epsilon')$-approximate largest eigenvector. Due to the slack in the geometry analysis of Lemma 3.3, we can set ϵ' to a constant (say $\epsilon' = 0.01$). In addition, matrix M is given implicitly by the positive SDP solver (e.g., [30]) as the sum of matrix exponentials $M = \frac{1}{T}\sum_{t=1}^T \frac{W_t}{\mathrm{tr}(W_t)}$ where $T = O(\log^2 N/\epsilon^3)$ is the number of iterations of the positive SDP solver and $W_t = \exp(\sum_{i=1}^N x_i^t A_i A_i^\top)$ for some $x^t \in \mathbb{R}^N$. To evaluate Mv in the power method, we multiply v with each W_t separately, where we use a degree $O(\log(1/\epsilon'))$ matrix polynomial of $\Phi_t = \sum_{i=1}^N x_i^t A_i A_i^\top$ to approximate $W_t = \exp(\Phi_t)$. It takes time $O(Nd)$ to compute $\Phi_t v$, and therefore it takes time $O(TNd) = O(Nd\log^2 N/\epsilon^3)$ to evaluate Mv.

[5] Recall that $X_i \in \mathbb{R}^{d\times 1}$ is the i-th sample, and $e_i \in \mathbb{R}^{N\times 1}$ is the i-th standard basis vector.

LEMMA 4.1. *Fix $\nu \in \mathbb{R}^d, 0 < \epsilon < 1/3$, and $\rho > 0$. If we have a solution w' of SDP (4.6) with parameters (ν, ϵ) such that $\|w'\|_1 \geq 1 - \frac{\epsilon}{10}$, then we can construct a solution w of SDP (3.2) with parameters $(\nu, 2\epsilon)$ whose objective value is at most $\frac{1}{\rho(1-\epsilon/10)}$. If we have a solution (M', y') of SDP (4.7) with parameters (ν, ϵ) such that $\operatorname{tr}(M') + \|y'\|_1 \leq 1$, then we can construct a solution M of SDP (3.3) with parameters (ν, ϵ) whose objective value is at least $1/\rho$.*

Proof. We first construct a solution w to SDP (3.2) with parameters $(\nu, 2\epsilon)$ given w'. Let $w = \frac{w'}{\|w'\|_1}$. Since $\|w'\|_1 \geq 1 - \frac{\epsilon}{10}$, we know that $w \in \Delta_{N,2\epsilon}$ is feasible for SDP (3.2). SDP (4.6) guarantees that $\rho \sum_i w'_i (X_i - \nu)(X_i - \nu)^\top \preceq I$, so the objective value of SDP (3.2) at w satisfies $\lambda_{\max}\left(\sum_i w_i (X_i - \nu)(X_i - \nu)^\top\right) \leq \frac{1}{\rho(1-\epsilon/10)}$.

Next, we will construct a solution M for the original dual SDP (3.3) with parameters (ν, ϵ) given (M', y'). We will work with the following SDP that is equivalent to the dual SDP (3.3).

$$\begin{array}{ll} \text{maximize} & z - \frac{\sum_{i=1}^N y_i}{(1-\epsilon)N} \\ \text{subject to} & M \succeq 0, \operatorname{tr}(M) \leq 1, y \geq 0, \\ & (X_i - \nu)^\top M (X_i - \nu) + y_i \geq z \end{array}$$

Given a dual solution $M \in \mathbb{R}^{d \times d}$, it is easy to find the optimal values of (y, z) variables: z should be at the $(1-\epsilon)$-th quantile for $(X_i - \nu)^\top M (X_i - \nu)$ and $y_i = \max\{z - (X_i - \nu)^\top M (X_i - \nu), 0\}$. Under these choices, we recover the objective value of SDP (3.3), which is the mean of the smallest $(1-\epsilon)N$ entries of $\left((X_i - \nu)^\top M (X_i - \nu)\right)_{i=1}^N$.

Let $M = \frac{M'}{\operatorname{tr}(M')}$, $y = \frac{(1-\epsilon)N}{\rho \operatorname{tr}(M')} y'$, and $z = \frac{1}{\rho \operatorname{tr}(M')}$. Note that y is well-defined, because we always have $M' \neq 0$, otherwise the objective value is at least $1/(1-\epsilon) > 1$. Note that (M, y, z) is a feasible solution to SDP (3.3): by the definition of (M, y, z), the constraint $\rho X_i^\top M' X_i + (1-\epsilon) N y'_i \geq 1$ translates to $\rho \operatorname{tr}(M') X_i^\top M X_i + \rho \operatorname{tr}(M') y_i \geq 1 = \rho \operatorname{tr}(M') z$, which is exactly $X_i^\top M X_i + y_i \geq z$. The objective value of (M, y, z) is $z - \frac{\|y\|_1}{(1-\epsilon)N} = \frac{1 - \|y'\|_1}{\rho \operatorname{tr}(M')} = 1/\rho$.

The plan is to use binary search to find a suitable $\rho > 0$, solve the packing/covering SDPs approximately, and then translate the solutions back using Lemma 4.1. The translated solutions will satisfy the conditions of Proposition 4.1. To make sure a suitable ρ exists, we use the following lemma which shows the optimal value of SDPs (4.6) (4.7) is continuous and monotone in ρ.

LEMMA 4.2. *Fix $\nu \in \mathbb{R}^d$ and $0 < \epsilon < 1/3$. Let $\mathrm{OPT}_\rho = \mathrm{OPT}_{\rho,\nu,\epsilon}$ be the optimal value of SDPs (4.6), (4.7) with parameters ν, ϵ, and $\rho > 0$. Then, OPT_ρ is continuous and non-increasing in ρ. Moreover, for $\rho^* = 1/\mathrm{OPT}_{\nu,\epsilon}$, we have $\mathrm{OPT}_{\rho^*} = 1$.*

Proof. To prove OPT_ρ is continuous and non-increasing in $\rho > 0$, it is sufficient to show that $\mathrm{OPT}_{\rho_1} \geq \mathrm{OPT}_{\rho_2} \geq (1-\gamma)\mathrm{OPT}_{\rho_1}$ for any $0 < \gamma < 1$ and $\rho_1 = (1-\gamma)\rho_2$. Let w_1, w_2 be the optimal solution that achieves OPT_{ρ_1} and OPT_{ρ_2} respectively. Because $\rho_1 < \rho_2$, w_2 is feasible for SDP (4.6) with parameter ρ_1, and therefore $\mathrm{OPT}_{\rho_1} \geq \|w_2\|_1 = \mathrm{OPT}_{\rho_2}$. Similarly, $(1-\gamma)w_1$ is feasible for SDP (4.6) with parameter ρ_2, because $(1-\gamma)w_1\rho_2 = w_1\rho_1$, and thus $\mathrm{OPT}_{\rho_2} \geq \|(1-\gamma)w_1\|_1 = (1-\gamma)\mathrm{OPT}_{\rho_1}$.

To prove $\mathrm{OPT}_{\rho^*} = 1$, we focus on the original primal SDP (3.2) and the packing SDP (4.6). We first prove $\mathrm{OPT}_{\rho^*} \geq 1$. Consider the optimal solution $w \in \mathbb{R}^N$ of SDP (3.2) with parameters (ν, ϵ). We can verify that w is feasible for SDP (4.6) with parameters (ρ^*, ν, ϵ): The constraint on the bottom-right block of SDP (4.6) states that $w_i \leq \frac{1}{(1-\epsilon)N}$ for all $i \in [N]$, and the constraint on the top-left block is equivalent to $\sum_i w_i X_i X_i^\top \leq \frac{1}{\rho^*} I = \mathrm{OPT}_{\nu,\epsilon} \cdot I$. Now assume if $\mathrm{OPT}_{\rho^*} > 1$. Let w' be the optimal solution to SDP (4.6) with $\|w'\|_1 > 1$. This leads to a contradiction, because $w = w'/\|w'\|_1$ is feasible for SDP (3.2), but its objective value is $\frac{1}{\rho^* \|w'\|_1} < \mathrm{OPT}_{\nu,\epsilon}$.

Now we are ready to prove Proposition 4.1 by putting Lemmas 4.2 and 4.1 together.

Proof. [Proof of Proposition 4.1] Fix $\nu \in \mathbb{R}^d$ and $0 < \epsilon < 1/3$. We first prove that it is sufficient to find some $\rho > 0$, such that we can compute both

(i) a solution $w \in \mathbb{R}^N$ to SDP (3.2) with parameters $(\nu, 2\epsilon)$, whose objective value is $\frac{1}{\rho(1-\epsilon/10)}$;

(ii) a solution $M \in \mathbb{R}^{d \times d}$ to SDP (3.3) with parameters (ν, ϵ), whose objective value is $\frac{1}{\rho}$.

The reason is as follows: Let $\mathrm{ALG}_P = \frac{1}{\rho(1-\epsilon/10)}$ and $\mathrm{ALG}_D = \frac{1}{\rho}$. It must be that either $\mathrm{ALG}_P \leq 1 + c_4\beta^2$ or $\mathrm{ALG}_D \geq 1 + \frac{9}{10}c_4\beta^2$. Assuming $\mathrm{ALG}_P > 1 + c_4\beta^2$ and $\mathrm{ALG}_D < 1 + \frac{9}{10}c_4\beta^2$ leads to the following contradiction: $1 - \frac{\epsilon}{10} = \frac{1/\rho}{1/(\rho(1-\epsilon/10))} = \frac{\mathrm{ALG}_D}{\mathrm{ALG}_P} < 1 - \frac{0.1c_4\beta^2}{1+c_4\beta^2} \leq 1 - \min(\frac{1}{20}, \frac{c_4}{20}\beta^2) \leq 1 - \frac{\epsilon}{10}$. Moreover, because ALG_P is the value of a solution to SDP (3.2) with parameters $(\nu, 2\epsilon)$, we know that $\mathrm{ALG}_D = (1 - \frac{\epsilon}{10})\mathrm{ALG}_P \geq \mathrm{OPT}_{\nu,2\epsilon}$ as needed. We will define a target interval $[\rho_1, \rho_2]$, such that solving SDPs (4.6), (4.7) for any parameters $(\rho \in [\rho_1, \rho_2], \nu, \epsilon)$ will allow us to compute a pair of solutions w' and (M', y') such that:

(i) w' is a solution to packing SDP (4.6) with $\|w'\|_1 \geq 1 - \frac{\epsilon}{10}$; and

(ii) (M', y') is a solution to covering SDP (4.7) with $\mathrm{tr}(M') + \|y'\|_1 \le 1$.

Then, by Lemma 4.1, we can convert these solutions to solutions of SDP (3.2) and (3.3) with values ALG_P and ALG_D.

We can first solve SDP (4.6) with $\rho = 1$, and check if the solution satisfies $\|w'\|_1 \ge 1 - \frac{\epsilon}{10}$. If so, we use Lemma 4.1 to convert w' back to a solution of SDP (3.2) whose objective value is $\frac{1}{1-\epsilon/10} \le 1 + c_4\beta^2$ and we are done. For the rest of the proof, we assume $\mathrm{OPT}_{\rho=1} < 1 - \frac{2\epsilon}{30}$. Fix any $\rho_1 \in \{\rho : \mathrm{OPT}_\rho = 1 - \frac{\epsilon}{30}\}$ and $\rho_2 \in \{\rho : \mathrm{OPT}_\rho = 1 - \frac{2\epsilon}{30}\}$. Note that they are well-defined because OPT_ρ is continuous, $\mathrm{OPT}_{\rho^*} = 1$, and $\mathrm{OPT}_{\rho=1} < 1 - \frac{2\epsilon}{30}$. For any $\rho \in [\rho_1, \rho_2]$, by monotonicity, we must have $\mathrm{OPT}_\rho \in [1 - \frac{2\epsilon}{30}, 1 - \frac{\epsilon}{30}]$. Therefore, if we can solve the packing/covering SDPs (4.6), (4.7) approximately up to a multiplicative factor of $(1 \pm O(\epsilon))$, we can find a primal solution w' with $\|w'\|_1 \ge 1 - \frac{\epsilon}{10}$, as well as a dual solution (M', y') with $\mathrm{tr}(M') + \|y'\|_1 \le 1$.

It remains to show that we can find a suitable ρ and solve the SDPs (4.6) (4.7) in time $\widetilde{O}(Nd/\epsilon^6)$. We can find $\rho \in [\rho_1, \rho_2]$ using binary search: if $\|w'\|_1 < 1 - \frac{\epsilon}{10}$ we will decrease ρ, and if $\mathrm{tr}(M') + \|y'\|_1 > 1$ we will increase ρ.

We first show that the binary search takes $O(\log(d/\epsilon))$ steps. Observe that by monotonicity, $0 < \rho^* \le \rho_1 \le \rho_2 < 1$. When $\|\nu - \mu^\star\|_2 \le O(\epsilon\sqrt{d})$, Lemma 3.1 implies that $\mathrm{OPT}_{\nu,\epsilon} \le O(d)$ and hence $\rho^* = 1/\mathrm{OPT}_{\nu,\epsilon} \ge \Omega(1/d)$. If $\rho_1 \ge (1 - \frac{\epsilon}{30})\rho_2$, then by the same argument in the proof of Lemma 4.2, we must have $\mathrm{OPT}_{\rho_2} \ge (1 - \frac{\epsilon}{30})\mathrm{OPT}_{\rho_1} = (1 - \frac{\epsilon}{30})^2 > \mathrm{OPT}_{\rho_2}$, which is a contradiction. Therefore, the interval $[\rho_1, \rho_2]$ has length at least $\rho_2 - \rho_1 \ge \frac{\epsilon}{30}\rho_2 \ge \Omega(\epsilon\rho^*) = \Omega(\epsilon/d)$. In summary, we start with an interval of length less than 1, and the target interval has length at least $\Omega(\epsilon/d)$, so binary search needs at most $O(\log(d/\epsilon))$ steps.

Finally, we bound from above the running time of the algorithm in this proposition. In each step of the binary search, we solve packing/covering SDPs (4.6), (4.7) for some ρ. We solve these SDPs to precision $(1 \pm O(\epsilon))$ as required in this proof, which takes time $\widetilde{O}(Nd/\epsilon^6)$, by Corollary 4.1 from Section 4.1. We repeat every use of Corollary 4.1 $O(\log\log(d/\epsilon))$ times, so that the failure probability is at most 1/10 when we take a union bound over all $O(\log(d/\epsilon))$ iterations. Eventually, when we have a suitable ρ, we can convert the solution back to solutions for SDPs (3.2) (3.3) using Lemma 4.1. Therefore, the total running time is

$$O(\log(d/\epsilon)) \cdot \widetilde{O}((Nd)/\epsilon^6) \cdot O(\log\log(d/\epsilon)) = \widetilde{O}(Nd/\epsilon^6) .$$

4.1 Positive SDP Solvers In this subsection, we show how to solve packing/covering SDPs (4.6), (4.7) in time $\widetilde{O}(Nd/\epsilon^6)$. It is known that positive (i.e., packing/covering) SDPs can be solved in nearly-linear time and poly-logarithmic number of iterations [22, 1, 30]. Because SDPs (4.6), (4.7) are packing/covering SDPs, we can apply the positive SDP solvers in [30] directly (Corollary 4.1).

Lemma 4.3 (Positive SDP Solver, [30]) *Let $A_1, \ldots, A_n$ be $m \times m$ PSD matrices given in factorized form $A_i = C_i C_i^\top$. Consider the following pair of packing and covering SDPs:*

$$\max_{x \ge 0} \mathbf{1}^\top x \quad \text{s.t. } \sum_{i=1}^n x_i A_i \preceq I \ .$$

$$\max_{Y \succeq 0} \mathrm{tr}(Y) \quad \text{s.t. } A_i \bullet Y \ge 1, \forall i \ .$$

We can compute, with probability at least 9/10, a feasible solution x to the packing SDP with $\mathbf{1}^\top x \ge (1 - \epsilon)\mathrm{OPT}$, and together a feasible solution Y to the covering SDP with $\mathrm{tr}(Y) \le (1 + \epsilon)\mathrm{OPT}$ in time $\widetilde{O}((n + m + q)/\epsilon^6)$, where q is the total number of non-zero entries in the C_i's.

An application of the above lemma yields the following corollary:

COROLLARY 4.1. *Fix $\nu \in \mathbb{R}^d$, $0 < \epsilon < 1/3$, and $0 < \rho \le 1$. We can compute in $\widetilde{O}(Nd/\epsilon^6)$ time, with probability at least 9/10,*

1. *a $(1+O(\epsilon))$-approximate solution w' for the packing SDP (4.6) with parameters (ν, ϵ, ρ); and*
2. *a $(1 - O(\epsilon))$-approximate solution (M', y') for the covering SDP (4.7) with parameters (ν, ϵ, ρ).*

Proof. The input matrices $A_i \in \mathbb{R}^{(d+N)\times(d+N)}$ in the SDPs can be factorized as $A_i = C_i C_i^\top$, where

$$C_i = \begin{bmatrix} \sqrt{\rho}(X_i - \nu) & 0_{d\times(d-1)} & 0_{N\times d} \\ 0_{d\times N} & & \sqrt{(1-\epsilon)N} \cdot e_i e_i^\top \end{bmatrix} .$$

The total number of non-zeros in all C_i's is $q = N(d+1) = O(Nd)$, so by Lemma 4.3, we can solve SDPs (4.6), (4.7) in time $\widetilde{O}(Nd/\epsilon^6)$ with probability 9/10. Note that the dual solution should be maintained implicitly to avoid writing down an $(N + d) \times (N + d)$ matrix: the top-left block of the dual solution Y is M', and the diagonals of the bottom-right block is y'.

5 Robust Mean Estimation under Second Moment Assumptions

In this section, we use the algorithmic ideas from Section 3 to establish Theorem 1.2. The algorithm in

this case is similar to the one for the sub-gaussian case with some important differences, due to the different concentration properties in the two settings.

Note that it suffices to prove Theorem 1.2 under the assumption that $\sigma = 1$, i.e., the covariance satisfies $\Sigma \preceq I$. This is without loss of generality: Given a distribution D with $\Sigma \preceq \sigma^2 I$, we can first divide every sample by σ, run the algorithm to learn the mean, and multiply the output by σ.

Recall that $N = |G|$, and $\Delta_{N,\epsilon}$ is the set $\Delta_{N,\epsilon} = \{w \in \mathbb{R}^N : \sum_i w_i = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N} \text{ for all } i\}$. We require the following condition to hold: there exists $G' \subseteq G$ with $|G'| \ge (1 - \frac{\epsilon}{10})|G|$, such that for all $w \in \Delta_{N,3\epsilon}$,

$$\begin{aligned} &\left\| \sum_{i \in G'} w_i (X_i - \mu^\star) \right\|_2 \le \delta \ , \\ &\left\| \sum_{i \in G'} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top \right\|_2 \le \delta_2 \ , \\ &\forall i \in G', \ \|X_i - \mu^\star\|_2 \le O(\sqrt{d/\epsilon}) \ . \end{aligned} \tag{5.8}$$

where $\delta = c_1\sqrt{\epsilon}$ and $\delta_2 = c_1$ for some universal constants c_1. It follows from Lemma A.18 of [10] that these conditions will be satisfied with high constant probability after $N = \Omega((d \log d)/\epsilon)$ samples. [6] In the rest of this section, we will abuse notation and use G to denote G', the set of good samples that are not too far from $\mu^\star$.

The high-level approach is the same as in learning the mean of sub-gaussian distributions: we maintain $\nu \in \mathbb{R}^d$ as our current guess for the unknown mean $\mu^\star$, and try to move ν closer to $\mu^\star$ by solving the dual SDP (3.3); eventually ν will be close enough, and the primal SDP (3.2) can provide good weights w so that we can output the weighted empirical mean $\widehat{\mu}_w$.

The algorithm will be almost identical to Algorithm 1. The only difference is in the "if" statement, where we need a different threshold to decide if the current primal SDP solution is good (or equivalently, whether our guess of ν is close enough to $\mu^\star$).

[6] We wish to throw away samples that are too far from $\mu^\star$. However, we do not know $\mu^\star$, so we run the following preprocessing step. We start with an ϵ-corrupted set of $2N$ samples and partition them into two sets of N samples, S_1 and S_2. We first compute the coordinate-wise median $\widetilde{\mu}$ of S_1, where $\|\widetilde{\mu} - \mu^\star\|_2 \le O(\epsilon\sqrt{d})$ with high probability. Let B be a ball of radius $O(\sqrt{d/\epsilon})$ around $\mu^\star$. Then we run our algorithm on all samples in $(S_2 \cap B)$. Let D' be the conditional distribution obtained by restricting the domain of D (the true distribution) to B. Notice that the mean of D' is close to that of D, and D' has bounded covariance. Moreover, the good samples in $(S_2 \cap B)$ are drawn i.i.d. from D', and most of the good samples in S_2 are in B, so $S_2 \cap B$ is an $O(\epsilon)$-corrupted set of samples for D'.

In this section, we use $c_1, \ldots, c_6$ to denote universal constants. They can be chosen in a way that is similar to how we set constants for Section 3 in Appendix A. We omit the details.

5.1 Optimal Values of the SDPs The following lemma is similar to Lemma 3.1. Specifically, Lemma 5.1 shows that when $\|\mu^\star - \nu\|_2 \ge c_2\beta$, OPT is approximately $\|\mu^\star - \nu\|_2^2$. The difference is that *(i)* in Section 3, OPT is roughly $1 + \|\mu^\star - \nu\|_2^2$, because we know the true covariance matrix is I, but in this section we only know the second moment matrix is bounded; and *(ii)* we need $\|\mu^\star - \nu\|_2 \ge c_2\beta$ in both settings, but in this section $\beta = \Theta(\sqrt{\delta^2/\epsilon}) = 1$ (rather than $\beta = \sqrt{\epsilon \ln(1/\epsilon)}$ as in Section 3) because the values of δ from the concentration bounds is different.

LEMMA 5.1. *Fix $0 < \epsilon < 1/3$ and $\nu \in \mathbb{R}^d$. Let $\delta = c_1\sqrt{\epsilon}$, $\delta_2 = c_1$, and $\beta = 1$. Let $\{X_i\}_{i=1}^N$ be a set of ϵ-corrupted samples drawn from a distribution on $\mathbb{R}^d$ with $\Sigma \preceq I$, where $N = \Omega((d \log d)/\epsilon)$. Let $\mathrm{OPT}_{\nu,\epsilon}$ denote the optimal value of the SDPs (3.2) (3.3) with parameters (ν, ϵ). Let $r = \|\nu - \mu^\star\|_2$. Then, we have*

$$r^2 - 2\delta r \le \mathrm{OPT}_{\nu,2\epsilon} \le \mathrm{OPT}_{\nu,\epsilon} \le \delta_2 + r^2 + 2\delta r \ .$$

In particular, when $\|\mu^\star - \nu\|_2 \ge c_2\beta$, we have

$$0.9r^2 \le \mathrm{OPT}_{\nu,2\epsilon} \le \mathrm{OPT}_{\nu,\epsilon} \le 1.1r^2 \ .$$

Proof. We take the same feasible primal/dual solutions as in the proof of Lemma 3.1. We get different upper/lower bounds because we use Conditions (5.8) in this section.

Consider a feasible primal solution w with $w_i = \frac{1}{|G|}$ for all $i \in G$ and $w_i = 0$ otherwise.

$$\begin{aligned} &\mathrm{OPT}_{\nu,\epsilon} \\ &\le \max_{y \in \mathbb{R}^d, \|y\|_2 = 1} \left(\sum_{i \in G} w_i \langle X_i - \mu^\star, y \rangle^2 + \langle \mu^\star - \nu, y \rangle^2 \right. \\ &\qquad \left. + 2\langle \sum_{i \in G} w_i (X_i - \mu^\star), y \rangle \langle \mu^\star - \nu, y \rangle \right) \\ &\le \delta_2 + \|\mu^\star - \nu\|_2^2 + 2\delta \|\mu^\star - \nu\|_2 \ . \end{aligned}$$

One feasible dual solution is $M = yy^\top$ where $y = \frac{\mu^\star - \nu}{\|\mu^\star - \nu\|_2}$. Let S denote the $(1-2\epsilon)N$ good samples with smallest $(X_i - \nu)^\top M (X_i - \nu)$. Let $w_i' = \frac{1}{(1-2\epsilon)N}$ for all

$i \in S$ and $w'_i = 0$ otherwise.

$$\begin{aligned}\mathrm{OPT}_{\nu,\epsilon} &\geq \sum_{i\in G} w'_i \langle X_i - \mu^\star, y\rangle^2 + w'_G \left\|\mu^\star - \nu\right\|_2^2 \\ &\quad + 2\sum_{i\in G} w'_i \langle X_i - \mu^\star, y\rangle \left\|\mu^\star - \nu\right\|_2 \\ &\geq \left\|\mu^\star - \nu\right\|_2^2 - 2\delta \left\|\mu^\star - \nu\right\|_2 \ .\end{aligned}$$

The same upper/lower bounds hold for $\mathrm{OPT}_{\nu,2\epsilon}$ as well.

5.2 When Primal SDP Has Good Solutions We prove that if the weighted empirical mean is far away from the true mean, then the value of SDP (3.2) must be large.

The next lemma is similar to Lemma 3.2. The same intuition still holds: if ϵ-fraction of the samples distort the mean by $\Omega(\delta)$, then they must introduce $\Omega(\delta^2/\epsilon)$ error to the second moment matrix. Note that because the true covariance matrix is no longer I, we cannot say anything about the contribution of the good samples.

LEMMA 5.2. *Fix $0 < \epsilon < 1/3$. Let $\delta = c_1\sqrt{\epsilon}$, $\delta_2 = c_1$, and $\beta = 1$. For all $w \in \Delta_{N,2\epsilon}$, if $\left\|\widehat{\mu}_w - \mu^\star\right\|_2 \geq c_3\delta$, then for all $\nu \in \mathbb{R}^d$, $\lambda_{\max}\left(\sum_{i=1}^N w_i(X_i-\nu)(X_i-\nu)^\top\right) \geq c_4\beta^2$.*

Proof. By Lemma 5.1, we know that if $\left\|\mu^\star - \nu\right\|_2 \geq c_5\beta$, we have $\mathrm{OPT}_{\nu,2\epsilon} \geq 0.9\left\|\mu^\star - \nu\right\|_2^2 \geq c_4\beta^2$. Therefore, we can assume that $\left\|\mu^\star - \nu\right\|_2 < c_5\beta$.

Let $w \in \Delta_{N,2\epsilon}$ denote the optimal primal solution. For $y = (\widehat{\mu}_w - \mu^\star)/\left\|\widehat{\mu}_w - \mu^\star\right\|_2$, we have

$$\left|\sum_{i\in B} w_i\langle X_i - \nu, y\rangle\right| \geq \left\|\widehat{\mu}_w - \mu^\star\right\|_2 - \delta - 2\epsilon c_5\beta \geq c_6\delta \ .$$

By Cauchy-Schwarz and $w_B \leq 2\epsilon$, $\sum_{i\in B} w_i\langle X_i - \nu, y\rangle^2 \geq \frac{c_6^2}{2}(\delta^2/\epsilon)$. We conclude the proof by observing that

$$\begin{aligned}\mathrm{OPT} &\geq \sum_{i=1}^N w_i\langle X_i - \nu, y\rangle^2 \\ &\geq \sum_{i\in B} w_i\langle X_i - \nu, y\rangle^2 \geq \frac{c_6^2\delta^2}{2\epsilon} \geq c_4\beta^2 \ .\end{aligned}$$

In summary, if we can find a solution $w \in \mathbb{R}^N$ to the primal SDP (3.2) whose objective value is $O(\beta^2)$, then we are done because Lemma 5.2 guarantees that (no matter what ν is) the weighted empirical mean $\widehat{\mu}_w$ is close to the true mean.

5.3 When Primal SDP Has No Good Solutions We show that when the primal SDP has no good solutions, we can solve the dual (approximately) and the dual will allow us to move ν closer to $\mu^\star$ by a constant factor. The next lemma is similar to Lemma 3.3. The first half changes slightly because we are using Condition (5.8), and the second half (the geometry part) is identical to that of Lemma 3.3.

LEMMA 5.3. *Fix $0 < \epsilon < 1/3$ and $\nu \in \mathbb{R}^d$. Assume $M \in \mathbb{R}^{d\times d}$ is a solution to dual SDP (3.3) with parameters (ν,ϵ), and the objective value of M is at least $\max\left(0.9c_4\beta^2, (1-\frac{\epsilon}{10})\mathrm{OPT}_{\nu,2\epsilon}\right)$. Then, we can find a vector ν' such that $\left\|\nu' - \mu^\star\right\|_2 \leq \frac{3}{4}\left\|\nu - \mu^\star\right\|_2$.*

Proof. By Lemma 5.1, we know that $\mathrm{OPT}_{\nu,\epsilon} \geq 0.9c_4\beta^2$ implies that $\left\|\mu^\star - \nu\right\|_2 \geq c_2\beta$ and $(1-\frac{\epsilon}{10})\mathrm{OPT}_{\nu,2\epsilon} \geq 0.85\left\|\nu - \mu^\star\right\|_2^2$. The dual objective is the mean of the smallest $(1-\epsilon)$-fraction of the entries $X_i^\top M X_i$. Because one way to choose $(1-\epsilon)$-fraction is to focus on the good samples, $\frac{1}{|G|}\sum_{i\in G}(X_i-\nu)^\top M(X_i-\nu) \geq 0.85\left\|\mu^\star - \nu\right\|_2^2$.

We know that $M \succeq 0$, $\mathrm{tr}(M) = 1$. Without loss of generality, we can assume M is symmetric. By Condition (5.8), we can prove $\langle M, (\mu^\star-\nu)(\mu^\star-\nu)^\top\rangle \geq \frac{3}{4}\left\|\mu^\star - \nu\right\|_2^2$ as follows.

$$\begin{aligned}&0.85\left\|\mu^\star - \nu\right\|_2^2 \\ &\leq \frac{1}{|G|}\sum_{i\in G}\langle M, (X_i-\mu^\star)(X_i-\mu^\star)^\top \\ &\qquad + 2(X_i-\mu^\star)(\mu^\star-\nu) + (\mu^\star-\nu)(\mu^\star-\nu)^\top\rangle \\ &\leq \delta_2 + 2\delta\left\|\mu^\star - \nu\right\|_2 + \langle M, (\mu^\star-\nu)(\mu^\star-\nu)^\top\rangle \\ &\leq 0.1\left\|\mu^\star - \nu\right\|_2^2 + \langle M, (\mu^\star-\nu)(\mu^\star-\nu)^\top\rangle \ .\end{aligned}$$

Therefore, we have matrix whose inner product with $(\mu^\star-\nu)(\mu^\star-\nu)^\top$ is approximately maximized, this implies that the top eigenvector of M aligns with $(\nu-\mu^\star)$. We omit the rest of the proof because the geometry analysis is identical to that of Lemma 3.3.

5.4 Proof of Theorem 1.2 By combining Lemmas 5.2 and 5.3, we can make progress by either finding a solution to the primal SDP (3.2) with objective value at most $c_4\beta^2$, or finding an approximately optimal solution to the dual SDP (3.3) whose objective value is at least $0.9c_4\beta^2$. This next proposition shows that this can be done in time $\widetilde{O}(Nd)/\operatorname{poly}(\epsilon)$.

PROPOSITION 5.1. *Fix $0 < \epsilon < 1/3$, and $\nu \in \mathbb{R}^d$ with $\left\|\nu - \mu^\star\right\|_2 \leq O(\sqrt{d/\epsilon})$. We can compute in time $\widetilde{O}(Nd)/\epsilon^6)$, with probability at least $9/10$, either (i) a solution w for primal SDP (3.2) with parameters $(\nu, 2\epsilon)$*

whose objective value is at most $c_4\beta^2$; or (ii) a solution M for dual SDP (3.3) *with parameter* (ν, ϵ) *whose objective is at least* $\max\left(0.9c_4\beta^2, (1-\frac{\epsilon}{10})\mathrm{OPT}_{\nu,2\epsilon}\right)$.

We omit the proof of Proposition 5.1 because its proof is almost identical to the proof of Proposition 4.1. The only difference is that the ratio between the objective values of the desired primal/dual solutions is now $\frac{0.9c_4\beta^2}{c_4\beta^2} = 0.9$, instead of $\frac{1+0.9c_4\beta^2}{1+c_4\beta^2}$ as in Proposition 4.1. The problem of computing a desired pair of solutions becomes easier since the gap is larger.

Theorem 1.2 follows directly from Lemmas 5.2, 5.3, and Proposition 5.1. The running time analysis is identical to that of Theorem 1.1, we can move our guess ν at most $O(\log(d/\epsilon))$ times, and for each guess we invoke Proposition 5.1 to obtain a good primal or dual solution. The overall running time is $\widetilde{O}(Nd\log(1/\tau)/\epsilon^6)$.

6 Conclusions and Future Directions

In this paper, we studied the problem of robust high-dimensional mean estimation for structured distribution families in the presence of a constant fraction of corruptions. As our main technical contribution, we gave the first algorithms with dimension-independent error guarantees for this problem that run in nearly-linear time. We hope that this work will serve as the starting point for the design of fast algorithms for high-dimensional robust estimation.

A number of natural directions suggest themselves: Do our techniques generalize to robust covariance estimation? We believe so, but we have not explored this direction in the current work. Can we obtain nearly-linear time robust algorithms for other inference tasks under sparsity assumptions [3] (e.g., for sparse mean estimation or sparse PCA)? Can we speed-up the convex programs obtained via the SoS hierarchy in this setting [19, 25]?

The running time of our algorithms is $\widetilde{O}(Nd)/\operatorname{poly}(\epsilon)$, i.e., the algorithms run in nearly-linear time only when ϵ is a constant. Can we avoid the extraneous $\operatorname{poly}(1/\epsilon)$ dependence in the runtime? This would require exploiting the problem structure even further, as even solving a single covering SDP incurs a $\operatorname{poly}(1/\epsilon)$ slowdown. We leave this is an interesting question for future work.

Acknowledgments

We would like to thank Alistair Stewart for useful discussions. Yu Cheng is supported in part by NSF grants CCF-1704656, CCF-1527084, CCF-1535972, CCF-1637397, IIS-1447554, and NSF CAREER Award CCF-1750140. Ilias Diakonikolas is supported by NSF CAREER Award CCF-1652862 and a Sloan Research Fellowship. Rong Ge is supported by NSF CCF-1704656.

References

[1] Z. Allen-Zhu, Y. Lee, and L. Orecchia, *Using optimization to obtain a width-independent, parallel, simpler, and faster positive SDP solver*, in Proc. 27th Annual Symposium on Discrete Algorithms (SODA), 2016, pp. 1824–1831.

[2] E. Amaldi and V. Kann, *The complexity and approximability of finding maximum feasible subsystems of linear relations*, Theoretical Computer Science, 147 (1995), pp. 181–210.

[3] S. Balakrishnan, S. S. Du, J. Li, and A. Singh, *Computationally efficient robust sparse estimation in high dimensions*, in Proc. 30th Annual Conference on Learning Theory (COLT), 2017, pp. 169–212.

[4] T. M. Chan, *An optimal randomized algorithm for maximum Tukey depth*, in Proc. 15th Annual Symposium on Discrete Algorithms (SODA), 2004, pp. 430–436.

[5] M. Charikar, J. Steinhardt, and G. Valiant, *Learning from untrusted data*, in Proc. 49th Annual ACM Symposium on Theory of Computing (STOC), 2017, pp. 47–60.

[6] M. Chen, C. Gao, and Z. Ren, *Robust covariance and scatter matrix estimation under Huber's contamination model*, CoRR, abs/1506.00691 (2015).

[7] Y. Cheng, I. Diakonikolas, D. M. Kane, and A. Stewart, *Robust learning of fixed-structure Bayesian networks*, in Proc. 33rd Annual Conference on Neural Information Processing Systems (NIPS), 2018.

[8] K. L. Clarkson, D. Eppstein, G. L. Miller, C. Sturtivant, and S.-H. Teng, *Approximating center points with iterated Radon points*, in Proc. 9th Annual Symposium on Computational Geometry (SoCG), New York, NY, USA, 1993, ACM, pp. 91–98.

[9] 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.

[10] 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, pp. 999–1008. Full version available at https://arxiv.org/abs/1703.00893.

[11] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, *Robustly learning a Gaussian: Getting optimal error, efficiently*, in Proc. 29th ACM-SIAM Symposium on Discrete Algorithms (SODA), 2018, pp. 2683–2702.

[12] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, J. Steinhardt, and A. Stewart, *Sever: A robust meta-algorithm for stochastic optimization*, arXiv preprint arXiv:1803.02815, (2018).

[13] 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.

[14] I. Diakonikolas, D. M. Kane, and A. Stewart, *Learning geometric concepts with nasty noise*, in Proc. 50th Annual ACM Symposium on Theory of Computing (STOC), 2018, pp. 1061–1073.

[15] 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, pp. 1047–1060.

[16] I. Diakonikolas, W. Kong, and A. Stewart, *Efficient algorithms and lower bounds for robust linear regression*, 2019.

[17] D. L. Donoho and M. Gasko, *Breakdown properties of location estimates based on halfspace depth and projected outlyingness*, Ann. Statist., 20 (1992), pp. 1803–1827.

[18] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics. The approach based on influence functions*, Wiley New York, 1986.

[19] 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, pp. 1021–1034.

[20] P. J. Huber, *Robust estimation of a location parameter*, Ann. Math. Statist., 35 (1964), pp. 73–101.

[21] P. J. Huber and E. M. Ronchetti, *Robust statistics*, Wiley New York, 2009.

[22] R. Jain and P. Yao, *A parallel approximation algorithm for positive semidefinite programming*, in Proc. 52nd IEEE Symposium on Foundations of Computer Science (FOCS), 2011, pp. 463–471.

[23] D. S. Johnson and F. P. Preparata, *The densest hemisphere problem*, Theoretical Computer Science, 6 (1978), pp. 93–107.

[24] A. Klivans, P. Kothari, and R. Meka, *Efficient algorithms for outlier-robust regression*, in Proc. 31st Annual Conference on Learning Theory (COLT), 2018, pp. 1420–1430.

[25] 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), 2018, pp. 1035–1046.

[26] 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.

[27] J. Li, *Principled Approaches to Robust Machine Learning and Beyond*, PhD thesis, Massachusetts Institute of Technology, 2018.

[28] L. Liu, Y. Shen, T. Li, and C. Caramanis, *High dimensional robust sparse regression*, CoRR, abs/1805.11643 (2018).

[29] G. L. Miller and D. Sheehy, *Approximate centerpoints with proofs*, Comput. Geom., 43 (2010), pp. 647–654.

[30] R. Peng, K. Tangwongsan, and P. Zhang, *Faster and simpler width-independent parallel algorithms for positive semidefinite programming*, arXiv preprint arXiv:1201.5135v3, (2016).

[31] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar, *Robust estimation via robust gradient estimation*, arXiv preprint arXiv:1802.06485, (2018).

[32] 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, pp. 45:1–45:21.

[33] J. W. Tukey, *Mathematics and picturing of data*, in Proceedings of ICM, vol. 6, 1975, pp. 523–531.

[34] D. Wang, M. W. Mahoney, N. Mohan, and S. Rao, *Faster parallel solver for positive linear programs via dynamically-bucketed selective coordinate descent*, arXiv preprint arXiv:1511.06468, (2015).

Appendix

A Setting Constants in Section 3

In this section, we describe how to set universal constants $c_1, \ldots, c_7$ in Section 3. The constants are set in the following order: c_1, c_2, c_4, c_5, c_7, c_6, and c_3. In this order, every c_i only depends on the constants set before it, and there are only lower bounds on the value of c_i, so we can set c_i to a sufficiently large constant. Note that c_3 is the last constant we choose, and our guarantee at the end of the day is to output some hypothesis vector $\widehat{\mu}$ that is close to the true mean $\mu^\star$: $\|\widehat{\mu} - \mu^\star\|_2 \le c_3\delta$.

Recall that in Section 3, $0 < \epsilon < 1/3$, $\delta = c_1\epsilon\sqrt{\ln(1/\epsilon)}$, $\delta_2 = c_1\epsilon\ln(1/\epsilon)$, and $\beta = \sqrt{\epsilon\ln(1/\epsilon)}$.

The constant c_1 appears in the concentration bounds for the good samples (Condition (3.4)), and it is related to the constants in Chernoff bounds and Hanson-Wright inequality (see, e.g., Section 2.1.3 of [27]). We can set c_1 to be any constant that Condition (3.4) holds with the right sample complexity.

The constant c_2 is a threshold on $r = \|\nu - \mu^\star\|_2$. When $r \ge c_2\beta$, we can show that OPT is roughly $1 + r^2$. We set c_2 to satisfy $\delta_2 + 2\delta(c_2\beta) \le 0.1(c_2\beta)^2$ as required by Lemma 3.1.

The constant c_4 shows up in the branching statement of Algorithm 1. If $\mathrm{OPT} \le 1 + c_4\beta^2$ we use the primal SDP solution, otherwise we use the dual SDP solution. We set c_4 to satisfy $0.9c_4^2 \ge 1.1c_2^2$ in the proof of Lemma 3.3, and $\frac{c_4}{20}\beta^2 \ge \frac{\epsilon}{10}$ in the proof of Proposition 4.1.

If we use the dual solution, we know that $r \geq c_2\beta$. If we use the primal solution, we have $r \leq c_5\beta$. We choose c_5 where $c_5 \geq c_2$ and $0.9c_5^2 \geq c_4$ as needed in the proof of Lemma 3.2.

In the proof of Lemma 3.2, the constants c_6 and c_7 appear when we argue that the bad samples contribute at least $\Omega(\delta^2/\epsilon)$ to the second-moment, and the good samples contribute at least $1 - O(\delta^2/\epsilon)$. We choose c_7 such that $c_7 \geq 1 + \frac{2c_5\beta}{\sqrt{\ln(1/\epsilon)}}$, and c_6 such that $\frac{c_1^2 c_6^2}{2} \geq c_4 + c_1 c_7$. Finally, because the good samples shift the mean by at most δ, if the empirical mean is off by more than $c_3\delta$ then most of the error are from the bad samples. We choose c_3 so that $c_3 \geq c_6 + 1 + \frac{2c_5\sqrt{\epsilon}}{c_1}$.