High-Dimensional Robust Mean Estimation via Gradient Descent

Yu Cheng*
University of Illinois at Chicago
yucheng2@uic.edu

Ilias Diakonikolas†
University of Wisconsin-Madison
ilias@cs.wisc.edu

Rong Ge‡
Duke University
rongge@cs.duke.edu

Mahdi Soltanolkotabi§
University of Southern California
soltanol@usc.edu

May 29, 2020

Abstract

We study the problem of high-dimensional robust mean estimation in the presence of a constant fraction of adversarial outliers. A recent line of work has provided sophisticated polynomial-time algorithms for this problem with dimension-independent error guarantees for a range of natural distribution families.

In this work, we show that a natural non-convex formulation of the problem can be solved directly by gradient descent. Our approach leverages a novel structural lemma, roughly showing that any approximate stationary point of our non-convex objective gives a near-optimal solution to the underlying robust estimation task. Our work establishes an intriguing connection between algorithmic high-dimensional robust statistics and non-convex optimization, which may have broader applications to other robust estimation tasks.

*Part of the work was done while visiting the Institute for Advanced Study.

†Supported by NSF Award CCF-1652862 (CAREER), a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant. Part of this work was done while visiting the Simons Institute for the Theory of Computing during the Summer 2019 program on the Foundations of Deep Learning.

‡Supported by NSF CCF1704656, NSF CCF-1845171 (CAREER), NSF CCF-1934964, a Sloan Fellowship, and a Google Faculty Research Award. Part of the work was done while visiting the Institute for Advanced Study.

§Supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, NSF CCF-CIF grants #1846369 and #1813877, AFOSR-YIP under award #FA9550-18-1-0078, DARPA Learning with Less Labels (LwLL) and Fast Network Interface Cards (FastNICs) programs, and a Google faculty research award. Part of the work was done while visiting the Simons Institute for the Theory of Computing.

1 Introduction

Learning in the presence of outliers is an important goal in machine learning that has become a pressing challenge in a number of high-dimensional data analysis applications, including data poisoning attacks [BNJT10, BNL12, SKL17] and exploratory analysis of real datasets with natural outliers, e.g., in biology [RPW$^+$02, PLJD10, LAT$^+$08]. In both these application domains, the outliers are not "random" but can be arbitrarily correlated, and could exhibit rather complex structures that is essentially impossible to accurately model. Hence, the goal in these settings is to design computationally efficient estimators that can tolerate a small constant fraction of arbitrary outliers.

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

Definition 1.1 (Strong Contamination Model)**.** *Given a parameter $0 < \epsilon < 1/2$ and a distribution family $\mathcal{D}$ 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 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 the above definition is the fraction of corrupted samples and quantifies the power of the adversary. Intuitively, among our samples, an unknown $(1 - \epsilon)$ fraction are generated from a distribution of interest and are called *inliers*, and the rest are called *outliers*.

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]. In contrast, until fairly recently, even the most basic algorithmic questions were poorly understood. Specifically, even for the basic task of high-dimensional mean estimation, all known robust estimators had runtime exponential in the dimension, rendering them ineffective in high-dimensional settings.

Recently, [DKK$^+$16, LRV16] gave the first efficiently computable robust estimators for high-dimensional unsupervised learning tasks, including mean and covariance estimation. Specifically, [DKK$^+$16] obtained the first polynomial-time robust estimators with *dimension-independent* error guarantees, i.e., with error scaling only with the fraction of corrupted samples ϵ and not with the dimensionality of the data. Since the dissemination of these works, there has been a flurry of research activity on algorithmic aspects of high-dimensional robust statistics; see, e.g., [DK19] for a recent survey on the topic.

Despite this exciting progress, the design of efficient robust estimators in high dimensions remains challenging. The difficulty, of course, lies in the non-convexity of the underlying optimization problem. Prior work developed fairly sophisticated algorithmic tools, even for the task of robust mean estimation. These include convex relaxations [DKK$^+$16] and quite subtle iterative spectral methods [DKK$^+$16, LRV16].

A natural and important goal is to understand to what extent such sophisticated methods are indeed necessary or whether much simpler robust learning algorithms exist. In this work, we take a direct optimization view of these problems and ask the following general question:

Is it possible to solve robust estimation tasks by standard first-order methods?

We believe that this question merits investigation in its own right. Moreover, its positive resolution may have significant implications in the practical adoption of robust estimation methods.

Particularly so since prior algorithms are either (1) computationally prohibitive (relying on large convex relaxations), (2) involve carefully crafted parameters that require precise tuning for practical deployment, or (3) are challenging to extend to more sophisticated robust estimation tasks. A tantalizing possibility is the following: For a range of high-dimensional robust estimation tasks, there exists a (natural) non-convex formulation such that gradient descent efficiently converges to a near-optimal solution.

In this paper, we show that this premise is true for the task of high-dimensional robust mean estimation. In robust mean estimation, we are given a set of N ϵ-corrupted samples from an unknown distribution D in a known family $\mathcal{D}$, and we want to output a hypothesis vector $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu^\star\|_2$ is as small as possible, where $\mu^\star$ is the mean of D. For simplicity, we will assume in this discussion that D is an unknown mean and identity covariance Gaussian on $\mathbb{R}^d$. We note that our results hold under more general distributional assumptions, as in [DKK+16, DKK+17].

The goal in robust mean estimation is to develop efficient algorithms whose ℓ_2-error guarantee scales only with ϵ and not with the dimension d. In particular, for the identity covariance Gaussian case, [DKK+16] gave polynomial-time algorithms for the problem that use $N = \widetilde{\Omega}(d/\epsilon^2)$ samples and guarantee error $O(\epsilon\sqrt{\log(1/\epsilon)})$. This error guarantee matches known Statistical Query (SQ) lower bounds [DKS17].

1.1 Overview of Results and Contributions

In this paper, we consider a natural non-convex optimization formulation of high-dimensional robust mean estimation, and show that gradient descent[1] efficiently converges to a near-optimal solution. Specifically, we show that gradient descent converges in a polynomial number of iterations and matches the error guarantee of the best known polynomial-time algorithms for the problem. Our technical contribution lies in showing that *any* approximate stationary point of our non-convex objective suffices – in the sense that it gives a near-optimal solution for the underlying estimation problem.

To describe our non-convex formulation, we require some background. We use the following framework for robust mean estimation, introduced in [DKK+16]. The idea is to assign a non-negative weight to each data point and then find an appropriate combination of weights such that the weighted empirical mean is close to the true mean. The constraint on the chosen weights is that they represent at least a $(1-\epsilon)$-density fractional subset of the dataset. More formally, given datapoints $X_1, \ldots, X_N \in \mathbb{R}^d$ with corresponding data matrix $X \in \mathbb{R}^{d\times N}$, the objective is to find a weight vector $w \in \mathbb{R}^N$ such that $\mu_w = Xw$ is close to $\mu^\star$. The constraint on w is that it belongs in the set

$$\Delta_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N}, \forall i \right\} ,$$

which is the convex hull of all uniform distributions over subsets $S \subseteq [N]$ of size $|S| = (1-\epsilon)N$.

[DKK+16] established a key structural lemma (Lemma 2.1), which formed the basis of their algorithms. Roughly speaking, the lemma states that any weight vector w is a good solution if the spectral norm of the weighted empirical covariance, $\Sigma_w = \sum_{i=1}^N w_i (X_i - \mu_w)(X_i - \mu_w)^\top$, is small. This lemma directly motivates the following non-convex optimization formulation:

$$\text{Min} \;\; \|\Sigma_w\|_2 \;\; \text{subject to} \;\; w \in \Delta_{N,2\epsilon} \tag{1}$$

[1] Throughout, we informally use the term "gradient descent" to refer to variations of gradient descent methods, which involve updates based on a generalized notion of a gradient, e.g., sub-gradient for non-differentiable functions.

It follows from the aforementioned structural lemma that a near-optimal solution w to (1) gives an μ_w that is close to μ^*. The challenge is that the objective function is not convex, hence it is unclear how to efficiently optimize. Faced with this difficulty, prior works on the topic [DKK+16, DKK+17] developed various sophisticated algorithms.

In this paper, we work directly with the natural formulation (1). Despite its non-convexity, we are able to leverage the structure of the problem to show that gradient descent efficiently converges to a good vector w. In more detail, we prove a novel result about the structure of approximate stationary points of this objective.

Theorem 1.2 (informal statement)**.** *Any approximate stationary point w of* (1) *defines an μ_w that is close to $\mu^\star$.*

See Theorem 3.1 for a detailed formal statement. Technically speaking, our statement is more subtle for various reasons, including the fact that the objective function is not differentiable and the domain is constrained. As a result, we require a careful definition of stationarity in our setting.

Given Theorem 1.2, we proceed to show that projected sub-gradient descent converges to an approximate stationary point in a polynomial number of iterations. This step is also somewhat intricate as the function is non-convex, non-smooth and the optimization problem (1) involves constraints. In summary, we establish the following theorem:

Theorem 1.3. *After $\widetilde{O}(N^2d^4)$ iterations, projected sub-gradient descent on* (1) *outputs a point w such that with high probability $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.*

The bound we establish on the convergence rate on the spectral norm objective (1) is polynomially bounded, but relatively slow. Our second main contribution involves considering the "softmax" version of the spectral norm, which has better smoothness properties. An analogous lemma about the structure of stationary points allows us to show a faster rate of convergence for this modified objective.

Theorem 1.4. *After $\widetilde{O}(Nd^3/\epsilon)$ iterations, projected gradient descent on the softmax objective outputs a point w such that with high probability $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.*

As evident from the above result, the additional smoothness of the "softmax" objective allows us to establish a significantly improved bound on the number of iterations.

1.2 Related Work

The algorithmic question of designing efficient robust mean estimators in high-dimensions has been extensively studied in recent years. After the initial papers [DKK+16, LRV16], a number of works [DKK+17, SCV18, CDG18, DHL19, DL19, CDGW19] have obtained algorithms with improved asymptotic worst-case runtimes that work under weaker distributional assumptions on the good data. Moreover, efficient high-dimensional robust mean estimators have been used as primitives for robustly solving a range of machine learning tasks that can be expressed as stochastic optimization problems [PSBR18, DKK+19a].

We compare our approach with the works of [CDG18] and [DHL19] that give the asymptotically fastest known algorithms for robust mean estimation. At a high-level, [CDG18], building on the convex programming relaxation of [DKK+16], proposed a primal-dual approach for robust mean estimation that reduces the problem to a poly-logarithmic number of packing and covering

SDPs. Each such SDP is known to be solvable in time $\widetilde{O}(Nd)$, using mirror descent [ALO16, PTZ16]. [DHL19] build on the iterative spectral approach of [DKK+16]. That work uses the matrix multiplicative weights update method with a specific regularization and dimension-reduction to improve the worst-case runtime.

In contrast to all of the above, we use a natural non-convex formulation of the robust mean estimation task, and show that a standard first-order method provably and efficiently converges to a near-optimal solution. Even though the convergence rates that we establish in this work do not yield the fastest known asymptotic runtimes for the problem, we believe that our approach is conceptually interesting for a number of reasons. First, our theorem regarding stationary points provides novel structural understanding about robust mean estimation and can be viewed as an explanation as to why this problem is polynomially solvable. Second, it is plausible that gradient descent applied in this context is more stable than previously known algorithms and may facilitate the adoption of robust estimation methods in practice. We hope that this work will serve as the starting point for solving other robust estimation tasks via first-order methods.

Finally, we note that there is an increasing literature on developing rigorous guarantees for non-convex optimization problems via gradient descent, e.g., see the recent survey [JK17] for a review of this literature. With a few exceptions [LW11, HSK17], this literature mostly focuses on showing that gradient descent converges to a global optimum starting from a spectral [KMO10, CLS15, TBS+15] or random initialization [GHJY15] in settings where there are no bad local optima. In contrast to most of this literature, in this paper we show that any *stationary* point has good approximation properties so that no specialized or random initialization is necessary. We believe that such a perspective may enable rigorous analysis of many other non-convex optimization problems.

1.3 Roadmap

In Section 2, we set up the necessary notation and provide some background on robust mean estimation. In the next two sections, we focus on the spectral norm objective. In Section 3, we prove our main structural result showing that any stationary point of the spectral norm objective yields a good solution. We also extend this result in Appendix B, showing that in fact, any approximate stationary point yields a sufficiently good solution. In Section 4, we show that gradient descent converges to an approximate stationary point and hence yields a good solution in a polynomial number of iterations. In Appendix C, we prove structural and algorithmic results for the softmax objective, showing that any approximate stationary point of the softmax objective yields a good solution, and we can find an approximate stationary point using projected gradient descent in a polynomial number of iterations. We conclude with future directions in Section 5.

2 Preliminaries and Background

Notation. For $N \in \mathbb{Z}_+$, we denote $[N] := \{1, \dots, N\}$. For a vector x, we use $\|x\|_1$, $\|x\|_2$, and $\|x\|_\infty$ to denote the ℓ_1, ℓ_2, and ℓ_∞ norm of x respectively. For a matrix A, we use $\|A\|_2$ to denote the spectral norm of A.

For two vectors $x, y \in \mathbb{R}^n$, we use $x^\top y = \sum_{i=1}^n x_i y_i$ to denote the inner product of x and y, and we use $x \odot y \in \mathbb{R}^n$ to denote entrywise product of x and y. For a vector $x \in \mathbb{R}^n$, let $\mathrm{diag}(x) \in \mathbb{R}^{n \times n}$ denote a diagonal matrix with x on the diagonal. For a matrix $A \in \mathbb{R}^{n \times n}$, let $\mathrm{diag}(A) \in \mathbb{R}^n$ denote a column vector with the diagonal entries of A.

Let I denote the identity matrix. For a matrix $A \in \mathbb{R}^{n\times n}$, let $\mathrm{tr}(A)$ denote the trace of A. For two matrices A and B of the same dimensions, let $A \bullet B = \langle A, B\rangle = \mathrm{tr}(A^\top B)$ be the entry-wise inner product of A and B. We use $\exp(A)$ to denote the matrix exponential of A.

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is said to be positive semidefinite (PSD) if $x^\top Ax \geq 0$ for all $x \in \mathbb{R}^n$. For two symmetric matrices A and B, we write $A \preceq B$ iff the matrix $B - A$ is positive semidefinite. Let $\Delta_{n\times n}$ be the set of all PSD matrices of trace 1.

Framework. We use N for the number of input samples, d for the dimension of the ground-truth distribution, and ϵ for the fraction of corrupted samples. Given N datapoints $X_1, \ldots, X_N \in \mathbb{R}^d$, we use $X \in \mathbb{R}^{d\times N}$ to denote the sample matrix, where the i-th column of X is X_i.

Given $w \in \mathbb{R}^N$, let $\mu_w = Xw = \sum_{i=1}^N w_i X_i$ denote the weighted empirical mean and let $\Sigma_w = \sum_{i=1}^N w_i (X_i - \mu_w)(X_i - \mu_w)^\top$ denote the weighted empirical covariance. Let $\Delta_{N,\epsilon}$ denote the convex hull of all uniform distributions over subsets $S \subseteq [N]$ of size $|S| = (1-\epsilon)N$:

$$\Delta_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \leq w_i \leq \frac{1}{(1-\epsilon)N}, \forall i \right\} .$$

Every weight vector $w \in \Delta_{N,\epsilon}$ corresponds to a fractional set of $(1-\epsilon)N$ samples.

Background on Robust Mean Estimation. As mentioned in the introduction, our non-convex formulation is directly motivated by the following structural lemma:

Lemma 2.1 ([DKK+16])**.** *Let S be an ϵ-corrupted set of $N = \widetilde{\Omega}(d/\epsilon^2)$ samples from an unknown $\mathcal{N}(\mu^\star, I)$ and $w \in \Delta_{N,2\epsilon}$. If $\lambda_{\max}(\Sigma_w) \leq 1 + \delta$, for some $\delta \geq 0$, then with high probability, we have that $\|\mu^\star - \mu_w\|_2 = O(\sqrt{\epsilon\delta} + \epsilon\sqrt{\log(1/\epsilon)})$.*

As in prior work, we will establish correctness for our algorithms under deterministic conditions on the inliers (good samples) that hold with high probability. Let $G^\star$ denote the original set of N good samples. Let $S = G \cup B$ denote the input samples after the adversary replaced ϵ-fraction of the samples, where $G \subset G^\star$ is the set of remaining good samples and B is the set of bad samples (outliers) added by the adversary. Note that $|G| = (1-\epsilon)N$ and $|B| = \epsilon N$. Given $w \in \mathbb{R}^N$, let $w_G = \sum_{i\in G} w_i$ be the total weight on good samples, and w_B be the total weight on bad samples.

We require the following concentration bounds to hold for the original N good samples $G^\star$ (which happens with high probability when $N = \widetilde{\Omega}(d/\epsilon^2)$). For all $\widehat{w} \in \Delta_{N,3\epsilon}$, we require the following condition to hold for $\delta = O(\epsilon \log(1/\epsilon))$:

$$\left\| \sum_{i\in G^\star} \widehat{w}_i (X_i - \mu^\star)(X_i - \mu^\star)^\top - I \right\|_2 \leq \delta . \tag{2}$$

Condition (2) on *original* samples $G^\star$ implies the following conditions on the *remaining* good samples G. For any weight vector $w \in \Delta_{N,2\epsilon}$ on the ϵ-corrupted set of samples $S = G \cup B$:

$$\left\| \sum_{i\in G} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top - I \right\|_2 \leq \delta . \tag{3}$$

This is because we can define $\widehat{w}$ as follows: $\widehat{w}_i = \frac{w_i}{w_G}$ for all $i \in G$ and $\widehat{w}_i = 0$ for all $i \in B$. Since $w \in \Delta_{N,2\epsilon}$, we have $\|\widehat{w}\|_\infty \leq \frac{\|w\|_\infty}{w_G} = \frac{\|w\|_\infty}{1-w_B} \leq \frac{\|w\|_\infty}{1-|B|\cdot\|w\|_\infty} \leq \frac{1}{(1-3\epsilon)N}$. In other words, $\widehat{w} \in \Delta_{N,3\epsilon}$ and Condition (3) follows directly from Condition (2).

Remark 2.2 (Distributional Assumptions)**.** For simplicity, in this paper we focus on the fundamental setting that the good data are drawn from an unknown mean and identity covariance Gaussian distribution. It should be noted that our structural and algorithmic results hold under more general distributional assumptions. Specifically, Theorem 4.1 immediately applies to identity covariance subgaussian distributions, with the same error guarantees, since it only relies on the concentration bounds (2) and (3) that only require subgaussian tails (see, e.g., [DKK+17].) Moreover, one can modify the proof of our structural results (Theorems 3.1 and 3.2), mutatis-mutandis, to apply (1) for distributions with bounded covariance (i.e., $\Sigma \preceq I$) and match the optimal $O(\sqrt{\epsilon})$ approximation to the mean [DKK+17]; and, (2) more generally, under the (ϵ, δ)-stability condition of [DK19] to yield an $O(\delta)$ ℓ_2-approximation to the mean.

Background and Definitions of Stationarity. Note that the spectral norm is not a differentiable function and therefore we need an alternative definition of stationarity. To address this issue, by the definition of spectral norm, we can define a function $F(w, u) = u^\top \Sigma_w u$ that takes two parameters as input: the weights $w \in \mathbb{R}^N$ and a unit vector $u \in \mathbb{R}^d$. Our non-convex objective $\min_w f(w) := \|\Sigma_w\|_2$ is then equivalent to solving the minimax problem $\min_w \max_u F(w, u)$. The function $\max_u F(w, u)$ is weakly-convex, and we use the following stationary point definition that is common in the weakly-convex optimization literature [Roc70, Roc81, Dru17, DD18, JNJ19].

Definition 2.3 (First-order stationary point)**.** *Let $F(w, u)$ be a function that is differentiable with respect to w for all u. Let $f(w) = \max_u F(w, u)$. Consider the constrained optimization problem $\min_{w \in K} f(w)$, where K is a closed convex set. We say that $w \in K$ is a first-order stationary point if there exists some $u \in \arg\max_v F(w, v)$ such that*

$$(\nabla_w F(w, u))^\top (\widetilde{w} - w) \geq 0 \ \text{ for all } \ \widetilde{w} \in K \ .$$

We also need a notion of an *approximate* stationary point in the sense that the updates from one iteration to the next do not change much. In the unconstrained and differentiable case, such a point can be characterized by the gradient being small. However, the objective function we consider is both non-differentiable and has constraints, so that a proper definition of approximate stationarity is much more subtle. To overcome this, we appeal to tools from conic geometry and notions of stationarity for weakly convex functions [Roc70, Roc81, Dru17, DD18] to define an appropriate notion of approximate stationarity.

To discuss the notion of approximate stationarity that we use, we need to work with a smoothed variant of the objective known as the *Moreau envelope*.

Definition 2.4 (Moreau envelope)**.** *For any function f and closed convex set $\mathcal{K}$, its associated Moreau envelope $f_\beta(w)$ is defined to be the function*

$$f_\beta(w) := \min_{\widetilde{w} \in \mathcal{K}} f(\widetilde{w}) + \beta \|w - \widetilde{w}\|_2^2 \ .$$

The Moreau envelope can be thought of as a form of convolution between the original function f and a quadratic, so as to smoothen the landscape. In particular, when $f(w)$ takes the form of a maximization problem ($f(w) = \max_u F(w, u)$) with F a mapping that is β-smooth in the u parameter ($|\nabla_w F(w, \widetilde{u}) - \nabla_w F(w, u)| \leq \beta \|\widetilde{u} - u\|_2$), the Moreau envelope is also β-smooth [Dru17]. Therefore, the approximate stationarity of the Moreau envelope can be easily defined through its gradient allowing us to define the following notion of approximate stationarity.

Definition 2.5 (Approximate first-order stationary point)**.** *For any function f and closed convex set $\mathcal{K}$ consider its associated Moreau envelope $f_\beta(w)$ per Definition 2.4. we say that a point w is a ρ-approximately stationary point if $\|\nabla f_\beta(w)\|_2 \leq \rho$.*

As mentioned earlier, the spectral norm admits a minimax formulation of the form $f(w) = \max_u F(w,u)$. Furthermore, as detailed in Appendix B, the corresponding function $F(w,u)$ is β-smooth with $\beta = 2\|X\|_2^2$, so that this notion of approximate stationarity can be applied to the objective of interest in this paper.

3 Structural Result: Any Approximate Stationary Point Suffices

In this section, we establish our main structural result, which says that every *approximate* stationary point of (1) must give a μ_w that is close to $\mu^\star$. For simplicity of the exposition, in the main body of this paper, we state and prove a simpler theorem showing that every (exact) stationary point is a good solution.

Theorem 3.1 (Any stationary point is a good solution)**.** *Let S denote an ϵ-corrupted set of N samples drawn from a d-dimensional Gaussian $\mathcal{N}(\mu^\star, I)$ with unknown mean $\mu^\star$. Suppose that S satisfies Lemma 2.1 and Condition* (3)*.*

Let $f(w)$ be the objective function defined in Equation (1)*. For any first-order stationary point $w \in \Delta_{N,2\epsilon}$ of $f(w)$, we have $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.*

We note that while Theorem 3.1 shows that any (exact) stationary point has small objective value, a stronger statement is required for our algorithmic results in the next section. Specifically, we require that any *approximate* stationary point — in the sense of Definition 2.5 — which gradient descent efficiently converges to, also has low objective value. This is accomplished in the next theorem which we prove in Appendix B. Specifically, by appealing to the gradient of the Moreau envelope from Definition 2.4, we extend the proof of Theorem 3.1 to show the following:

Theorem 3.2 (Any approximate stationary point suffices)**.** *Consider the same setting as in Theorem 3.1. Consider the spectral norm objective $f(w) = \|\Sigma_w\|_2$ with $f_\beta(w)$ denoting the corresponding Moreau envelope function per Definition 2.4 with $\beta = 2\|X\|_2^2$. Then, for any $w \in \Delta_{N,2\epsilon}$ satisfying*

$$\|\nabla f_\beta(w)\|_2 = O(\log(1/\epsilon)) \ ,$$

we have $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.

In the remainder of this section, we focus on proving Theorem 3.1 and briefly discuss how this proof can be generalized to prove Theorem 3.2. Our proof is carried out in two steps: (1) We establish a structural lemma which states that every stationary point w must satisfy a *bimodal subgradient* property; (2) We show any point satisfying such property must have a small objective value. Given these two steps, we can conclude any stationary points μ_w is close to $\mu^\star$, by Lemma 2.1.

For the first step, the bimodal subgradient property states that there exists a vector $\nu \in \partial f(w)$ (in the sub-gradient of the function at that stationary point) whose entries divided in two groups of indices such that for any $i \in S^-$ and any $j \in S^+$ we have $\nu_i \leq \nu_j$. Intuitively, S^- contains all indices with positive w_i, so they can potentially be decreased; while S^+ contains all indices with $w_i < \frac{1}{(1-2\epsilon)N}$, so they can potentially be increased. If the bimodal sub-gradient property is violated,

there must be indices $i \in S^-$, $j \in S^+$, where $\nu_i > \nu_j$. In this case, decreasing w_i and increasing w_j would decrease the objective and thus violate stationarity.

For the second step, recall that

$$\Sigma_w = \left(X \operatorname{diag}(w) X^\top - X w w^\top X^\top \right)$$

and $F(w,u) = u^\top \Sigma_w u$. Let us first compute the sub-gradient $\nabla_w F(w,u)$ with respect to a vector u:

$$\nabla_w F(w,u) = X^\top u \odot X^\top u - 2(u^\top X w) X^\top u \; . \tag{4}$$

Our key observation is that the sub-gradient at direction u is equivalent to the gradient of w for the one-dimensional problem with input $(X_i^\top u)_{i=1}^N$. This allows us to effectively reduce our problem to a one-dimensional robust mean estimation problem. This reduction allows us to show that when the objective function is large, then there must be some non-zero weights associated with the corrupted points that are far away from the mean (these points will be in S^-); while on the other hand, S^+ must contain at least ϵ-fraction of the good points. One can then select indices from these two sets to violate the bimodal sub-gradient property.

Fix a first-order stationary point $w \in \Delta_{N,2\epsilon}$. Definition 2.3 implies that there is a corresponding unit vector $u \in \mathbb{R}^d$ such that w is a stationary point of $F(w,u)$. We first state the bimodal sub-gradient property.

Lemma 3.3 (Bimodal sub-gradient property at stationarity)**.** *Fix $w \in \Delta_{N,2\epsilon}$ and a unit vector u with $u^\top \Sigma_w u = \|\Sigma_w\|_2$. Let $S_- = \{i : w_i > 0\}$ and $S_+ = \{i : w_i < \frac{1}{(1-2\epsilon)N}\}$ denote the coordinates of w that can decrease and increase respectively. If w is a first-order stationary point of $F(w,u)$, then*

$$\nabla_w F(w,u)_i \le \nabla_w F(w,u)_j \; ,$$

for all $i \in S_-$ and $j \in S_+$.

Proof. Suppose there is some $i \in S_-$ and $j \in S_+$ such that $\nabla_w F(w,u)_i > \nabla_w F(w,u)_j$, then intuitively we can make $f(w)$ smaller by decreasing w_i and increasing w_j. Formally, let $\widetilde{w} = w + \min(w_i, \frac{1}{(1-2\epsilon)N} - w_j)(e_j - e_i)$ where e_i is the i-th basis vector. We have $\widetilde{w} \in \Delta_{N,2\epsilon}$ and $(\nabla_w F(w,u))^\top (\widetilde{w} - w) < 0$, which violates the assumption that w is a stationary point (Definition 2.3). □

Given Lemma 3.3, we prove Theorem 3.1 by contradiction. We show that if μ_w is far from $\mu^\star$, then w violates the property stated in Lemma 3.3 and therefore cannot be a stationary point. More specifically, we show that, if μ_w is far from $\mu^\star$, then there exists a bad sample with index $j \in S_-$ whose gradient is large (Lemma 3.4). Meanwhile, the concentration bounds in Condition (3) guarantee that there exists a good sample with index $i \in S_+$ whose gradient is small (Lemma 3.5).

Lemma 3.4 (Bad sample with large gradient)**.** *Assume that Condition* (3) *and Lemma 2.1 hold. Fix $w \in \Delta_{N,2\epsilon}$ and a unit vector u with $u^\top \Sigma_w u = \|\Sigma_w\|_2$. Let $r = \|\mu_w - \mu^\star\|_2$ and suppose $r \ge c_2 \epsilon \sqrt{\ln(1/\epsilon)}$. Then there exists some $i \in (B \cap S_-)$ such that*

$$\nabla_w F(w,u)_i - u^\top \mu^\star (\mu^\star - 2\mu_w)^\top u > 2c_3 \cdot \frac{r^2}{\epsilon^2} \; .$$

Here, c_2 and c_3 are universal positive constants.

Lemma 3.5 (Good sample with small gradient)**.** *Consider the same setting as in Lemma 3.4. There is some $j \in (G \cap S_+)$ such that*

$$\nabla_w F(w,u)_j - u^\top \mu^\star (\mu^\star - 2\mu_w)^\top u \le c_3 \cdot \frac{r^2}{\epsilon^2} \, .$$

We defer the proofs of Lemmas 3.4 and 3.5 to Sections 3.1 and 3.2, and we first use these two lemmas to prove Theorem 3.1.

Proof of Theorem 3.1. Suppose that $w \in \Delta_{N,2\epsilon}$ is a first-order stationary point of $f(w)$, and moreover, w is a bad solution where $\|\mu_w - \mu^\star\|_2 \ge c_2\epsilon\sqrt{\ln(1/\epsilon)}$. By Definition 2.3, there exists a unit vector $u \in \mathbb{R}^d$ such that w is a stationary point of $F(w,u)$.

Fix such a vector u. Since Condition (3) and Lemma 2.1 both hold, we can invoke Lemmas 3.4 and 3.5 on (w,u) to find two coordinates $i \in S_-$ and $j \in S_+$ that violate the bimodal subgradient condition in Lemma 3.3. Consequently, w cannot be a stationary point of $F(w,u)$. This leads to a contradiction, and therefore, all first-order stationary points of $f(w)$ are good solutions. □

We now briefly comment on the modifications required to prove Theorem 3.2 (see Appendix B). Theorem 3.2 is proven by first showing (using conic geometry) that for such an approximate stationary point an approximate bimodal sub-gradient property holds. Specifically, we show that the bimodal sub-gradient property (Lemma 3.3) is *stable* in the sense that for an approximate stationary point an *approximate bimodal sub-gradient* property holds, i.e., $\nu_i \le \nu_j + \delta$. Further, for any point obeying such an approximate bimodal property, the objective is small and has good approximation guarantees. The last two steps when combined show that any approximate stationary point has good approximation guarantees (similar to the proof of Theorem 3.1 for exact stationary points).

3.1 Finding a Bad Sample With Large Gradient

In this subsection, we prove Lemma 3.4.

Lemma 3.4 states that when μ_w is far from $\mu^\star$, there exists an index $i \in (B \cap S_-)$ such that the gradient $\nabla_w F(w,u)_i$ is relatively large.

Recall that $\nabla_w F(w,u)$ in Equation (4) is the same as the gradient of the variance (weighted by w) of the one-dimensional samples $\left(X_i^\top u\right)_{i=1}^N$. Roughly speaking, for this one-dimensional problem, a sample far from the (projected) true mean should have large gradient. Our objective is to find such a sample with positive weight.

More specifically, since w is a bad solution and u is in the top eigenspace of Σ_w, the weighted empirical variance of the projected samples is very large. Because the good samples cannot have this much variance, most of the variance comes from the bad samples. We show that among the bad samples that contribute a lot to the variance, one of them must be very far from the (projected) true mean.

In this section and Section 3.2, we use $c_1, \ldots, c_4$ to denote universal constants that are independent of N, d, and ϵ. We give a detailed description of how to set these constants in Appendix A.

Proof of Lemma 3.4. We first show that the variance of one-dimensional samples $\left(X_i^\top u\right)_{i=1}^N$ is relatively large.

By Lemma 2.1, we know that if $\|\mu_w - \mu^\star\|_2 \geq r$ and $r \geq c_2\epsilon\sqrt{\ln(1/\epsilon)}$, then

$$\lambda_{\max}(\Sigma_w) \geq 1 + c_4 \cdot \frac{r^2}{\epsilon}$$

for some universal constant c_4.

Because u is a unit vector that maximizes $u^\top \Sigma_w u$, we have

$$u^\top \Sigma_w u = \lambda_{\max}(\Sigma_w) \geq 1 + \frac{c_4 r^2}{\epsilon} \, .$$

Recall that $\Sigma_w = \sum_{i=1}^N w_i (X_i - \mu_w)(X_i - \mu_w^\top)$. If we replace μ_w with $\mu^\star$, we have

$$\sum_{i=1}^N w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top \succeq \Sigma_w \, ,$$

and therefore,

$$u^\top \left(\sum_{i=1}^N w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top \right) u \geq 1 + \frac{c_4 r^2}{\epsilon} \, .$$

Next we show that most of this variance is due to bad samples. By Condition (3),

$$u^\top \left(\sum_{i \in G} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top \right) u \leq 1 + c_1 \epsilon \ln(1/\epsilon) \, .$$

Consequently,

$$u^\top \left(\sum_{i \in B} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top \right) u \geq \frac{c_4 r^2}{\epsilon} - c_1 \epsilon \ln(1/\epsilon) \geq 0.98 \cdot c_4 \cdot \frac{r^2}{\epsilon} \, .$$

The last step is because $r \geq c_2 \cdot \epsilon\sqrt{\ln(1/\epsilon)}$ and we can choose c_4 to be sufficiently large.

Now that we know most of the variance is due to the bad samples, observe that the total weight w_B on the bad samples is at most $\epsilon N \cdot \frac{1}{(1-2\epsilon)N} \leq 2\epsilon$. Therefore, there must be some $i \in B$ with $w_i > 0$ such that

$$u^\top \left((X_i - \mu^\star)(X_i - \mu^\star)^\top \right) u \geq \frac{0.98 \cdot c_4 \cdot r^2 \cdot \epsilon^{-1}}{w_B} \geq 0.49 \cdot c_4 \cdot \frac{r^2}{\epsilon^2} \, .$$

In other words,

$$\left| u^\top (X_i - \mu^\star) \right| \geq 0.7 \cdot \sqrt{c_4} \cdot \frac{r}{\epsilon} \, .$$

By definition, $i \in B \cap S_-$. It remains to show that $\nabla_w F(w,u)_i$ is large.

$$\begin{aligned}
&\nabla_w F(w,u)_i - u^\top \mu^\star (\mu^\star - 2\mu_w)^\top u \\
&= u^\top \left((X_i - \mu^\star)(X_i - \mu^\star)^\top \right) u - 2u^\top \left((X_i - \mu^\star)(\mu_w - \mu^\star)^\top \right) u \\
&\geq \left(u^\top (X_i - \mu^\star) \right)^2 - 2 \left| u^\top (X_i - \mu^\star) \right| \cdot \|\mu_w - \mu^\star\|_2 \\
&\geq \frac{0.49 \cdot c_4 \cdot r^2}{\epsilon^2} - 2 \cdot \frac{0.7 \cdot \sqrt{c_4} \cdot r}{\epsilon} \cdot r > 2c_3 \cdot \frac{r^2}{\epsilon^2} \, .
\end{aligned}$$

The first inequality is by Cauchy-Schwarz. The last step uses the fact that ϵ is sufficiently small. $\square$

3.2 Finding a Good Sample With Small Gradient

In this subsection, we prove Lemma 3.5.

Lemma 3.5 states that there exists an index $j \in (G \cap S_+)$ such that the gradient $\nabla_w F(w,u)_j$ is relatively small. Similar to the previous section, a sample close to the (projected) true mean should have small gradient. Our goal is to find such a sample for which we can increase its weight.

Recall that S^+ contains all samples whose weight can be increased. We first prove that there are at least ϵN good samples in S^+. Among these ϵN good samples, the concentration bounds imply that some X_j must be very close to the (projected) true mean.

Proof of Lemma 3.5. Recall that S^+ contains every coordinate i where $w_i < \frac{1}{(1-2\epsilon)N}$. Since at most $(1-2\epsilon)N$ samples can have the maximum weight $\frac{1}{(1-2\epsilon)N}$, we know that $|S^+| \geq 2\epsilon N$. Combining this with $|G| = (1-\epsilon)N$, we know that $|G \cap S^+| \geq \epsilon N$.

Fix a subset $G^+ \subseteq (G \cap S^+)$ of size $|G^+| = \epsilon N$. We first show that on average, samples in G^+ do not contribute much to the variance.

Let w' be the uniform weight vector on G, i.e., $w'_i = \frac{1}{(1-\epsilon)N}$ for all $i \in G$ and $w'_i = 0$ otherwise. Since $w' \in \Delta_{N,2\epsilon}$, by Condition (3),

$$\left\| \sum_{i \in G} \frac{1}{|G|} (X_i - \mu^\star)(X_i - \mu^\star)^\top - I \right\|_2 \leq c_1 \cdot \epsilon \ln(1/\epsilon) \; .$$

Let w'' be the uniform weight vector on $S \setminus G^+ = (G \setminus G^+) \cup B$, i.e., $w''_i = \frac{1}{(1-\epsilon)N}$ for all $i \in ((G \setminus G^+) \cup B)$ and $w''_i = 0$ otherwise. Since $w'' \in \Delta_{N,2\epsilon}$, again by Condition (3), we have

$$\left\| \sum_{i \in G \setminus G^+} \frac{1}{|G|} (X_i - \mu^\star)(X_i - \mu^\star)^\top - I \right\|_2 \leq c_1 \epsilon \ln(1/\epsilon).$$

Combining the previous two concentration bounds,

$$\begin{aligned}
&\left\| \sum_{i \in G^+} \frac{1}{|G|} (X_i - \mu^\star)(X_i - \mu^\star)^\top \right\|_2 \\
&\leq \left\| \sum_{i \in G} \frac{1}{|G|} (X_i - \mu^\star)(X_i - \mu^\star)^\top - I \right\|_2 + \left\| \sum_{i \in G \setminus G^+} \frac{1}{|G|} (X_i - \mu^\star)(X_i - \mu^\star)^\top - I \right\|_2 \\
&\leq 2c_1 \cdot \epsilon \ln(1/\epsilon) \; .
\end{aligned}$$

Consequently, because u is a unit vector,

$$u^\top \left(\sum_{i \in G^+} \frac{1}{|G|} (X_i - \mu^\star)(X_i - \mu^\star)^\top \right) u \leq 2c_1 \epsilon \ln(1/\epsilon) \; .$$

At this point, we know samples in G^+ do not contribute much to the variance. We now proceed to show that one of these samples satisfies the lemma.

Let $j = \arg\min_{i \in G^+} \left|u^\top (X_i - \mu^\star)\right|$. We have

$$u^\top \left((X_j - \mu^\star)(X_j - \mu^\star)^\top\right) u \le \frac{|G|}{|G^+|} \cdot 2c_1 \cdot \epsilon \ln(1/\epsilon) \le 2c_1 \ln(1/\epsilon) \; .$$

Finally, because $\left|u^\top (X_j - \mu^\star)\right| \le \sqrt{2c_1 \ln(1/\epsilon)}$, we can show that $\nabla_w F(w,u)_j$ is small:

$$\begin{aligned}
&\nabla f(w)_j - \mu^{\star\top} Y(\mu^\star - 2\mu_w) \\
&= u^\top \left((X_j - \mu^\star)(X_j - \mu^\star)^\top\right) u + 2u^\top \left((X_j - \mu^\star)(\mu_w - \mu^\star)^\top\right) u \\
&\le 2c_1 \ln(1/\epsilon) + 2\sqrt{2c_1 \ln(1/\epsilon)} \cdot r \\
&\le \frac{c_3}{2} \cdot \frac{r^2}{\epsilon^2} + \frac{c_3}{2} \cdot \frac{r}{\epsilon} \cdot r \le c_3 \cdot \frac{r^2}{\epsilon^2} \; .
\end{aligned}$$

The last step uses that c_3 is sufficiently large, as well as the fact that $\ln(1/\epsilon) \le \frac{r^2}{\epsilon^2}$ because $r \ge c_2 \epsilon \sqrt{\ln(1/\epsilon)}$. □

4 Algorithmic Result: Finding a Stationary Point via Gradient Descent

In this section, we show that a simple Projected Gradient Descent (PGD) algorithm (Algorithm 1) can efficiently find an approximate stationary point w of our spectral norm objective, and that w is a good solution to our robust mean estimation task.

Algorithm 1 Robust Mean Estimation via PGD

Input: ϵ-corrupted set of N samples $\{X_i\}_{i=1}^N$ on $\mathbb{R}^d$ satisfying Condition (3), and $\epsilon < \epsilon_0$.
Output: $w \in \mathbb{R}^N$ with $\|\mu_w - \mu^\star\|_2 \le O(\epsilon\sqrt{\log(1/\epsilon)})$.
Let $F(w,u) = u^\top \Sigma_w u$.
Let w_0 be an arbitrary weight vector in $\Delta_{N,2\epsilon}$.
Let $T = \widetilde{O}(N^2 d^4)$.
for $\tau = 0$ **to** $T-1$ **do**
  Find a unit vector $u_\tau \in \mathbb{R}^d$ such that $F(w_\tau, u_\tau) \ge (1-\epsilon)\max_u F(w_\tau, u)$.
  $w_{\tau+1} = \mathcal{P}_{\Delta_{N,2\epsilon}}(w_\tau - \eta \nabla_w F(w_\tau, u_\tau))$, where $\mathcal{P}_{\mathcal{K}}(\cdot)$ is the ℓ_2 projection operator onto $\mathcal{K}$.
end for
return $w_{\tau^\star}$ where $\tau^\star = \arg\min_{0 \le \tau < T} \|\Sigma_{w_\tau}\|_2$.

We note that finding the unit vector u_τ required in the for loop of Algorithm 1 can be done in time $O(Nd\log(d)/\epsilon)$. Given the PSD matrix $A = \Sigma(w_\tau)$, we want to find a unit vector $u \in \mathbb{R}^d$ such that $u^\top A u \ge (1-\epsilon)\max_v (v^\top A v)$. This is the (approximate) largest eigenvector problem which can be solved via power method in $O(\log(d)/\epsilon)$ iterations. Since the matrix-vector multiplication $Av = \Sigma_{w_\tau} v = \left(X \operatorname{diag}(w_\tau) X^\top - X w_\tau w_\tau^\top X^\top\right) v$ can be computed in time $O(Nd)$, the running time for finding such a vector u_τ is $O(Nd\log(d)/\epsilon)$.

The main result of this section is the following theorem:

Theorem 4.1 (Gradient descent finds a good solution)**.** *Let S be an ϵ-corrupted set of $N = \widetilde{\Omega}(d/\epsilon^2)$ samples from a d-dimensional Gaussian $\mathcal{N}(\mu^\star, I)$ with unknown mean $\mu^\star$. Suppose S satisfies Condition* (3) *and Lemma 2.1. Then, after $\widetilde{O}(N^2 d^4)$ iterations, Algorithm 1 outputs a weight vector $w \in \mathbb{R}^N$ such that $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.*

We first give a high-level overview of the proof. Our proof of Theorem 4.1 can be divided into two steps:

1. The first step is an immediate consequence of Theorem 3.2, which allows us to conclude that any approximate stationary point (in the sense of Definition 2.5) has good approximation guarantees.

2. To finalize the proof, in the second step we show that simple iterative procedures such as (sub)gradient descent can converge in a polynomial number of iterations to such an approximate stationary point. We prove such a result by utilizing a simple and well-known observation: a minimax optimization problem which is smooth in the minimization parameter is weakly convex (after maximization) in the minimization parameter. This connection allows us to leverage recent literature [Dru17, DD18] that provides convergence guarantees for weakly convex optimization problems to prove our algorithm finds an approximate stationary point in a polynomial number of iterations.

To elaborate further, in the second step of our proof, we utilize and slightly generalize[2] the analysis of [DD18] and prove that projected sub-gradient descent can find an approximate stationary point.

Lemma 4.2. *Let $\mathcal{K}$ be a closed convex set. Let $F(w, u)$ be a function which is L-Lipschitz and β-smooth with respect to w. Consider the following optimization problem $\min_{w\in\mathcal{K}} \max_{\|u\|_2=1} F(w, u)$.*

Starting from any initial point $w_0 \in \mathcal{K}$, we run iterative updates of the form:

$$\begin{aligned} &\text{find } u_\tau \text{ with } F(w_\tau, u_\tau) \geq (1-\epsilon') \max_u F(w_\tau, u_\tau) \\ &w_{\tau+1} = \mathcal{P}_{\mathcal{K}}(w_\tau - \eta \nabla_w F(w_\tau, u_\tau) \end{aligned}$$

for T iterations with step size $\eta = \frac{\gamma}{\sqrt{T}}$. Then, we have

$$\min_{0\leq\tau<T} \|\nabla f_\beta(w_\tau)\|_2^2 \leq \frac{2}{\sqrt{T}} \left(\frac{f_\beta(w_0) - \min_w f(w)}{\gamma} + \gamma\beta L^2 \right) + 4\beta\epsilon'$$

where $f_\beta(w)$ is the Moreau envelope as in Definition 2.4.

As shown in Appendix B, $F(w, u)$ associated with $f(w)$ obeys the required Lipschitz and smoothness property, with $L = \widetilde{O}(\sqrt{N}d)$ and $\beta = \widetilde{O}(Nd)$. In addition, we have $0 \leq f(w) \leq \widetilde{O}(d)$ for all $w \in \Delta_{N,2\epsilon}$. Thus, we can apply the result above with the constraint $\mathcal{K} = \Delta_{N,2\epsilon}$. Theorem 4.1 follows by combining Theorem 3.2 and Lemma 4.2. We defer the proofs to Appendix B.

[2]The generalization is to deal with constraints and handle the fact that the inner maximization is not solved precisely.

5 Discussion

The main conceptual contribution of this work is to establish an intriguing connection between algorithmic high-dimensional robust statistics and non-convex optimization. Specifically, we showed that high-dimensional robust mean estimation can be efficiently solved by directly applying a first-order method to a natural non-convex formulation of the problem.

The main technical contribution of this paper is in showing that any approximate stationary point of our non-convex objective suffices to solve the underlying learning problem. Our novel structural result may be viewed as an explanation as to why robust mean estimation can be solved efficiently in high dimensions, despite its non-convexity. Specifically, we establish that the optimization landscape of our non-convex objective is well-behaved, in a precise sense.

There are a number of directions along which our results could be improved. At the technical level, it would be interesting to obtain faster convergence rates for gradient descent (or other first-order methods), with linear convergence as the ultimate goal. We note that our upper bound is fairly loose and we did not make an explicit effort to optimize the polynomial dependence.

A natural direction is to extend our approach to more general robust estimation tasks, including covariance estimation [DKK+16, CDGW19], sparse PCA [BDLS17, DKK+19b], and robust regression [KKM18, DKS19]. Such generalizations will appear in a followup work.

6 Acknowledgments

We thank Jelena Diakonikolas for sharing her expertise in optimization.

References

[ALO16] 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)*, pages 1824–1831, 2016.

[BDLS17] 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*, pages 169–212, 2017.

[Bec17] A. Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[BNJT10] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar. The security of machine learning. *Machine Learning*, 81(2):121–148, 2010.

[BNL12] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2012.

[CDG18] Y. Cheng, I. Diakonikolas, and R. Ge. High-dimensional robust mean estimation in nearly-linear time. *CoRR*, abs/1811.09380, 2018. Conference version in SODA 2019, p. 2755-2771.

[CDGW19] Y. Cheng, I. Diakonikolas, R. Ge, and D. P. Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory, COLT 2019*, pages 727–757, 2019.

[CLS15] E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[DD18] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

[DHL19] Y. Dong, S. B. Hopkins, and J. Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *CoRR*, abs/1906.11366, 2019. Conference version in NeurIPS 2019.

[DK19] I. Diakonikolas and D. M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019.

[DKK+16] 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)*, pages 655–664, 2016.

[DKK+17] 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)*, pages 999–1008, 2017.

[DKK+19a] 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)*, pages 1596–1606, 2019.

[DKK+19b] I. Diakonikolas, S. Karmalkar, D. Kane, E. Price, and A. Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. In *Advances in Neural Information Processing Systems 33, NeurIPS 2019*, pages 10688–10699, 2019.

[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)*, pages 73–84, 2017.

[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)*, pages 2745–2754, 2019.

[DL19] J. Depersin and G. Lecue. Robust subgaussian estimation of a mean vector in nearly linear time. *CoRR*, abs/1906.03058, 2019.

[Dru17] D. Drusvyatskiy. The proximal point method revisited. *arXiv preprint arXiv:1712.06038*, 2017.

[GHJY15] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[HSK17] H. Hassani, M. Soltanolkotabi, and A. Karbasi. Gradient methods for submodular maximization. In *Advances in Neural Information Processing Systems*, pages 5841–5851, 2017.

[Hub64] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[JK17] P. Jain and P. Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

[JNJ19] C. Jin, P. Netrapalli, and M. I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*, 2019.

[KKM18] A. Klivans, P. Kothari, and R. Meka. Efficient algorithms for outlier-robust regression. In *Proc. 31st Annual Conference on Learning Theory (COLT)*, pages 1420–1430, 2018.

[KMO10] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[LAT+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. *Science*, 319:1100–1104, 2008.

[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)*, pages 665–674, 2016.

[LW11] P. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

[PLJD10] P. Paschou, J. Lewis, A. Javed, and P. Drineas. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of Medical Genetics*, 47:835–847, 2010.

[PSBR18] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[PTZ16] R. Peng, K. Tangwongsan, and P. Zhang. Faster and simpler width-independent parallel algorithms for positive semidefinite programming. *arXiv preprint arXiv:1201.5135v3*, 2016.

[Roc70] R. T. Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.

[Roc81] R. T. Rockafellar. Favorable classes of lipschitz continuous functions in subgradient optimization. 1981.

[Roc15] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 2015.

[RPW+02] N. Rosenberg, J. Pritchard, J. Weber, H. Cann, K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. Genetic structure of human populations. *Science*, 298:2381–2385, 2002.

[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)*, pages 45:1–45:21, 2018.

[SKL17] J. Steinhardt, P. W. Koh, and P. S. Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems 30*, pages 3520–3532, 2017.

[TBS+15] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

[Tuk60] J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

[Wil67] R. M. Wilcox. Exponential operators and parameter differentiation in quantum physics. *Journal of Mathematical Physics*, 8(4):962–982, 1967.

A Setting Constants in Section 3

In this section, we describe how to appropriately set the universal constants $c_1, \ldots, c_4 \geq 1$ in Section 3. These constants are set in the following order: c_1, c_3, c_4, c_2. In this order, each c_i only depends on the constants set before it, and there is only a lower bound requirement on the value of each c_i so we can set c_i to a sufficiently large constant.

The constant c_1 appears in Condition (3). and is related to the constants involved in the concentration inequalities required to establish this condition. With the right sample complexity, Condition (3) holds with high probability for $\delta = c_1\epsilon\ln(1/\epsilon)$.

For the remaining three constants, recall that by assumption $r = \|\mu_w - \mu^\star\|_2 \geq c_2\epsilon\sqrt{\ln(1/\epsilon)} \geq \epsilon\sqrt{\ln(1/\epsilon)}$.

Next we choose c_3 such that $c_3 \geq 5c_1$. This is to guarantee that, in the proof of Lemma 3.5, we have $2c_1\ln(1/\epsilon) + 2\sqrt{2c_1\ln(1/\epsilon)}\cdot r \leq c_3 \cdot \frac{r^2}{\epsilon^2}$.

The constant c_4 appears in the proof of Lemma 3.4. There are two inequalities related to c_4. We need $c_4 \geq 50c_1$ so that $\frac{c_4 r^2}{\epsilon} - c_1\epsilon\ln(1/\epsilon) \geq 0.98 \cdot c_4 \cdot \frac{r^2}{\epsilon}$, and we require $c_4 \geq \max(100, 6c_3)$ so that $\frac{0.49\cdot c_4 \cdot r^2}{\epsilon^2} - \frac{1.4\cdot\sqrt{c_4}\cdot r^2}{\epsilon} > 2c_3 \cdot \frac{r^2}{\epsilon^2}$.

Finally, we set the value of c_2, which appears in our final guarantee: we show that any stationary point w of $f(w)$ satisfies $\|\mu_w - \mu^\star\|_2 \leq c_2\epsilon\sqrt{\ln(1/\epsilon)}$. The constant c_2 only depends on c_4. At the beginning of the proof of Lemma 3.4, we need that if $\|\mu_w - \mu^\star\|_2 \geq c_2\epsilon\sqrt{\ln(1/\epsilon)}$, then $\|\Sigma_w\|_2 \geq 1 + c_4 \cdot \frac{r^2}{\epsilon}$. By Lemma 2.1 from [DKK+16], we know that this is possible if we set c_2 to be sufficiently large.

B Missing Proofs from Section 4

In this section, we prove Theorem 3.2 and Lemma 4.2 from Section 4. These two statements play an important role in showing that projected sub-gradient descent efficiently finds an approximate stationary point w, and that w is a good solution to our robust mean estimation task.

We briefly recall our notation. We use $X \in \mathbb{R}^{d\times N}$ to denote the sample matrix, $\Sigma_w = \big(X\,\mathrm{diag}(w)X^\top - Xww^\top X^\top\big)$, $F(w,u) = u^\top \Sigma_w u$, $f(w) = \max_u F(w,u) = \|\Sigma_w\|_2$, and $\Delta_{N,\epsilon} = \Big\{w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N} \forall i\Big\}$.

Note that we can assume without loss of generality that no input samples have very large ℓ_2-norm. This is because we can perform a standard preprocessing step that centers the input samples at the coordinate-wise median, which does not affect our mean estimation task. We can then throw away all samples that are $\Omega(\sqrt{d\log d})$ far from the coordinate-wise median. With high probability, the coordinate-wise median of all good samples are $O(\sqrt{d\log d})$ far from the true mean. Assuming this happens, then no good samples are thrown away and the remaining samples satisfy $\max_i \|X_i\|_2 = O(\sqrt{d\log d})$. Consequently, we have $\|\mu_w\|_2 = O(\sqrt{d\log d})$ for any $w \in \Delta_{N,\epsilon}$.

In Lemma B.1, we show that the function $F(w,u) = u^\top \Sigma_w u$ is Lipschitz and smooth with respect to w.

Lemma B.1. *The function $F(w,u)$ is L-Lipschitz and β-smooth for $L = \widetilde{O}(\sqrt{N}d)$ and $\beta = \widetilde{O}(Nd)$. That is,*

$$|F(w,u) - F(\widetilde{w},u)| \le L\,\|\widetilde{w} - w\|_2 \quad \text{for all } w,\widetilde{w},\in \Delta_{N,2\epsilon} \text{ and all unit vectors } u \in \mathbb{R}^d$$
$$\|\nabla_w F(w,u) - \nabla_w F(\widetilde{w},u)\|_2 \le \beta\,\|\widetilde{w} - w\|_2 \quad \text{for all } w,\widetilde{w},\in \Delta_{N,2\epsilon} \text{ and all unit vectors } u \in \mathbb{R}^d\,.$$

Proof. We use the ℓ_2-norm of the gradient to bound L from above. We have

$$\begin{aligned}\|\nabla_w F(w,u)\|_2 &= \Big\|X^\top u \odot X^\top u - 2(u^\top Xw)X^\top u\Big\|_2 \\ &\le \sqrt{N}\max_i (X_i^\top u)^2 + 2\Big\|u^\top X\Big\|_\infty \|w\|_1\,\|X\|_2\,\|u\|_2 \\ &\le \sqrt{N}\max_i \|X_i\|_2^2 + 2\max_i \|X_i\|_2\,\|X\|_2\ .\end{aligned}$$

To bound from above the smoothness parameter, we have

$$\|\nabla_w F(w,u) - \nabla_w F(\widetilde{w},u)\|_2 = 2\Big|u^\top X(w-\widetilde{w})\Big|\,\Big\|X^\top u\Big\|_2 \le 2\,\|X\|_2^2\,\|w-\widetilde{w}\|_2\ .$$

We conclude the proof by observing that, after the preprocessing step, we have $\max_i \|X_i\|_2 = O(\sqrt{d\log d})$ and consequently $\|X\|_2 = O(\sqrt{Nd\log d})$. Therefore, $L = O(\sqrt{N}d\log d)$ and $\beta = O(Nd\log d)$. □

Recall that the Moreau envelope $f_\beta(w)$ is defined as

$$f_\beta(w) = \min_{\widetilde{w}} \mathcal{I}_{\mathcal{K}}(\widetilde{w}) + F(\widetilde{w}) + \beta\,\|\widetilde{w} - w\|_2^2 = \min_{\widetilde{w}\in\mathcal{K}} f(\widetilde{w}) + \beta\,\|\widetilde{w} - w\|_2^2\ ,$$

where $\mathcal{I}_{\mathcal{K}}(\cdot)$ is the support function of $\mathcal{K}$.

We restate Theorem 3.2 before proving it.

Theorem 3.2. *Consider the spectral norm loss $f(w) = \|\Sigma_w\|_2$ with $f_\beta(w)$ denoting the corresponding Moreau envelope function per Definition 2.4 with $\beta = 2\|X\|_2^2$. Then, for any $w \in \Delta_{N,2\epsilon}$ obeying*

$$\|\nabla f_\beta(w)\|_2 = O(\log(1/\epsilon)),$$

we have $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.

Proof. Let $\delta = \frac{c_3 c_2^2 \ln(1/\epsilon)}{\sqrt{2}}$, where c_2 and c_3 are the positive universal constants from Lemma 3.4. We show that any $w \in \Delta_{N,2\epsilon}$ obeying $\|\nabla f_\beta(w)\|_2 \le \delta$ must satisfy that $\|\mu_w - \mu^\star\|_2 \le O(\epsilon\sqrt{\log(1/\epsilon)})$.

The condition $\|\nabla f_\beta(w)\|_2 \le \delta$ implies that there exists a vector $\widehat{w}$ such that (see, e.g., [Roc15]):

$$\|\widehat{w} - w\|_2 = \frac{\delta}{2\beta} \quad \text{and} \quad \min_{g \in \partial f(\widehat{w}) + \partial \mathcal{I}_{\mathcal{K}}(\widehat{w})} \|g\|_2 \le \delta \ .$$

We first show that $\widehat{w}$ is a good solution.

It is well known that the subdifferential of the support function is the normal cone, which is in turn the polar of the tangent cone. That is,

$$\partial \mathcal{I}_{\mathcal{K}}(\widehat{w}) = \mathcal{N}_{\mathcal{K}}(\widehat{w}) = (\mathcal{C}_{\mathcal{K}}(\widehat{w}))^\circ \ .$$

Thus, there exists a vector $g = \nu + v$ with $\|g\|_2 \le \delta$ such that $\nu \in \partial f(\widehat{w})$ and $v \in (\mathcal{C}_{\mathcal{K}}(\widehat{w}))^\circ$. Now consider any unit vector $u \in \mathcal{C}_{\mathcal{K}}(\widehat{w})$:

$$-\delta \le u^\top g = u^\top \nu + u^\top v \le u^\top \nu \ ,$$

where the last step follows from the definition of the polar set. In other words, there exists a vector $\nu \in \partial f(\widehat{w})$ such that

$$-\nu^\top u \le \delta \quad \text{for all unit vectors } \ u \in \mathcal{C}_{\mathcal{K}}(\widehat{w}) \ . \tag{5}$$

Suppose $\|\mu_{\widehat{w}} - \mu^\star\|_2 \ge c_2\epsilon\sqrt{\ln(1/\epsilon)}$. Then for the $v \in \partial f(\widehat{w})$ in question, we can use Lemmas 3.4 and 3.5 to find two coordinates i and j such that

$$\widehat{w}_i > 0, \ \widehat{w}_j < \frac{1}{(1-2\epsilon)N}, \ \text{ and } \ \nu_i - \nu_j > c_3 \frac{\|\mu_{\widehat{w}} - \mu^\star\|_2^2}{\epsilon^2} \ge c_3 c_2^2 \ln(1/\epsilon) = \sqrt{2}\delta \ .$$

However, this contradicts Condition (5), because for the unit vector $u = \frac{1}{\sqrt{2}}(e_j - e_i)$, where e_i is the i-th basis vector, we have $u \in \mathcal{C}_{\Delta_{N,2\epsilon}}(\widehat{w})$ but

$$-\nu^\top u = \frac{\nu_i - \nu_j}{\sqrt{2}} > \delta \ .$$

Therefore, $\widehat{w}$ must satisfy $\|\mu_{\widehat{w}} - \mu^\star\|_2 < c_2\epsilon\sqrt{\ln(1/\epsilon)}$.

We conclude the proof by noticing that w is very close to $\widehat{w}$, so if $\widehat{w}$ is a good solution, then w must also be a good solution:

$$\begin{aligned}
\|\mu_w - \mu^\star\|_2 &\le \|\mu_w - \mu_{\widehat{w}}\|_2 + \|\mu_{\widehat{w}} - \mu^\star\|_2 \\
&\le \|X\|_2 \|w - \widehat{w}\|_2 + c_2\epsilon\sqrt{\ln(1/\epsilon)} \\
&= O(\beta^{-1/2}\delta + \epsilon\sqrt{\log(1/\epsilon)}) = O(\epsilon\sqrt{\log(1/\epsilon)}) \ .
\end{aligned}$$

In the last two steps, we used the fact that $\|\widehat{w} - w\|_2 = \frac{\delta}{2\beta}$ and $\beta = 2\|X\|_2^2$ (see Lemma B.1). This completes the proof of Theorem 3.2. □

We restate Lemma 4.2 before proving it. We note that the proof of Lemma 4.2 is directly inspired by the proof of Theorem 2.1 in [DD18].

Lemma 4.2. *Let $\mathcal{K}$ be a closed convex set. Let $F(w,u)$ be a function which is L-Lipschitz and β-smooth with respect to w. Consider the following optimization problem $\min_{w\in\mathcal{K}} \max_{\|u\|_2=1} F(w,u)$.*

Starting from any initial point $w_0 \in \mathcal{K}$, we run iterative updates of the form:

$$\begin{aligned}&\text{Find } u_\tau \text{ with } F(w_\tau, u_\tau) \geq (1-\epsilon') \max_u F(w_\tau, u_\tau);\\ &w_{\tau+1} = \mathcal{P}_{\mathcal{K}}(w_\tau - \eta \nabla_w F(w_\tau, u_\tau)\ ,\end{aligned}$$

for T iterations with step size $\eta = \frac{\gamma}{\sqrt{T}}$. Then, we have

$$\begin{aligned}&\min_{0\leq\tau<T} \|\nabla f_\beta(w_\tau)\|_2^2\\ &\quad\leq \frac{2}{\sqrt{T}}\left(\frac{f_\beta(w_0) - \min_w f(w)}{\gamma} + \gamma\beta L^2\right) + 4\beta\epsilon'\ ,\end{aligned}$$

where $f_\beta(w)$ is the Moreau envelope, as in Definition 2.4.

Proof. Note that since f is β-smooth with respect to w and u_τ is an approximate maximizer for w_τ, for any $\widetilde{w} \in \mathcal{K}$, we have that

$$\begin{aligned}f(\widetilde{w}) \geq F(\widetilde{w}, u_\tau) &\geq F(w_\tau, u_\tau) + (\nabla_w F(w_\tau, u_\tau))^\top (\widetilde{w} - w_\tau) - \frac{\beta}{2}\|\widetilde{w} - w_\tau\|_2^2\\ &\geq f(w_\tau) - \epsilon' + (\nabla_w F(w_\tau, u_\tau))^\top (\widetilde{w} - w_\tau) - \frac{\beta}{2}\|\widetilde{w} - w_\tau\|_2^2\ . \qquad (6)\end{aligned}$$

To continue, define the proximal function

$$\text{prox}_{f_\beta}(w) = \arg\min_{\widetilde{w}\in\mathcal{K}} \left(f(\widetilde{w}) + \beta\|\widetilde{w} - w\|_2\right)\ ,$$

and let $\widehat{w}_\tau = \text{prox}_{f_\beta}(w_\tau)$.

Now we have

$$\begin{aligned}f_\beta(w_{\tau+1}) &\leq f(\widehat{w}_\tau) + \beta\|\widehat{w}_\tau - w_{\tau+1}\|_2\\ &= f(\widehat{w}_\tau) + \beta\|\widehat{w}_\tau - \Pi_{\mathcal{K}}(w_\tau - \eta\nabla_w F(w_\tau, u_\tau))\|_2\\ &\leq f(\widehat{w}_\tau) + \beta\|\widehat{w}_\tau - w_\tau + \eta\nabla_w F(w_\tau, u_\tau)\|_2 && \text{(convexity of } \mathcal{K})\\ &= f(\widehat{w}_\tau) + \beta\|\widehat{w}_\tau - w_\tau\|_2^2 + 2\eta\beta(\nabla_w F(w_\tau, u_\tau))^\top(\widehat{w}_\tau - w_\tau) + \eta^2\beta\|\nabla_w F(w_\tau, u_\tau)\|_2^2\\ &= f_\beta(w_\tau) + 2\eta\beta(\nabla_w F(w_\tau, u_\tau))^\top(\widehat{w}_\tau - w_\tau) + \eta^2\beta\|\nabla_w F(w_\tau, u_\tau)\|_2^2 && (\widehat{w}_\tau = \text{prox}_{f_\beta}(w_\tau))\\ &\leq f_\beta(w_\tau) + 2\eta\beta(\nabla_w F(w_\tau, u_\tau))^\top(\widehat{w}_\tau - w_\tau) + \eta^2\beta L^2 && (F(w,u) \text{ is } L\text{-Lipschitz in } w)\\ &\leq f_\beta(w_\tau) + 2\eta\beta\left(f(\widehat{w}_\tau) - f(w_\tau) + \epsilon' + \frac{\beta}{2}\|\widehat{w}_\tau - w_\tau\|_2^2\right) + \eta^2\beta L^2\ . && \text{(by Inequality (6))}\end{aligned}$$

Summing the above over τ, we obtain

$$f_\beta(w_T) \leq f_\beta(w_0) + 2\eta\beta\sum_{\tau=0}^{T-1}\left(f(\widehat{w}_\tau) - f(w_\tau) + \frac{\beta}{2}\|\widehat{w}_\tau - w_\tau\|_2^2\right) + \eta^2\beta L^2 T + 2\eta\beta T\epsilon'\ .$$

Dividing by $2\eta\beta T$, we get

$$\frac{1}{T}\sum_{\tau=0}^{T-1}\left(f(w_\tau)-f(\widehat{w}_\tau)-\frac{\beta}{2}\left\|\widehat{w}_\tau-w_\tau\right\|_2^2\right)\leq\frac{f_\beta(w_0)-f_\beta(w_T)}{2\eta\beta T}+\frac{\eta L^2}{2}+\epsilon'$$
$$\leq\frac{f_\beta(w_0)-\min_w f(w)}{2\eta\beta T}+\frac{\eta L^2}{2}+\epsilon'\ .$$

Observe that the function $w\to f(w)+\beta\left\|w-w_\tau\right\|_2^2$ is β-strongly convex, therefore

$$\begin{aligned}
&f(w_\tau)-f(\widehat{w}_\tau)-\frac{\beta}{2}\left\|\widehat{w}_\tau-w_\tau\right\|_2^2\\
&=\left(f(w_\tau)+\beta\left\|w_\tau-w_\tau\right\|_2^2\right)-\left(f(\widehat{w}_\tau)+\beta\left\|w_\tau-\widehat{w}_\tau\right\|_2^2\right)+\frac{\beta}{2}\left\|w_\tau-\widehat{w}_\tau\right\|_2^2\\
&\geq\frac{\beta}{2}\left\|\widehat{w}_\tau-w_\tau\right\|_2^2+\frac{\beta}{2}\left\|\widehat{w}_\tau-w_\tau\right\|_2^2 && \text{(strong convexity)}\\
&=\beta\left\|\widehat{w}_\tau-w_\tau\right\|_2^2=\frac{1}{4\beta}\left\|\nabla f_\beta(w_\tau)\right\|_2^2\ .
\end{aligned}$$

In the above, we used the fact that for a β-strongly convex function $h(w)=\mathcal{I}_{\mathcal{K}}(w)+f(w)+\beta\left\|w-w_\tau\right\|_2^2$, we have $g(w_\tau)-g(\widehat{w}_\tau)\geq\frac{\beta}{2}\left\|w_\tau-\widehat{w}_\tau\right\|_2^2$.

Combining the two inequalities above, we arrive at

$$\frac{1}{T}\sum_{\tau=0}^{T-1}\left\|\nabla f_\beta(w_\tau)\right\|_2^2\leq 2\frac{f_\beta(w_0)-\min_w f(w)}{\eta T}+2\eta\beta L^2+4\beta\epsilon'\ .$$

Finally, setting the step size $\eta=\frac{\gamma}{\sqrt{T}}$, we conclude that

$$\min_{0\leq\tau<T}\left\|\nabla f_\beta(w_\tau)\right\|_2^2\leq\frac{2}{\sqrt{T}}\left(\frac{f_\beta(w_0)-\min_w f(w)}{\gamma}+\gamma\beta L^2\right)+4\beta\epsilon'\ .$$

This completes the proof of Lemma 4.2. □

C Minimizing Softmax of Spectral Norm

In this section, we analyze our alternate non-convex formulation that replaces the spectral norm with a softmax. Note that when the largest eigenvalue of Σ_w is not unique, the spectral norm of Σ_w may not be differentiable with respect to w. Instead of considering sub-gradients, we can minimize the softmax of the eigenvalues of Σ_w, which is a smoothed version of spectral norm that is differentiable everywhere.

Formally, we minimize the following non-convex objective function:

$$f(w)=\operatorname{smax}_\rho\left(\Sigma_w\right)=\frac{1}{\rho}\ln\operatorname{tr}(\exp(\rho\Sigma_w))\qquad\text{for}\qquad\rho=\frac{\ln d}{\epsilon}\ ,\tag{7}$$

where $X\in\mathbb{R}^{d\times N}$ is the sample matrix, and $\Sigma_w=X\operatorname{diag}(w)X^\top-Xww^\top X^\top$ is the weighted empirical covariance matrix.

The structure of this section is as follows: In Section C.1, we start by recording some useful properties of the softmax objective. In Section C.2, we prove our key structural result for this section (Theorem C.5), establishing that any approximate stationary point w of $f(w)$ provides a good estimate μ_w of the true mean $\mu^\star$. In Section C.3, we present our algorithmic result (Theorem 1.4), which states that we can efficiently find an approximate stationary point of $f(w)$ via projected gradient descent.

C.1 Basic Properties of Softmax

Lemma C.1 (Duality of softmax)**.** *For any $Z \in \mathbb{R}^{n\times n}$ and $\rho > 0$, let $\mathrm{smax}_\rho(Z) := \frac{1}{\rho}\ln \mathrm{tr}(\exp(\rho Z))$. We have the following identity*

$$\mathrm{smax}_\rho(Z) = \max_{Y\in\Delta_{n\times n}} \left(Y \bullet Z - \frac{1}{\rho} Y \bullet \log Y \right) .$$

Proof. Fix $Z \in \mathbb{R}^{n\times n}$. Let $f(Y) = Y \bullet Z - \frac{1}{\rho} Y \bullet \log Y$. Using the KKT conditions, we know that when $f(Y)$ is maximized, we have $\frac{\partial f}{\partial Y} = \lambda I$, for some $\lambda \in \mathbb{R}$. Combining this with $\frac{\partial f}{\partial Y} = Z - \frac{1}{\rho}(\log Y + I)$, it follows that $f(Y)$ is maximized at

$$Y^\star = \exp(\rho Z - (\rho\lambda + 1)I) = \frac{\exp(\rho Z)}{\mathrm{tr}(\exp(\rho Z))} ,$$

where the second equality holds because $Y^\star \in \Delta_{n\times n}$. One can substitute $Y^\star$ into the definition of $f(Y)$ and verify that $f(Y^\star) = \mathrm{smax}_\rho(Z)$. □

Corollary C.2 (Softmax and max)**.** *For any PSD matrix $Z \in \mathbb{R}^{n\times n}$ and $\rho > 0$, we have that $\lambda_{\max}(Z) \le \mathrm{smax}_\rho(Z) \le \lambda_{\max}(Z) + \frac{\ln n}{\rho}$. Moreover, for $Y = \frac{\exp(\rho Z)}{\mathrm{tr}(\exp(\rho Z))}$, we have that $Y \bullet Z \ge \mathrm{smax}_\rho(Z) - \frac{\ln n}{\rho}$.*

Proof. Observe that

$$\mathrm{smax}_\rho(Z) = \frac{1}{\rho}\ln \mathrm{tr}(\exp(\rho Z)) \ge \frac{1}{\rho}\ln \lambda_{\max}(\exp(\rho Z)) = \lambda_{\max}(Z) ,$$

and

$$\mathrm{smax}_\rho(Z) = \frac{1}{\rho}\ln \mathrm{tr}(\exp(\rho Z)) \le \frac{1}{\rho}\ln(n \cdot \lambda_{\max}(\exp(\rho Z))) = \lambda_{\max}(Z) + \frac{\ln n}{\rho} .$$

For the second claim, by Lemma C.1, we know that $\mathrm{smax}_\rho(Z) = Y \bullet Z - \frac{1}{\rho} Y \bullet \log Y$. The claim then follows from the fact that $Y \bullet \log Y \ge -\ln n$ for all $Y \in \Delta_{n\times n}$. □

When working with the matrix exponentials in our softmax objective function f, the following chain rule formula will be useful to compute the Hessian of f (see, e.g., [Wil67]).

Lemma C.3 (Derivative of matrix exponential)**.** *For a symmetric matrix function $X(t)$ that depends on a scalar t, we have that*

$$\frac{d}{dt}\exp(X(t)) = \int_0^1 \exp(\alpha X(t)) \frac{dX(t)}{dt} \exp((1-\alpha)X(t)) d\alpha .$$

C.2 Structural Result: Any Approximate Stationary Point Suffices

The gradient of our softmax objective function is

$$\nabla f(w) = \mathrm{diag}(X^\top Y X) - 2X^\top Y X w \;, \quad \text{where} \quad Y = \frac{\exp(\rho \Sigma_w)}{\mathrm{tr}(\exp(\rho \Sigma_w))} \;. \tag{8}$$

Notice that $Y \in \Delta_{N\times N}$ is a convex combination of directions. That is, we can write $Y = \sum_{k=1}^{d} \lambda_k u_k u_k^\top$, where $u_k \in \mathbb{R}^d$ and $\sum_k \lambda_k = 1$. The gradient $\nabla f(w)$ is the same as the gradient of w for the one-dimensional problem, where the input samples are $(X_i^\top Y^{1/2})_{i=1}^N$. Equivalently, $\nabla f(w)$ tries to move w towards minimizing the average variance

$$\sum_k \lambda_k \left(\sum_i w_i (X_i^\top u_k)^2 - \left(\sum_i w_i (X_i^\top u_k) \right)^2 \right)$$

of the projections of X along the directions $\{u_k\}$.

The intuition is as follows: The goal is to show that $\lambda_{\max}(\Sigma_w)$ is small at any stationary point w of $\mathrm{smax}_\rho(\Sigma_w)$. Now fix some $w \in \Delta_{N,2\epsilon}$, where $\lambda_{\max}(\Sigma_w)$ is large. Then $\mathrm{smax}_\rho(\Sigma_w)$ must be large. By the duality of softmax, there is a combination of directions Y such that: (1) the one-dimensional samples $(X_i^\top Y^{1/2})_{i=1}^N$ weighted by w have large variance, and (2) the derivative of $\mathrm{smax}_\rho(\Sigma_w)$ is the same as the derivative for minimizing variance on this one-dimensional instance. We proceed by examining this one-dimensional instance, which is easier to analyze. We show that w cannot be a stationary point, because we can always reduce the variance by increasing the weight on one of the good samples and reducing the weight on one of the bad samples.

Formally, we use the following notion of approximate stationarity for our constrained non-convex minimization problem.

Definition C.4. *Fix a convex set $\mathcal{K}$. For $\delta > 0$, we say $x \in \mathcal{K}$ is a δ-stationary point of f if the following condition holds: For any unit vector u where $x + \alpha u \in \mathcal{K}$ for some $\alpha > 0$, we have $u^\top \nabla f(x) \geq -\delta$.*

Our main structural result in this section is the following theorem.

Theorem C.5 (Any stationary point of $f(w)$ is a good solution)**.** *Let S be an ϵ-corrupted set of $N = \widetilde{\Omega}(d/\epsilon^2)$ samples drawn from a d-dimensional Gaussian $\mathcal{N}(\mu^\star, I)$ with unknown mean $\mu^\star$. Suppose S satisfies Condition* (3) *and Lemma 2.1.*

Let $f(w)$ be the softmax objective as defined in Equation (7)*. Let $\delta = c\ln(1/\epsilon)$ for some universal constant c. For any $w \in \Delta_{N,2\epsilon}$ that is a δ-stationary point of $f(w)$, we have $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.*

Theorem C.5 follows directly from Lemmas C.6, C.7, and C.8.

For the rest of this subsection, we assume the input samples satisfy Condition (3) and Lemma 2.1, and we fix an approximate stationary point $w \in \Delta_{N,2\epsilon}$ of the softmax objective. We establish the following bimodal sub-gradient property which holds at all (approximate) stationary points.

Lemma C.6 (Bimodal sub-gradient property at stationary points)**.** *Fix $w \in \Delta_{N,2\epsilon}$. Let $S_- = \{i : w_i > 0\}$ and $S_+ = \{i : w_i < \frac{1}{(1-2\epsilon)N}\}$ denote the set of coordinates of w that can decrease and increase respectively. If w is a δ-stationary point of $f(w)$, then $\nabla f(w)_i \leq \nabla f(w)_j + \sqrt{2}\delta$ for all $i \in S_-$ and $j \in S_+$.*

Proof. Suppose there is some $i \in S_-$ and $j \in S_+$ such that $\nabla f(w)_i > \nabla f(w)_j + \sqrt{2}\delta$.

Consider the unit vector $u = \frac{1}{\sqrt{2}}(e_j - e_i)$, where e_i is the i-th basis vector. We have $w + \alpha u \in \Delta_{N,2\epsilon}$ for $\alpha = \min(w_i, \frac{1}{(1-2\epsilon)N} - w_j) > 0$, but

$$u^\top \nabla f(x) = \frac{\nabla f(w)_j - \nabla f(w)_i}{\sqrt{2}} < -\delta \,,$$

which violates the assumption that w is a δ-approximate stationary point (Definition C.4). □

At a high level, we prove Theorem C.5 by showing that if μ_w is far from $\mu^\star$, then w violates Lemma C.6. More specifically, if μ_w is far from $\mu^\star$, then there exists a bad sample with index $j \in S_-$ whose gradient is large (Lemma C.7). Meanwhile, the concentration bound in Condition (3) guarantees that there exists a good sample with index $i \in S_+$ whose gradient is small (Lemma C.8).

We frequently use the partial derivative of $f(w)$ with respect to w_i in our analysis:

$$\begin{aligned}\nabla f(w)_i &= X_i^\top Y X_i - 2X_i^\top Y \mu_w \\ &= (X_i - \mu^\star)^\top Y (X_i - \mu^\star) - 2(X_i - \mu^\star)^\top Y (\mu_w - \mu^\star) \\ &\quad + \mu^{\star\top} Y (\mu^\star - 2\mu_w) \,.\end{aligned}$$

Notice that the last term in $\nabla f(w)_i$ is the same for all i. Since our goal is to identify $i \in S_-$ and $j \in S_+$ such that $\nabla f(w)_i > \nabla f(w)_j$, we can focus on the first two terms.

We have the following lemmas:

Lemma C.7. *Fix $w \in \Delta_{N,2\epsilon}$ and assume that Condition (3) and Lemma 2.1 hold. Let c_2 and c_3 be universal constants. Let $r = \|\mu_w - \mu^\star\|_2$ and suppose $r \geq c_2\epsilon\sqrt{\ln(1/\epsilon)}$. Then, there exists $i \in (B \cap S_-)$ such that*

$$\nabla f(w)_i - \mu^{\star\top} Y (\mu^\star - 2\mu_w) > 2c_3 \cdot \frac{r^2}{\epsilon^2} \,.$$

Lemma C.8. *Consider the same setting as in Lemma C.7. There exists $j \in (G \cap S_+)$ such that*

$$\nabla f(w)_j - \mu^{\star\top} Y (\mu^\star - 2\mu_w) \leq c_3 \cdot \frac{r^2}{\epsilon^2} \,.$$

We defer the proofs of Lemmas C.7 and C.8 to Section C.2.1, and we first use them to prove Theorem C.5.

Proof of Theorem C.5. Suppose that w is a bad solution where $\|\mu_w - \mu^\star\|_2 \geq c_2\epsilon\sqrt{\ln(1/\epsilon)}$. Since we assume Condition (3) and Lemma 2.1 both hold on the input samples, we can use Lemmas C.7 and C.8 to find two coordinates $i \in S_-$ and $j \in S_+$, such that the bimodal sub-gradient property in Lemma C.6 does not hold at w. Therefore, w is not a δ-approximate stationary point for some $\delta = \sqrt{2}c_3\frac{\|\mu_w - \mu^\star\|_2^2}{\epsilon^2} \geq \sqrt{2}c_3c_2^2\ln(1/\epsilon)$, that is, we can set $c = \sqrt{2}c_3c_2^2$. □

C.2.1 Proofs of Lemmas C.7 and C.8

In this section, we prove Lemmas C.7 and C.8.

The proofs of these lemmas are conceptually similar to the proofs of related lemmas (Lemmas 3.4 and 3.5) in Section 3. We include their proofs here to make this section self-contained.

The main difference is that we switch to the softmax objective, and consequently, we need to work with multiple directions simultaneously. That is, we consider the projections using Y instead of the projections along the maximum eigenvector of Σ_w.

Lemma C.7 states that when μ_w is far from $\mu^\star$, there exists an index $i \in (B \cap S_-)$ such that the gradient $\nabla f(w)_i$ is relatively large.

Recall that the gradient $\nabla f(w)$ in Equation (8) is the same as the gradient of the variance (weighted by w) of the one-dimensional samples $\left(X_i^\top Y^{1/2}\right)_{i=1}^N$. For this one-dimensional problem, a sample far from the (projected) true mean must have large gradient. Our objective is to find such a sample for which we can decrease its weight. More specifically, since w is assumed to be a bad solution, and the softmax objective is close to the spectral norm of Σ_w, the weighted empirical variance of the projected samples is very large. Because the good samples cannot have this much variance, most of the variance comes from the bad samples. We prove that among these bad samples that contribute a lot to the variance, one of them must be very far from the (projected) true mean and hence has a large gradient, which satisfies Lemma C.7.

We use $c_1, \ldots, c_4$ to denote universal positive constants that are independent of N, d, and ϵ. These constants can be set in a way that is similar to that in Section 3 (see Appendix A). The universal constant c in Theorem C.5 can be set as $c = \sqrt{2}c_3c_2^2$ after we set c_2 and c_3.

Proof of Lemma C.7. We first show that $\Sigma_w \bullet Y$ is relatively large. By Lemma 2.1, we know that if $\|\mu_w - \mu^\star\|_2 \geq r$ and $r \geq c_2\epsilon\sqrt{\ln(1/\epsilon)}$, then

$$\lambda_{\max}(\Sigma_w) \geq 1 + c_4 \cdot \frac{r^2}{\epsilon} \,.$$

By Corollary C.2, for $Y = \frac{\exp(\rho\Sigma_w)}{\operatorname{tr}(\exp(\rho\Sigma_w))}$ and $\rho = \frac{\ln d}{\epsilon}$, we have

$$\Sigma_w \bullet Y \geq \operatorname{smax}_\rho(\Sigma_w) - \epsilon \geq \lambda_{\max}(\Sigma_w) - \epsilon \geq 1 - \epsilon + \frac{c_4 r^2}{\epsilon} \,.$$

Recall that $\Sigma_w = \sum_{i=1}^N w_i(X_i - \mu_w)(X_i - \mu_w^\top)$. If we replace μ_w with $\mu^\star$, we have

$$\sum_{i=1}^N w_i(X_i - \mu^\star)(X_i - \mu^\star)^\top \succeq \Sigma_w \,,$$

and therefore,

$$\left(\sum_{i=1}^N w_i(X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \bullet Y \geq \Sigma_w \bullet Y \geq 1 - \epsilon + \frac{c_4 r^2}{\epsilon} \,.$$

Next we show that most of the variance is due to bad samples. By Condition (3),

$$\left(\sum_{i\in G} w_i(X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \bullet Y \leq 1 + c_1 \cdot \epsilon\ln(1/\epsilon) \,.$$

Consequently,

$$\left(\sum_{i\in B} w_i(X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \bullet Y \geq \frac{c_4 r^2}{\epsilon} - \epsilon - c_1\epsilon\ln(1/\epsilon) \geq 0.98 \cdot c_4 \cdot \frac{r^2}{\epsilon} \,.$$

The last step is because $r \geq c_2 \cdot \epsilon\sqrt{\ln(1/\epsilon)}$ and we can choose c_2 and c_4 to be sufficiently large.

At this point, we know that when $r = \|\mu_w - \mu^\star\|_2$ is large, most of the variance is due to the bad samples. However, the total weight w_B on the bad samples is at most $\epsilon N \cdot \frac{1}{(1-2\epsilon)N} \leq 2\epsilon$. Therefore, there must be some $i \in B$ with $w_i > 0$ and

$$\left((X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \bullet Y \geq \frac{0.98 \cdot c_4 \cdot r^2 \cdot \epsilon^{-1}}{w_B} \geq 0.49 \cdot c_4 \cdot \frac{r^2}{\epsilon^2} \,.$$

By definition, $i \in B \cap S_-$. It remains to show that $\nabla f(w)_i$ is large.

$$\begin{aligned}
\nabla f(w)_i - \mu^{\star\top} Y(\mu^\star - 2\mu_w) &= \left((X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \bullet Y - 2\left((X_i - \mu^\star)(\mu_w - \mu^\star)^\top\right) \bullet Y \\
&\geq \left\|Y^{1/2}(X_i - \mu^\star)\right\|_2^2 - 2\left\|Y^{1/2}(X_i - \mu^\star)\right\|_2 \cdot \left\|Y^{1/2}\right\|_2 \cdot \|\mu_w - \mu^\star\|_2 \\
&\geq \frac{0.49 \cdot c_4 \cdot r^2}{\epsilon^2} - 2 \cdot \frac{0.7 \cdot \sqrt{c_4} \cdot r}{\epsilon} \cdot 1 \cdot r \\
&> 2c_3 \cdot \frac{r^2}{\epsilon^2} \,.
\end{aligned}$$

The first inequality is because $Y \in \Delta_{d\times d}$. The last step uses the fact that c_4 can be sufficiently large. This completes the proof of Lemma C.7. □

Lemma C.8 states that there exists an index $j \in (G \cap S_+)$ such that the gradient $\nabla f(w)_j$ is relatively small. Similar to the proof of Lemma C.7, for the projected one-dimensional instance, a sample close to the (projected) true mean should have small gradient. Our goal is to find such a sample for which we can increase its weight. Recall that S^+ contains the samples whose weight can be increased. We first prove that there are at least ϵN good samples in S^+. Among these ϵN good samples, the concentration bounds imply that there must exist some X_j that is close to the (projected) true mean. The derivative $\nabla f(w)_j$ satisfies Lemma C.8.

Proof of Lemma C.8. Recall that S^+ contains every coordinate i where $w_i < \frac{1}{(1-2\epsilon)N}$. Since at most $(1-2\epsilon)N$ samples can have the maximum weight $\frac{1}{(1-2\epsilon)N}$, we know that $|S^+| \geq 2\epsilon N$. Combining this with $|G| = (1-\epsilon)N$, we know that $|G \cap S^+| \geq \epsilon N$.

Fix a subset $G^+ \subseteq (G \cap S^+)$ of size $|G^+| = \epsilon N$. We first show that, on average, samples in G^+ do not contribute much to the variance.

Let w' be the uniform weight vector on G, i.e., $w'_i = \frac{1}{(1-\epsilon)N}$ for all $i \in G$ and $w'_i = 0$ otherwise. Since $w' \in \Delta_{N,2\epsilon}$, by Condition (3), we have that

$$\left\|\sum_{i\in G} \frac{1}{|G|}(X_i - \mu^\star)(X_i - \mu^\star)^\top - I\right\|_2 \leq c_1 \cdot \epsilon \ln(1/\epsilon) \,.$$

Let w'' be the uniform weight vector on $S \setminus G^+ = (G \setminus G^+) \cup B$, i.e., $w''_i = \frac{1}{(1-\epsilon)N}$ for all $i \in ((G \setminus G^+) \cup B)$ and $w''_i = 0$ otherwise. Since $w'' \in \Delta_{N,2\epsilon}$, again by Condition (3), we have that

$$\left\|\sum_{i\in G\setminus G^+} \frac{1}{|G|}(X_i - \mu^\star)(X_i - \mu^\star)^\top - I\right\|_2 \leq c_1 \cdot \epsilon \ln(1/\epsilon) \,.$$

Combining the previous two concentration bounds, we obtain that

$$\left\|\sum_{i\in G^+}\frac{1}{|G|}(X_i-\mu^\star)(X_i-\mu^\star)^\top\right\|_2 \le \left\|\sum_{i\in G}\frac{1}{|G|}(X_i-\mu^\star)(X_i-\mu^\star)^\top - I\right\|_2 + \left\|\sum_{i\in G\setminus G^+}\frac{1}{|G|}(X_i-\mu^\star)(X_i-\mu^\star)^\top - I\right\|_2 \le 2c_1\cdot\epsilon\ln(1/\epsilon)\ .$$

As a result, because $Y\in\Delta_{d\times d}$, it follows that

$$\left(\sum_{i\in G^+}\frac{1}{|G|}(X_i-\mu^\star)(X_i-\mu^\star)^\top\right)\bullet Y \le 2c_1\cdot\epsilon\ln(1/\epsilon)\ .$$

Now we know that, on average, samples in G^+ do not contribute much to the variance. We continue to show that one of these samples satisfies the lemma.

Let $j=\arg\min_{i\in G^+}\left(Y\bullet(X_i-\mu^\star)(X_i-\mu^\star)^\top\right)$. We have that

$$\left((X_j-\mu^\star)(X_j-\mu^\star)^\top\right)\bullet Y \le \frac{|G|}{|G^+|}\cdot 2c_1\cdot\epsilon\ln(1/\epsilon) \le 2c_1\ln(1/\epsilon)\ .$$

Finally, because $(X_j-\mu^\star)^\top Y(X_j-\mu^\star)\le 2c_1\ln(1/\epsilon)$, we can bound $\nabla f(w)_j$ from above as follows:

$$\begin{aligned}\nabla f(w)_j-\mu^{\star\top}Y(\mu^\star-2\mu_w) &= \left((X_j-\mu^\star)(X_j-\mu^\star)^\top\right)\bullet Y - 2\left((X_j-\mu^\star)(\mu_w-\mu^\star)^\top\right)\bullet Y\\ &\le \left\|Y^{1/2}(X_j-\mu^\star)\right\|_2^2 + 2\left\|Y^{1/2}(X_j-\mu^\star)\right\|_2\cdot\left\|Y^{1/2}\right\|_2\cdot\|\mu_w-\mu^\star\|_2\\ &\le 2c_1\ln(1/\epsilon)+2\sqrt{2c_1\ln(1/\epsilon)}\cdot 1\cdot r\\ &\le \frac{c_3}{2}\cdot\frac{r^2}{\epsilon^2}+\frac{c_3}{2}\cdot\frac{r}{\epsilon}\cdot r \le c_3\cdot\frac{r^2}{\epsilon^2}\ .\end{aligned}$$

The last step uses that c_3 is sufficiently large, as well as the fact that $\ln(1/\epsilon)\le\frac{r^2}{\epsilon^2}$, because $r\ge c_2\epsilon\sqrt{\ln(1/\epsilon)}$. This completes the proof of Lemma C.8. □

C.3 Convergence Rate of Minimizing Softmax

In this section, we prove our algorithmic result for the softmax objective (Theorem 1.4). We show that the projected gradient descent algorithm (Algorithm 2) on f can efficiently find an approximate stationary point w, and that w is a good solution to our robust mean estimation task.

We first restate Theorem 1.4 (correctness and iteration count of Algorithm 2).

Theorem 1.4. *Let S be an ϵ-corrupted set of $N=\widetilde{\Omega}(d/\epsilon^2)$ samples drawn from a d-dimensional Gaussian $\mathcal{N}(\mu^\star,I)$ with unknown mean $\mu^\star$. Suppose S satisfies Condition (3) and Lemma 2.1.*

Let $f(w)$ be the softmax objective as defined in Equation (7). After $\widetilde{O}(Nd^3/\epsilon)$ iterations, projected gradient descent on $f(w)$ outputs a point w such that $\|\mu_w-\mu^\star\|_2=O(\epsilon\sqrt{\log(1/\epsilon)})$.

Theorem 1.4 follows immediately from Lemmas C.9, C.10, and C.11.

Algorithm 2 Robust Mean Estimation via Projected Gradient Descent on the Softmax Objective

Input: ϵ-corrupted set of N samples $\{X_i\}_{i=1}^N$ on $\mathbb{R}^d$ satisfying Condition (3), and $\epsilon < \epsilon_0$.
Output: $w \in \mathbb{R}^N$ with $\|\mu_w - \mu^\star\|_2 \le O(\epsilon\sqrt{\log(1/\epsilon)})$.
Let $\rho = \ln d/\epsilon$.
Let $\beta = \widetilde{O}(Nd^2/\epsilon)$ be the smoothness parameter of the softmax objective $f(w) = \mathrm{smax}_\rho(\Sigma_w)$.
Let w_0 be an arbitrary weight vector in $\Delta_{N,2\epsilon}$.
Let $T = \widetilde{O}(Nd^3/\epsilon)$ and $\eta = 1/\beta$.
for $\tau = 0$ **to** $T-1$ **do**
  $w_{\tau+1} = \mathcal{P}_{\Delta_{N,2\epsilon}}(w_\tau - \eta\nabla f(w))$, where $\mathcal{P}_{\mathcal{K}}(\cdot)$ is the ℓ_2-projection operator onto $\mathcal{K}$.
end for
return $w_{\tau^\star}$ where $\tau^\star = \arg\min_{0\le\tau<T} \|w_{\tau+1} - w_\tau\|_2$.

Lemma C.9 analyzes the convergence rate of (nonconvex) projected gradient descent. The number of iterations in Lemma C.9 depends on the range and smoothness of the objective function. Lemmas C.10 and C.11 upper bounds these two parameters for our softmax objective.

We note that Lemma C.9 appears to be folklore in the optimization literature, see, e.g., [Bec17]. For the sake of completeness, we provide a self-contained proof in the following subsection.

Lemma C.9. *Fix a (possibly non-convex) function f and a convex set $\mathcal{K}$. Suppose f is β-smooth on $\mathcal{K}$ and $0 \le f(x) \le B$ for all $x \in \mathcal{K}$. If we run projected gradient descent with step size $\eta = \frac{1}{\beta}$ starting from an arbitrary $x_0 \in \mathcal{K}$:*

$$x_{\tau+1} = \Pi_{\mathcal{K}}\left(x_\tau - \eta\nabla f(x_\tau)\right) ,$$

where $\Pi_{\mathcal{K}}$ is the projection onto $\mathcal{K}$, we can compute a δ-stationary point of f in $O(\frac{\beta\cdot B}{\delta^2})$ iterations.

Recall that the softmax objective is $f(w) = \mathrm{smax}_\rho(\Sigma_w) = \frac{1}{\rho}\ln\mathrm{tr}(\exp(\rho\Sigma_w))$ with $\rho = \frac{\ln d}{\epsilon}$. A differentiable function f is β-smooth on $\mathcal{K}$ if $\|\nabla f(x) - \nabla f(y)\|_2 \le \beta\|x - y\|_2$ for all $x, y \in \mathcal{K}$.

Lemma C.10 (Smoothness of f)**.** *The softmax objective f is β-smooth on $\Delta_{N,2\epsilon}$ for $\beta = \widetilde{O}(Nd^2/\epsilon)$.*

Lemma C.11 (Range of f)**.** *The softmax objective f satisfies that $0 \le f(w) \le \widetilde{O}(d)$ for all $w \in \Delta_{N,2\epsilon}$.*

We defer the proofs of Lemmas C.9, C.10, and C.11 to the next subsections and first use them to prove Theorem 1.4.

Proof of Theorem 1.4. We first prove the correctness of Algorithm 2. Let c be the universal constant in Theorem C.5 and let $\delta = c\ln(1/\epsilon)$. We run Algorithm 2 to obtain a δ-stationary point w. Since we assume the input samples satisfy Condition (3) and Lemma 2.1, Theorem C.5 states that w is a good solution with $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\ln(1/\epsilon)})$.

We now analyze the number of iterations T. By Lemma C.9, it is sufficient to set $T = O(\frac{\beta\cdot B}{\delta^2})$, as in Algorithm 2. Substituting the upper bounds on β and B from Lemmas C.10 and C.11, and our choice of δ, we get

$$T = O(\beta\cdot B\cdot\delta^{-2}) = \widetilde{O}(Nd^2/\epsilon)\cdot\widetilde{O}(d)\cdot O(\log^{-2}(1/\epsilon)) = \widetilde{O}(Nd^3/\epsilon) ,$$

as claimed. □

C.4 Proof of Lemma C.9

In this section, we prove Lemma C.9.

Lemma C.9 analyzes the convergence rate of projected gradient descent, when we use it to minimize a smooth non-convex function with constraints. Lemma C.9 follows directly from Lemmas C.12 and C.13.

Lemma C.12 defines a "truncated gradient" mapping g and relates the progress in the τ-th iteration with $\|g(x_\tau)\|_2^2$. Because we cannot keep decreasing $f(x)$, we know that after many iterations, there exists some τ such that $\|g(x_\tau)\|_2$ is very small. Lemma C.13 shows that if $\|g(x_\tau)\|_2$ is very small, that is, if projected gradient descent moves very little between x_τ and $x_{\tau+1}$, then $x_{\tau+1}$ is an approximate stationary point.

Lemma C.12. *Fix a convex set $\mathcal{K}$. Suppose f is β-smooth on $\mathcal{K}$ and $0 \le f(x) \le B$ for all $x \in \mathcal{K}$. Suppose we run projected gradient descent with step size $\eta = \frac{1}{\beta}$ starting from an arbitrary $x_0 \in \mathcal{K}$, i.e.,*

$$x_{\tau+1} = \Pi_{\mathcal{K}}\left(x_\tau - \eta \nabla f(x_\tau)\right) \ ,$$

where $\Pi_{\mathcal{K}}$ is the ℓ_2-projection onto $\mathcal{K}$. Then we have that

$$\min_{0 \le \tau < T} \frac{1}{\eta} \left\|\Pi_{\mathcal{K}}\left(x_\tau - \eta \nabla f(x_\tau)\right) - x_\tau\right\|_2 \le \sqrt{\frac{2\beta B}{T}} \ .$$

Proof. Define the mapping

$$g(x) = \frac{x - \Pi_{\mathcal{K}}(x - \eta \nabla f(x))}{\eta} \ .$$

Let $y_{\tau+1} = x_\tau - \eta \nabla f(x_\tau)$. Notice that $x_{\tau+1} = \Pi_{\mathcal{K}}(y_{\tau+1}) = x_s - \eta g(x_\tau)$.

By the convexity of $\mathcal{K}$, we have

$$(x_{\tau+1} - x_\tau)^\top (x_{\tau+1} - y_{\tau+1}) \le 0 \ ,$$

which is equivalent to

$$\nabla f(x_\tau)^\top (x_{\tau+1} - x_\tau) \le g(x_\tau)^\top (x_{\tau+1} - x_\tau) \ .$$

Using the quadratic upper bound combined with the above inequality, we have

$$\begin{aligned}
f(x_{\tau+1}) &\le f(x_\tau) + \nabla f(x_\tau)^\top (x_{\tau+1} - x_\tau) + \frac{\beta}{2} \|x_{\tau+1} - x_\tau\|_2^2 \\
&\le f(x_\tau) + g(x_\tau)^\top (x_{\tau+1} - x_\tau) + \frac{\beta}{2} \|x_{\tau+1} - x_\tau\|_2^2 \\
&= f(x_\tau) - \eta \|g(x_\tau)\|_2^2 + \frac{\eta^2 \beta}{2} \|g(x_\tau)\|_2^2 \\
&= f(x_\tau) - \frac{1}{2\beta} \|g(x_\tau)\|_2^2 \ .
\end{aligned}$$

Therefore, after T iterations, we have

$$\min_{0 \le \tau < T} \|g(x_\tau)\|_2^2 \le \frac{1}{T} \sum_{\tau=0}^{T-1} \|g(x_\tau)\|_2^2 \le \frac{2\beta}{T} \left(f(x_0) - f(x_T)\right) \le \frac{2\beta B}{T} \ . \qquad \square$$

Lemma C.13. *Consider the same setting as in Lemma C.12. Define the tangent cone of $\mathcal{K}$ at a point $x \in \mathcal{K}$ as $\mathcal{C}_{\mathcal{K}}(x) = \mathrm{cone}(\mathcal{K} - \{x\})$. If for some τ we have*

$$\left\|\Pi_{\mathcal{K}}\left(x_\tau - \eta \nabla f(x_\tau)\right) - x_\tau\right\|_2 \le \frac{\delta}{2} ,$$

then for all unit vector $u \in \mathcal{C}_{\mathcal{K}}(x)$,

$$\nabla f(x_{\tau+1})^\top u \le \delta .$$

Proof. By the convexity of $\mathcal{K}$, we know that for any $z \in \mathcal{K}$,

$$(y_{\tau+1} - x_{\tau+1})^\top (z - x_{\tau+1}) \le 0 .$$

Consequently, for any $u \in \mathcal{C}_{\mathcal{K}}(x_{\tau+1})$, we have

$$(y_{\tau+1} - x_{\tau+1})^\top u \le 0 ,$$

which is equivalent to

$$-\nabla f(x_\tau)^\top u \le -g(x_\tau)^\top u .$$

Using the fact that u is a unit vector together with the above inequality, we get

$$\begin{aligned} -\nabla f(x_{\tau+1})^\top u &\le -\nabla f(x_{\tau+1})^\top u + \nabla f(x_\tau)^\top u - g(x_\tau)^\top u \\ &\le \left\|f(x_{\tau+1}) - \nabla f(x_\tau)\right\|_2 + \left\|g(x_\tau)\right\|_2 \\ &\le \beta \left\|x_{\tau+1} - x_\tau\right\|_2 + \left\|g(x_\tau)\right\|_2 \\ &= 2\left\|g(x_\tau)\right\|_2 \le \delta . \end{aligned}$$

□

Proof of Lemma C.9. As in Algorithm 2, we run projected gradient descent, track the value of $\|g(x_\tau)\|_2$ in each iteration, and return the x_τ that has the minimum $\|g(x_\tau)\|_2$. Combining Lemmas C.12 and C.13, if we want a δ-stationary point, we should set T such that $\sqrt{2\beta B/T} \le \delta/2$, i.e., $T \ge 8\beta B\delta^{-2} = O(\beta B \delta^{-2})$. □

C.5 Proofs of Lemmas C.10 and C.11

In this subsection, we bound from above the smoothness and maximum value of the softmax objective.

For these two lemmas, we can assume without loss of generality that no input samples have very large ℓ_2-norm. This is because we can perform a standard preprocessing step that centers the input samples at the coordinate-wise median, which does not affect our mean estimation task. We then throw away all samples that are $\Omega(\sqrt{d \log d})$ far from the coordinate-wise median. With high probability, the coordinate-wise median and all good samples are $O(\sqrt{d \log d})$ far from the true mean. Assuming this happens, then no good samples are thrown away and all remaining samples satisfies $\max_i \|X_i\|_2 = O(\sqrt{d \log d})$. Consequently, we have $\|\mu_w\|_2 = O(\sqrt{d \log d})$ for any $w \in \Delta_{N,\epsilon}$.

Proof of Lemma C.10. We proceed to bound from above the spectral norm of the Hessian of f. Recall that $X \in \mathbb{R}^{d \times N}$ and the partial derivative of f with respect to w_i is

$$\nabla f(w)_i = X_i^\top Y X_i - 2X_i^\top Y \mu_w = \left(X_i X_i^\top - X_i \mu_w^\top - \mu_w X_i^\top\right) \bullet Y,$$

where $Y = \frac{\exp(\rho\Sigma_w)}{\operatorname{tr}\exp(\rho\Sigma_w)}$ is a PSD matrix. Observe that $Y \succeq 0$, $\operatorname{tr}(Y) = 1$, and Y depends on w.

We can compute the (i,j)-th entry in the Hessian matrix of f, as follows

$$\nabla^2 f(w)_{i,j} = \frac{df(w)_i}{dw_j} = \left(X_i X_i^\top - X_i\mu_w^\top - \mu_w X_i^\top\right) \bullet \frac{dY}{dw_j} - \left(X_i X_j^\top + X_j X_i^\top\right) \bullet Y \ .$$

By the chain rule, we have

$$\begin{aligned}\frac{dY}{dw_j} &= \frac{1}{\operatorname{tr}(\exp(\rho\Sigma_w))^2}\left[\frac{d\exp(\rho\Sigma_w)}{dw_j}\operatorname{tr}(\exp(\rho\Sigma_w)) - \frac{d\operatorname{tr}(\exp(\rho\Sigma_w))}{dw_j}\exp(\rho\Sigma_w)\right] \\ &= \frac{1}{\operatorname{tr}(\exp(\rho\Sigma_w))}\left[\frac{d\exp(\rho\Sigma_w)}{dw_j} - \frac{d\operatorname{tr}(\exp(\rho\Sigma_w))}{dw_j}\cdot Y\right] \ .\end{aligned}$$

Using Lemma C.3 to compute the derivative of matrix exponential, we have

$$\begin{aligned}\frac{dY}{dw_j} &= \frac{1}{\operatorname{tr}(\exp(\rho\Sigma_w))}\left[\frac{d\exp(\rho\Sigma_w)}{dw_j} - \frac{d\operatorname{tr}(\exp(\rho\Sigma_w))}{dw_j}\ Y\right] \\ &= \frac{1}{\operatorname{tr}\exp(\rho\Sigma_w)}\left[\int_{\alpha=0}^1 \exp(\alpha\rho\Sigma_w)\frac{d(\rho\Sigma_w)}{dw_j}\exp((1-\alpha)\rho\Sigma_w)d\alpha - \left(\frac{d(\rho\Sigma_w)}{dw_j} \bullet \exp(\rho\Sigma_w)\right)Y\right] \\ &= \frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\int_{\alpha=0}^1 \exp(\alpha\Sigma_w)\frac{d\Sigma_w}{dw_j}\exp((1-\alpha)\rho\Sigma_w)d\alpha - \rho\left(\frac{d\Sigma_w}{dw_j} \bullet Y\right)Y \ .\end{aligned}$$

Since $\frac{d\Sigma_w}{dw_j} = X_j X_j^\top - X_j\mu_w^\top - \mu_w X_j^\top$, putting it all together, we have,

$$\begin{aligned}\nabla^2 f(w)_{i,j} = &-\left(X_i^\top Y(X_i - 2\mu_w)\right)\left(X_j^\top Y(X_j - 2\mu_w)\right) - 2X_i^\top Y X_j \\ &+ \frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\cdot \\ &\int_{\alpha=0}^1 \operatorname{tr}\left(\left(X_i X_i^\top - X_i\mu_w^\top - \mu_w X_i^\top\right)\exp(\alpha\rho\Sigma_w)\left(X_j X_j^\top - X_j\mu_w^\top - \mu_w X_j^\top\right)\exp((1-\alpha)\rho\Sigma_w)\right)d\alpha \ .\end{aligned}$$

Let $R = \max(\|\mu_w\|_2, \max_i \|X_i\|_2)$. From the preprocessing step, we know that $R = \widetilde{O}(d^{1/2})$. Using this fact, we obtain

$$\left|\nabla^2 f(w)_{i,j}\right| \le 9R^4 + 2R^2 + 9\rho R^4 = \widetilde{O}(\rho d^2) \ .$$

This is because the first term can be bounded from above by

$$\begin{aligned}-\left(X_i^\top Y(X_i - 2\mu_w)\right)\left(X_j^\top Y(X_j - 2\mu_w)\right) &\le \|X_i\|_2 \|Y\|_2 \|X_i - 2\mu_w\|_2 \|X_j\|_2 \|Y\|_2 \|X_j - 2\mu_w\|_2 \\ &\le 9R^4 \ .\end{aligned}$$

Similarly, the second term is at most $2R^2$. The third term can be split into 9 terms of the form

$$\begin{aligned}&\frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\int_{\alpha=0}^1 \operatorname{tr}\left(\left(X_i X_i^\top\right)\exp(\alpha\rho\Sigma_w)\left(X_j X_j^\top\right)\exp((1-\alpha)\rho\Sigma_w)\right)d\alpha \\ &= \frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\int_{\alpha=0}^1 \left(X_i^\top\exp(\alpha\rho\Sigma_w)X_j\right)\left(X_j^\top\exp((1-\alpha)\rho\Sigma_w)X_i\right)d\alpha \\ &\le \frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\int_{\alpha=0}^1 \|X_i\|_2\|\exp(\alpha\rho\Sigma_w)\|_2\|X_j\|_2\|X_j\|_2\|\exp((1-\alpha)\rho\Sigma_w)\|_2\|X_i\|_2 d\alpha \\ &= \frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\cdot R^4\cdot\|\exp(\rho\Sigma_w)\|_2 \le \rho R^4 \ .\end{aligned}$$

To conclude the proof, we bound from above the smoothness parameter by the spectral norm of the Hessian matrix. For any $w \in \Delta_{N,2\epsilon}$,

$$\left\|\nabla^2 f(w)\right\|_2 \leq N \cdot \max_{ij} \left|\nabla^2 f(w)_{ij}\right| \leq O(N\rho d^2) = \widetilde{O}(Nd^2/\epsilon) \ ,$$

where the last step uses that $\rho = \ln d/\epsilon$. □

Proof of Lemma C.11. Fix any $w \in \Delta_{N,2\epsilon}$. By Corollary C.2 and our choice of $\rho = \frac{\ln d}{\epsilon}$, we have

$$f(w) = \mathrm{smax}_\rho(\Sigma_w) \leq \lambda_{\max}(\Sigma_w) + \epsilon.$$

Therefore, it is sufficient to bound from above $\lambda_{\max}(\Sigma_w)$ by $O(d \log d)$.

The preprocessing step guarantees that all samples have ℓ_2-norm at most $\widetilde{O}(d^{1/2})$, consequently, the weighted empirical mean μ_w has ℓ_2-norm is at most $\widetilde{O}(d^{1/2})$ as well. Consequently,

$$\begin{aligned}\|\Sigma_w\|_2 &= \left\|\sum_{i=1}^N w_i (X_i - \mu_w)(X_i - \mu_w)^\top\right\|_2 \\ &\leq \sum_{i=1}^N w_i \left\|(X_i - \mu_w)(X_i - \mu_w)^\top\right\|_2 \leq \max_{i\in[N]} \|X_i - \mu_w\|_2^2 \leq \widetilde{O}(d) \ .\end{aligned}$$

The proof is now complete. □