High-dimensional Robust Covariance Estimation via Gradient Descent

Jay Sarva

May 2024

Abstract

We explore the relationship between outlier-robust high-dimensional covariance estimation and non-convex optimization. We seek to develop a simple gradient-descent-based formulation for this problem such that any approximate stationary point yields a near-optimal solution for the underlying robust estimation task. In particular, we look for an algorithm which is simple and practical to implement and compute.

1 Introduction

Estimating the covariance of a high-dimensional dataset is a fundamental statistical task. The empirical covariance is known to be a strong estimator for uncorrupted datasets sampled from standard distributions. In particular, suppose N samples are drawn i.i.d. from $\mathcal{N}(0, \Sigma)$ on $\mathbb{R}^d$ where Σ is unknown. We know that the empirical covariance has the best possible Frobenius error from the true Σ: $O(d/\sqrt{N})\|\Sigma\|_2$.
In this paper, we investigate the setting when a small fraction of the initial N samples get corrupted arbitrarily. The model is described in Def 1.1.

Definition 1.1 (Strong Contamination Model, [4]). *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 a number of samples n, and n samples are drawn from some unknown $D \in \mathcal{D}$. The adversary is allowed to inspect the samples, replace up to ϵn of 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.*

We study the following problem: Given an ϵ-corrupted set of N samples from an unknown $\mathcal{N}(0, \Sigma)$ on $\mathbb{R}^d$, compute an estimate of Σ in Frobenius norm. There are a number of prior works that discuss this problem but they generally require solving complex subproblems (see Section 2 for details on other algorithms). We search for a first-order algorithm to learn the empirical covariance by running gradient descent over some singular objective function. We utilize an approach similar to [1], where we search for some non-convex formulation which yields a near-optimal solution for the underlying robust estimation task.
Specifically, the goal is to find some simple function for this task such that all local optima are global optima.

In this work, we demonstrate partial progress towards proving that the following function satisfies such requirements:

$$f(w) = \max_{\|M\|_F = 1, M \in \mathbb{S}^{d^2}} vec(M)^\top (\Sigma_{w,Y} - (2 + O(\delta^4/\epsilon^3))\Sigma_{w,X}^{\otimes 2}) vec(M)$$

2 Related Work

2.1 Robust Statistical Learning in High-Dimensions

Statistical learning in presence of outliers has been a relevant goal since the 1960s [5]. More recently, sample-efficient and robust estimators were discovered [4]. These works provided robust estimators for high-dimensional statistical learning tasks like robust mean and covariance estimation.
In recent years, many high-dimensional robust estimation problems have been reduced to optimizing simple non-convex functions which have the property that all local minima are global minima. In particular, recent work showed this for high dimensional robust mean estimation [2, 6]. In particular, any stationary point of some proposed non-convex function would provide a near-optimal estimator of the mean.
Our work is directly inspired by another work which targets robust sparse mean and PCA estimation [1]. We describe our roadmap in a future section.

2.2 Algorithms for Robust High-dimensional Covariance Estimation

The first robust high-dimensional covariance estimator frameworks were proposed recently [4], but shortly afterwards, faster covariance estimation algorithms were proposed [3]. This work proposes an iterative refinement algorithm, where the bounds on the covariance estimate are slowly tightened by repeatedly solving some robust mean estimation subproblems.
Another recent work also proposes a first-order method for solving robust covariance estimation, although it requires solving a complex SDP to determine the gradient of the function [6]. We search for a simpler and more direct method which is computationally simpler.

3 Desired Results

Our fundamental goal is to find a function denoted $f(w)$ such that the following holds.

Theorem 3.1. *Fix $0 < \epsilon < \epsilon_0$ and $\delta > \epsilon$. Let G^* be a set of n samples that is (ϵ, δ)-stable as in Definition 5.1 with respect to a d-dimensional ground-truth distribution with unknown covariance Σ. Denote $S = (X_i)_{i=1}^n$ a ϵ-corrupted set of samples from G^*. Let $\gamma = O(n^{1/2}\delta^2\epsilon^{-3/2}$. If $\lambda = O(\delta^2/\epsilon)$, then for any γ-stationary point of $f(w)$ denoted as $\widehat{w}$, we can find $\|\Sigma^{-1/2}\Sigma_{\widehat{w},S}\Sigma^{-1/2}\|_F \le O(\delta)$ where $\Sigma_{\widehat{w},X}$ is the weighted covariance matrix as described in Section 4.1.*

3.1 Roadmap

In Section 4, we will introduce notation and the necessary stability conditions on the initial uncorrupted samples for this result to hold. In section 6, we will show that a negative objective function value is sufficient to guarantee a good estimate of the covariance for a sufficiently large λ. In Section 7, we show partial progress towards all local minima of the objective function have a negative objective value.

4 Preliminaries and Background

I will represent the identity matrix. For vector v, we use $\|v\|_0, \|v\|_1, \|v\|_2$ for the number of non-zeros, ℓ_1, ℓ_2 norms of v respectively. For matrix M, we use $\|M\|_2, \|M\|_F, \mathrm{tr}(M)$ for the spectral norm, Frobenius norm, and trace of M respectively. Matrix A is positive semidefinite if and only if $x^\top Ax \geq 0$ for all x. We write $A \preceq B$ if and only if $B - A$ is PSD. We will use $\otimes$ for the Kronecker product between matrices.
We denote $M^\flat = vec(M)$ to be the canonical flattening of matrix M. Similarly, we denote the matricization operation as $mat(M) = M^\#$.

4.1 Sample Reweighting Framework

We denote n as number of samples, d as the dimension, ϵ for fraction of corrupted samples. Let G^* be the original set of n good samples. We call $S = (G^* - L) \cup B = G \cup B$, where L is the removed good samples which are replaced by adversarily-added bad samples B and G is the remaining good samples. Here, $|G| = (1-\epsilon)n$ and $|B| = \epsilon n$.
Denote the given samples in S as $S = \{X_i\}_{i=1}^n$. Let $Y_i = X_i^{\otimes 2}$ for $i \in [n]$ where $\otimes$ denotes the Kronecker product. Also denote set $T = \{Y_i\}_{i=1}^n$.
$\Delta_{n,\epsilon}$ is the convex hull of all uniform distributions over subsets $T \subseteq [n]$ of size $|T| = (1-\epsilon)n$. Specifically:

$$\Delta_{n,\epsilon} = \{w \in \mathbb{R}^d \mid \|w\|_1 = 1 \text{ and } 0 \leq w_i \leq \frac{1}{(1-\epsilon)n} \forall i\}$$

We denote weighted empirical mean and weighted empirical covariance as follows, where $w \in \triangle_{n,\epsilon}$ and $S = \{X_i\}_{i=1}^n$.

$$\mu_{w,S} = \Sigma_i w_i X_i$$

$$\Sigma_{w,S} = \Sigma_i w_i (X_i - \mu_{w,X})(X_i - \mu_{w,X})^\top$$

$\lambda_{max}(A)$ denotes the maximum eigenvalue of matrix A. Let w^* be the uniform weight vector on the good samples.

4.2 Initial Naive Normalization and Filtering

We first normalize all elements by $\|\widehat{\Sigma}\|_2$ where $\widehat{\Sigma}$ is estimated from median of $\|X_i\|_2$:

$$\widehat{\Sigma} = \sum_{i=1}^{n} \frac{1}{n} (X_i - med(X))(X_i - med(X))^\top$$

Then, remove all points which are farther than $O(d \log(d))$ from the coordinate-wise median. This gives concrete bounds on $\|X_i\|_2$ and $\|Y_i\|_2$ which are used in Section 7.

5 Stability Conditions

Definition 5.1. *A set of n samples $G = \{X_i\}_{i=1}^n$ is said to be (ϵ, δ)-stable with respect to a distribution with mean μ and covariance Σ iff for any weight vector $w \in \Delta_{n,2\epsilon}$, we have*

1. $\|\mu_{w,G} - \mu\|_2 \leq O(\delta)$

2. $\|\Sigma_{w,G} - \Sigma\|_2 \leq O(\delta^2/\epsilon)\|\Sigma\|_2$

3. $\|\Sigma_{w,\{X_i \otimes X_i\}_{i=1}^n} - 2\Sigma^{\otimes 2}\|_{\mathcal{S}} \leq O(\delta^4/\epsilon^3)\|\Sigma\|_2^2$

6 Any negative objective suffices

Definition 6.1 (Robust-Covariance Guarantee). *Set $G = \{X_i\}_{i=1}^n$ is (A, ϵ, δ)-guaranteed w.r.t covariance Σ if the following holds, where $Z_i = A^{-1/2}X_i \otimes A^{-1/2}X_i$ and $G' = \{Z_i\}_{i=1}^n$:*

1. $\|\mu_{w,G'} - vec(A^{-1/2}\Sigma A^{-1/2})\|_2 \leq \delta$

2. $\|\Sigma_{w,G'} - 2(A^{-1/2}\Sigma A^{-1/2})^{\otimes 2}\|_{\mathcal{S}} \leq \delta^2/\epsilon$

This section goes about proving robust-covariance guarantee conditions on a particular normalization of the good samples, specifically $\{vec(\Sigma_{w,S}^{-1/2}X_iX_i^\top\Sigma_{w,S}^{-1/2})\}_{i\in G}$. Note that $\Sigma_{w,S}$ is calculated over the corrupted samples, while the set itself is only contains rotated good samples. Then, if this condition holds alongside the objective function being small enough, $\|\Sigma^{-1/2}\Sigma_{w,S}\Sigma^{-1/2} - I\|_F$ is small.

Lemma 6.2 (Negative $f(w)$ suffices). *Suppose $S = \{X_i\}_{i=1}^n$ is ϵ-corrupted copy of $G^* \subset \mathbb{R}^d$ which is (ϵ, δ)-stable set. Also denote set $T = \{X_i \otimes X_i\}_{i=1}^n$. Let $w \in \Delta_{n,\epsilon}$ be such that $f(w) = \lambda_{max}(\Sigma_{w,T} - (2+\lambda)\Sigma_{w,S}^{\otimes 2}) < 0$. With $\lambda = O(\delta^2/\epsilon)$, this implies $\|\Sigma^{-1/2}\Sigma_{w,S}\Sigma^{-1/2} - I\|_F \leq O(\delta)$.*

Proof. Define $Z_i = \Sigma_{w,S}^{-1/2}X_i \otimes \Sigma_{w,S}^{-1/2}X_i$ and set $T_{\Sigma_{w,S}} = \{Z_i\}_{i=1}^n$.

$$\begin{aligned}
\Sigma_{w,T} - (2+\lambda)\Sigma_{w,S} &\preceq 0 \\
\Sigma_{i\in S}w_i[((X_i \otimes X_i) - \Sigma_{j\in S}w_j[X_j \otimes X_j])((X_i \otimes X_i) - \Sigma_{j\in S}w_j[X_j \otimes X_j])^\top] &\preceq (2+\lambda)\Sigma_{w,S} \\
\Sigma_{i\in S}w_i[(Z_i - \Sigma_{j\in S}w_j[Z_j])(Z_i - \Sigma_{j\in S}w_j[Z_j])^\top] &\preceq (2+\lambda)I \\
\lambda_{max}(\Sigma_{i\in S}w_i[(Z_i - \Sigma_{j\in S}w_j[Z_j])(Z_i - \Sigma_{j\in S}w_j[Z_j])^\top]) &\leq 2+\lambda \\
\lambda_{max}(\Sigma_{w,T_{\Sigma_{w,S}}}) &\leq 2+\lambda
\end{aligned}$$

The second step is due to Fact 2.
Apply Lemma 6.3. Then, we have $(\Sigma_{w,S}, \epsilon, O(\delta + \sqrt{\epsilon}))$-guarantee on the set G^*. Repeatedly apply Lemma 6.6. Eventually, the guarantee condition converges to $(\Sigma_{w,S}, \epsilon, O(\delta + \epsilon^{3/4}\lambda^{1/4}))$. By our definition of $\lambda = O(\delta^2/\epsilon)$, we can claim the following from use of the robust mean certificate on set $T_{\Sigma_{w,S}}$ with respect to mean $vec(\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2})$ (Lemma 2.4 from [**?**])

$$\|\Sigma_{w,S}^{-1/2}(\Sigma - \Sigma_{w,S})\Sigma_{w,S}^{-1/2}\|_F \leq O(\delta)$$

The desired multiplicative error is at most a multiple of 2 off, by Fact 3 so

$$\|\Sigma^{-1/2}(\Sigma - \Sigma_{w,S})\Sigma^{-1/2}\|_F \leq O(\delta)$$

which gives the desired inequality:

$$\|\Sigma^{-1/2}\Sigma_{w,S}\Sigma^{-1/2}\|_F \leq O(\delta)$$

□

6.1 Initial Stability Conditions

Lemma 6.3. *Suppose $w \in \Delta_{w,\epsilon}$. G^* is $(\Sigma_{w,S}, \epsilon, O(\delta + \sqrt{\epsilon}))$-guaranteed.*

Proof. $\Sigma_{w,S} \succeq \Sigma/2$ from Lemma 6.4.

$$\Sigma_{w,S} \succeq \Sigma/2 \tag{1}$$
$$2I \succeq \Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2} \tag{2}$$
$$I \succeq \Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2} - I \tag{3}$$
$$1 \geq \|\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2} - I\|_2 \tag{4}$$

Now, apply Lemma 6.5 for $(\Sigma_{w,S}, \epsilon, O(\delta + \sqrt{\epsilon}))$-guarantee. □

Lemma 6.4. *Suppose $w \in \Delta_{w,\epsilon}$. Then, $\Sigma_{w,S} \succeq \Sigma/2$*

Proof. Let $\alpha = \|w_G\|_1$ and $\beta = \|w_B\|_1$. Let $\overline{w} = w_G/\alpha$ and $\widehat{w} = w_B/\beta$. Notice that $w = \alpha\overline{w} + \beta\widehat{w}$. Fact 1 states that

$$\Sigma_{w,S} = \alpha\Sigma_{\overline{w},S} + \beta\Sigma_{\widehat{w},S} + \alpha\beta(\mu_{\overline{w},S} - \mu_{\widehat{w},S})(\mu_{\overline{w},S} - \mu_{\widehat{w},S})^\top$$

We can assume WLOG that the original (good) samples are from a centered distribution (mean 0), so

$$\Sigma_{w,S} = \alpha\Sigma_{i\in G}w_iX_iX_i^\top + \beta\Sigma_{\widehat{w},S} + \alpha\beta(\mu_{\overline{w},S} - \mu_{\widehat{w},S})(\mu_{\overline{w},S} - \mu_{\widehat{w},S})^\top$$

Notice that $\alpha\Sigma_{i\in G}w_iX_iX_i^\top \succeq (1 - O(\delta))\Sigma$. Thus, it holds that $\Sigma_{w,S} \succeq (1 - O(\delta))\Sigma \succeq \Sigma/2$. □

6.2 Recursive Iteration on Stability Conditions

Lemma 6.5. *Consider $S = \{X_i\}_{i=1}^n$ to be a ϵ-corrupted set copy of (ϵ, δ)-stable set G^*. Suppose $\|\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2} - I\|_2 < 1$ where Σ is the true covariance of distribution G^* is sampled from. G^* is $(\epsilon, O(\delta + \sqrt{\epsilon\|\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2} - I\|_2}))$-guaranteed.*

Proof. $\Sigma_{w,S} \succeq \Sigma/2$ by Lemma 6.4.
Define linear operator A such that $\Sigma_{w,S}^{-1/2}X_i \otimes \Sigma_{w,S}^{-1/2}X_i = A(\Sigma^{-1/2}X_i \otimes \Sigma^{-1/2}X_i)$. A is PSD and has eigenvalues $(a_ia_j)^{-1}$ where a_k are eigenvalues of $\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2}$. This is justified in Lemma 6.7. For convenience, define a_{max} and a_{min} to be the largest and smallest eigenvalues of $\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2}$ respectively.
We are concerned with covariance for the set of $Z_i = \Sigma_{w,S}^{-1/2}X_i \otimes \Sigma_{w,S}^{-1/2}X_i$ where $i \in G^*$, or set $H_{\Sigma_{w,S}} = \{Z_i\}_{i\in G^*}$. Also denote $H_\Sigma = \{\Sigma^{-1/2}X_i \otimes \Sigma^{-1/2}X_i\}_{i\in G^*}$ where Σ is the true covariance. Then, it must hold that $\Sigma_{w,H_{\Sigma_{w,S}}} = A\Sigma_{w,H_\Sigma}A^\top$.
We care how far eigenvalues of $\Sigma_{w,H_{\Sigma_{w,S}}}$ from 1. Notice that $\{X_i\}_{i\in G^*}$ is $(\Sigma_{w^*,S}, \epsilon, \delta)$-guaranteed, following from stability of G^*.

$$\begin{aligned} m^\top\Sigma_{w,H_{\Sigma_{w,S}}}m - 1 &\leq a_{max}(1 + O(\delta^2/\epsilon)) - 1 \\ &\leq \|\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2}\|_2(1 + O(\delta^2/\epsilon)) - 1 \\ &\leq \|\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2} - I\|_2 + O(\delta^2/\epsilon) \end{aligned}$$

This proves the second part of stability from Defn 6.1. The first part of stability follows directly. Thus, we have $(\epsilon, O(\delta + \sqrt{\epsilon \|\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2} - I\|_2}))$-stability. $\square$

Lemma 6.6. *Suppose $w \in \Delta_{n,\epsilon}$. Suppose also that $\lambda_{max}(\Sigma_{w,T_{\Sigma_{w,S}}}) \leq 2 + \lambda$ where set $T_{\Sigma_{w,S}} = \{Z_i\}_{i=1}^n$. Suppose also that $(\Sigma_{w,S}, \epsilon, \xi)$-guarantee conditions on G^*. Then, it is also $(\Sigma_{w,S}, \epsilon, O(\delta + \sqrt{\epsilon}(\sqrt{\xi} + \epsilon^{1/4}\lambda^{1/4})))$-guaranteed.*

Proof. $\Sigma_{w,S} \succeq \Sigma/2$ by Lemma 6.4. Apply a variant of Lemma 2.4 from [**?**]. Notice that $\mu_{w,T_{\Sigma_{w,S}}} = vec(\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2})$ and similarly, $\mu = vec(\Sigma_{w,S}^{-1/2}\Sigma_{w,S}\Sigma_{w,S}^{-1/2})$.
Then,

$$\|vec(\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2}) - vec(\Sigma_{w,S}^{-1/2}\Sigma_{w,S}\Sigma_{w,S}^{-1/2})\|_2 = \sqrt{2}\|\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2} - I\|_F = O(\xi + \sqrt{\epsilon\lambda}).$$

Apply Lemma 6.5 to get desired result. $\square$

Lemma 6.7. *Suppose there is operator $F(a) = vec(\Sigma_{w,S}^{-1/2}\Sigma^{1/2}mat(a)\Sigma^{1/2}\Sigma_{w,S}^{-1/2})$. This is a linear operator with matrix representation $A := \Sigma_{w,S}^{-1/2}\Sigma^{1/2} \otimes \Sigma_{w,S}^{-1/2}\Sigma^{1/2}$ which is PSD, with eigenvalues $(k_i k_j)^{-1/2}$ where k_i are eigenvalues of $\Sigma_{w,S}^{-1/2}\Sigma\Sigma_{w,S}^{-1/2}$.*

Proof. Suppose we have operator B on $\mathbb{R}^{n\times n} \to \mathbb{R}^{n^2}$ where $B \longmapsto vec(\Sigma^{-1/4}B\Sigma^{-1/4})$. Then,

$$F(a) = B((\Sigma^{1/4}\Sigma_{w,S}^{-1/2}\Sigma^{1/4})B^{-1}(a)(\Sigma^{1/4}\Sigma_{w,S}^{-1/2}\Sigma^{1/4}))$$

There is conjugate operator C defined by $M \longmapsto (\Sigma^{1/4}\Sigma_{w,S}^{-1/2}\Sigma^{1/4})M(\Sigma^{1/4}\Sigma_{w,S}^{-1/2}\Sigma^{1/4})$. Therefore, A and C have same eigenvalues, which must be $(k_i k_j)^{-1/2}$. Now, we show A is PSD and find its matrix representation.

$$\begin{aligned}
F(a) &= vec(\Sigma_{w,S}^{-1/2}\Sigma^{1/2}mat(a)\Sigma^{1/2}\Sigma_{w,S}^{-1/2}) \\
&= (\Sigma_{w,S}^{-1/2}\Sigma^{1/2} \otimes \Sigma_{w,S}^{-1/2}\Sigma^{1/2})vec(mat(a)) \\
&= (\Sigma_{w,S}^{-1/2}\Sigma^{1/2} \otimes \Sigma_{w,S}^{-1/2}\Sigma^{1/2})a
\end{aligned}$$

Thus, matrix operator is $A = \Sigma_{w,S}^{-1/2}\Sigma^{1/2} \otimes \Sigma_{w,S}^{-1/2}\Sigma^{1/2}$. This is trivially PSD, as $\Sigma_{w,S}$ and Σ are both PSD as covariance matrices. $\square$

7 All Local Minima are Negative

Lemma 7.1. $f(w^*) < -O(\delta^2/\epsilon)\|\Sigma\|_2^2$

Proof. Note $\Sigma_{w^*,Y} = 2\Sigma_{w^*,X}^{\otimes 2}$ Therefore

$$\begin{aligned}
f(w^*) &= \lambda_{max}(\Sigma_{w^*,Y} - (2+\lambda)\Sigma_{w^*,X}^{\otimes 2}) \\
&\le \lambda_{max}(\Sigma_Y + O(\delta^2/\epsilon)I - (2+\lambda)\Sigma_{w^*,X}^{\otimes 2}) \\
&\le \lambda_{max}(2\Sigma_{w^*,X}^{\otimes 2} - (2+\lambda)\Sigma_{w^*,X}^{\otimes 2}) \\
&\le \lambda_{max}(2\Sigma_{w^*,X}^{\otimes 2} - (2+\lambda)\Sigma_{w^*,X}^{\otimes 2}) \\
&\le \lambda_{max}(-\lambda\Sigma_{w^*,X}^{\otimes 2}) \\
&\le -\lambda\|\Sigma_{w^*,X}^{\otimes 2}\|_2^2 \\
&\le -\lambda
\end{aligned}$$

where the last step holds because of the naive filtering argument provided initially. $\square$

Lemma 7.2. *All local minima $w \in \Delta_{n,\epsilon}$ of the objective function f satisfies that $f(w) < 0$.*

Proof. We will prove this by showing that if $f(w) > 0$, the derivative of $f((1-t)w + tw^*)$ with respect to t at $t = 0$ is negative.
First, we consider $f((1-t)w + tw^*)$. Both terms can be expanded and simplified via Fact 1:

$$\Sigma_{(1-t)w+tw^*,T} = (1-t)\Sigma_{w,T} + t\Sigma_{w^*,T} + t(1-t)(\mu_{w,T} - \mu_{w^*,T})(\mu_{w,T} - \mu_{w^*,T})^\top$$

$$\begin{aligned}
\Sigma_{(1-t)w+tw^*,S}^{\otimes 2} &= ((1-t)\Sigma_{w,S} + t\Sigma_{w^*,S} + t(1-t)(\mu_{w,S} - \mu_{w^*,S})(\mu_{w,S} - \mu_{w^*,S})^\top)^{\otimes 2} \\
&\succeq ((1-t)\Sigma_{w,S} + t\Sigma_{w^*,S})^{\otimes 2} \\
&= (1-t)\Sigma_{w,S}^{\otimes 2} + t\Sigma_{w^*,S}^{\otimes 2} - t(1-t)(\Sigma_{w,S} - \Sigma_{w^*,S})^{\otimes 2}
\end{aligned}$$

At this point, there is a roadblock – we would like to show that $f((1-t)w + tw^*) < f(w)$ to show that as $w \to w*$, the objective function will always decrease for w such that $f(w) > 0$. However, following the above calculations, we can only show:

$$f((1-t)w + tw^*) \le (1-t)f(w) + tf(w^*) + t(1-t)\|(\mu_{w,T} - \mu_{w^*,T})(\mu_{w,T} - \mu_{w^*,T})^\top + \overline{\lambda}(\Sigma_{w,S} - \Sigma_{w^*,S})^{\otimes 2}\|_2$$

We must bound the 3rd term. It is still open as to how to bound this term. $\square$

7.1 Key Structural Lemma: Good 4th Moment Implies Good 2nd Moment

Lemma 7.3. *Suppose $w \in \Delta_{n,\epsilon}$. Then,*

$$\Sigma_{w,T} - 2\Sigma^{\otimes 2} \succeq O(1/\epsilon)((\mu_{w,T} - \mu_{w^*,T})(\mu_{w,T} - \mu_{w^*,T})^\top) - O(\delta^2/\epsilon)\|\Sigma\|_2^2 I$$

Proof. First, we define some notation. Given $w \in \mathbb{R}^d$, we write w_G for a vector with $w_G(i) = w(i)$ if $i \in G$ and 0 if $i \in B$. We define w_B similarly by restricting to w to B.
Let $\alpha = \|w_G\|_1$ and $\beta = \|w_B\|_1$. Let $\overline{w} = w_G/\alpha$ and $\widehat{w} = w_B/\beta$. Notice that $w = \alpha\overline{w} + \beta\widehat{w}$.
Fact 1 states that

$$\Sigma_{w,T} = \alpha\Sigma_{\overline{w},T} + \beta\Sigma_{\widehat{w},T} + \alpha\beta(\mu_{\overline{w},T} - \mu_{\widehat{w},T})(\mu_{\overline{w},T} - \mu_{\widehat{w},T})^\top$$

For all vectors $v \in \mathbb{R}^d$:

$$
\begin{aligned}
v^\top(\Sigma_{w,T} - 2\Sigma^{\otimes 2})v &= \alpha v^\top \Sigma_{\overline{w},T} v + \beta v^\top \Sigma_{\widehat{w},T} v + \alpha\beta((\mu_{\overline{w},T} - \mu_{\widehat{w},T})^\top v)^2 - \|2\Sigma^{\otimes 2}\|_2 \\
&\geq \alpha v^\top \Sigma_{\overline{w},T} v - \|2\Sigma^{\otimes 2}\|_2 + \alpha\beta((\mu_{\overline{w},T} - \mu_{\widehat{w},T})^\top v)^2 \\
&\approx \alpha\beta((\mu_{\overline{w},T} - \mu_{\widehat{w},T})^\top v)^2 \\
&= \frac{\alpha}{\beta}(\beta(\mu_{\overline{w},T} - \mu_{\widehat{w},T})^\top v)^2
\end{aligned}
$$

where the last approximately equals is due to stability conditions: $\alpha v^\top \Sigma_{\overline{w},T} v \approx \|2\Sigma^{\otimes 2}\|_2$.
Also,

$$
\begin{aligned}
\beta(\mu_{\widehat{w},T} - \mu_{\overline{w},T}) &= \beta\mu_{\widehat{w},T} + \alpha\mu_{\overline{w},T} - \mu_{\overline{w},T} \\
&= \mu_{w,T} - \mu_{\overline{w},T} \\
&= (\mu_{w,T} - \mu_{w^*,T}) + (\mu_{w^*,T} - \mu_{\overline{w},T})
\end{aligned}
$$

so we can use stability conditions to say that:

$$
\begin{aligned}
|(\mu_{w^*,T} - \mu_{\overline{w},T})^\top v| &\leq \|\mu_{w^*,T} - \mu_{\overline{w},T}\|_2 \\
&\leq \|\mu_{w^*,T} - \mu_Y\|_2 + \|\mu_Y - \mu_{\overline{w},T}\|_2 \\
&\leq O(\delta)\|\Sigma\|_2
\end{aligned}
$$

We will use this, in turn, to have the following inequality:

$$
\begin{aligned}
(\beta(\mu_{\widehat{w},T} - \mu_{\overline{w},T})^\top v)^2 &= ((\mu_{w,T} - \mu_{w^*,T})^\top v + (\mu_{w^*,T} - \mu_{\overline{w},T})^\top v) \\
&\geq \frac{1}{2}((\mu_{w,T} - \mu_{w^*,T})^\top v)^2 - ((\mu_{\overline{w},T} - \mu_{w^*,T})^\top v)^2 \\
&\geq \frac{1}{2}((\mu_{w,T} - \mu_{w^*,T})^\top v)^2 - O(\delta^2)\|\Sigma\|_2^2
\end{aligned}
$$

Now we can finally prove the desired statement:

$$
\begin{aligned}
v^\top(\Sigma_{w,T} - 2\Sigma^{\otimes 2})v &\geq \frac{\alpha}{\beta}\left(\beta(\mu_{\overline{w},Y} - \mu_{\widehat{w},T})^\top v\right)^2 \\
&\geq \frac{1-2\epsilon}{\epsilon}\left(\frac{1}{2}((\mu_{w,T} - \mu_{w^*,T})^\top v)^2 - O(\delta^2)\|\Sigma\|_2^2\right) \\
&\geq \frac{1}{4\epsilon}\left(((\mu_{w,T} - \mu_{w^*,T})^\top v)^2 - O(\delta^2)\|\Sigma\|_2^2\right) \\
&\geq O(1/\epsilon)v^\top(\mu_{w,T} - \mu_{w^*,T})(\mu_{w,T} - \mu_{w^*,T})^\top v - O(\delta^2/\epsilon)\|\Sigma\|_2^2 \qquad \square
\end{aligned}
$$

7.2 Algorithmic Results: Finding Stationary Points

We show that one can find an approximate stationary point for the suggested objective function in a polynomial number of iterations.
Note the analysis is generally just a sketch and not optimal.

Lemma 7.4. *$F(w,m) = m^\top(\Sigma_{w,Y} - \lambda\Sigma_{w,X}^{\otimes 2})m$ We show $F(w,m)$ is L-Lipschitz and β-smooth for $L = O(\sqrt{N}d^2\log^2(d))$, $\beta = O(Nd^2\log^2(d))$*

Proof. First, notice that $F(w, M) = vec(M)^\top(\Sigma_{w,Y} - \lambda\Sigma_{w,X}^{\otimes 2})vec(M) = vec(M)^\top\Sigma_{w,Y}vec(M) - \|\Sigma_{w,X}^{1/2}M\Sigma_{w,X}^{1/2}\|_2$. For notational simplicity, we can define $B = X^\top MX$. The gradient of $F(w, M)$ with respect to w is:

$$Y^\top vec(M)\odot Y^\top vec(M) - 2vec(M)^\top YwY^\top vec(M) - \lambda(2(B^\top\odot B)w + 2(3B^\top w\odot Bw) + 4(w^\top Bw)B^\top w)$$

Without loss of generality, we will let $\max\|X_i\|_2 \le O(\sqrt{d\log(d)})$ and $\max\|Y_i\|_2 \le O(d\log(d))$ which we can say by normalizing our samples by an estimate of $\|\Sigma\|_2$ via the median of $\|X_i\|_2$ and then removing all samples which have $\|Y_i\|_2$ farther than $O(d\log(d))$ from the origin.

$$\begin{aligned}\nabla_w F(w, M) &\le \sqrt{N}\max_i\|Y_i\|_2^2 + 2\max_i\|Y_i\|_2\|Y\|_2 + \lambda\|(B^\top\odot B)w\|_2 + \lambda\|B^\top w\odot Bw\|_2 + \lambda\|w^\top BwB^\top w\|_2 \\ &\le O(\sqrt{N}d^2\log^2(d))\end{aligned}$$

Thus, $L = O(\sqrt{N}d^2\log^2(d))$.
The hessian of $F(w, M)$ with respect to w is:

$$-2Y^\top vec(M)Y^\top vec(M) - \lambda(2(B^\top\odot B) + \frac{3}{2}Bdiag(Bw) + 12w^\top BwB)$$

Thus,

$$\|\nabla_w^2 F(w, M)\|_2 \le O(Nd^2\log^2(d))$$

So $\beta = O(Nd^2\log^2(d))$

$\square$

Lemma 7.5. *Fix ϵ. Let $S = \{X_i\}_{i=1}^n$ be ϵ-corrupted set of $n = \widetilde{O}(d^2/\epsilon^2)$ samples drawn from a centered gaussian distribution with unknown covariance. Let $Y_i = X_i\otimes X_i$. Consider optimization problem $\min_{w\in\Delta_{n,\epsilon}}\max_{\|M\|_F=1} vec(M)^\top(\Sigma_{w,Y} - \lambda\Sigma_{w,X}^{\otimes 2})vec(M)$. After $\widetilde{O}(d^8/\epsilon^2)$ iterations, projected subgradient descent will output $w\in\Delta_{n,\epsilon}$ such that $\|\Sigma^{-1/2}\Sigma_{w,X}\Sigma^{-1/2} - I\| \le O(\epsilon\log(1/\epsilon))$*

Proof. We showed previously that $\gamma = -O(n^{1/2}\delta^2\epsilon^{-3/2})$.
From Lemma 7.4, $L = O(\sqrt{N}d^2\log^2(d))$, $\beta = O(Nd^2\log^2(d))$. Also note that

$$B := f_\beta(w_0) - \min_w f(w) \le f_\beta(w_0) \le f(w_0) \le \|\Sigma_{w,Y}\|_2^2 \le \widetilde{O}(d^2)$$

Therefore, by Lemma C.4 of [1], the iteration count is $T \ge O(B\beta L^2/\gamma^4) = O(d^8/\epsilon^2)$ to guarantee that $\|\nabla f_\beta(w)\|_2 \le \gamma$. $\square$

8 Ongoing and Future Work

In the future, we will continue trying to resolve the gaps in the proof of Section 7. There also may be a simpler formulation of the objective function which could lead to a less complex proof in Section 6. We have some initial work into such a simpler formulation via theorems in 1-dimensions – it remains to prove these theorems in the high-dimensional setting.

References

[1] Yu Cheng, Ilias Diakonikolas, Rong Ge, Shivam Gupta, Daniel M. Kane, and Mahdi Soltanolkotabi. Outlier-robust sparse estimation via non-convex optimization, 2022.

[2] Yu Cheng, Ilias Diakonikolas, Rong Ge, and Mahdi Soltanolkotabi. High-dimensional robust mean estimation via gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1768–1778. PMLR, 13–18 Jul 2020.

[3] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David Woodruff. Faster algorithms for high-dimensional robust covariance estimation, 2019.

[4] Ilias Diakoniklas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability, 2019.

[5] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

[6] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Robust estimation via generalized quasi-gradients, 2020.

A List of Facts

Fact 1. *Fix n samples $X_1, \ldots, X_n \in \mathbb{R}^d$ in set S. Let $\overline{w}, \widehat{w} \in \mathbb{R}^d$ be two non-negative weight vectors with $\|\overline{w}\|_2 = \|\widehat{w}\|_2 = 1$. Then, $\forall \alpha, \beta \geq 0$ where $\alpha + \beta = 1$, the following holds with $w = \alpha\overline{w} + \beta\widehat{w}$*

$$\Sigma_{w,S} = \alpha\Sigma_{\overline{w},S} + \beta\Sigma_{\widehat{w},S} + \alpha\beta(\mu_{\overline{w},S} - \mu_{\widehat{w},S})(\mu_{\overline{w},S} - \mu_{\widehat{w},S})^\top .$$

Fact 2. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.

Fact 3. *Suppose $\delta < 1$ and A, B positive semidefinite. Then, $\|A^{-1/2}BA^{-1/2} - I\|_F \leq \delta$ implies $\|B^{-1/2}AB^{-1/2} - I\|_F \leq 2\delta$.*