Robust Second-Order Nonconvex Optimization and Its Application to Low Rank Matrix Sensing

Shuyao Li
University of Wisconsin-Madison
shuyao.li@wisc.edu

Yu Cheng
Brown University
yu_cheng@brown.edu

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

Jelena Diakonikolas
University of Wisconsin-Madison
jelena@cs.wisc.edu

Rong Ge
Duke University
rongge@cs.duke.edu

Stephen Wright
University of Wisconsin-Madison
swright@cs.wisc.edu

Abstract

Finding an approximate second-order stationary point (SOSP) is a well-studied and fundamental problem in stochastic nonconvex optimization with many applications in machine learning. However, this problem is poorly understood in the presence of outliers, limiting the use of existing nonconvex algorithms in adversarial settings.

In this paper, we study the problem of finding SOSPs in the strong contamination model, where a constant fraction of datapoints are arbitrarily corrupted. We introduce a general framework for efficiently finding an approximate SOSP with *dimension-independent* accuracy guarantees, using $\widetilde{O}(D^2/\epsilon)$ samples where D is the ambient dimension and ϵ is the fraction of corrupted datapoints.

As a concrete application of our framework, we apply it to the problem of low rank matrix sensing, developing efficient and provably robust algorithms that can tolerate corruptions in both the sensing matrices and the measurements. In addition, we establish a Statistical Query lower bound providing evidence that the quadratic dependence on D in the sample complexity is necessary for computationally efficient algorithms.

1 Introduction

Learning in the presence of corrupted data is a significant challenge in machine learning (ML) with many applications, including ML security [Bar+10; BNL12; SKL17; Dia+19] and exploratory data analysis of real datasets, e.g., in biological settings [Ros+02; Pas+10; Li+08; Dia+17]. The goal in such scenarios is to design efficient learning algorithms that can tolerate a small constant fraction of outliers and achieve error guarantees independent of the dimensionality of the data. Early work in robust statistics [Ham+86; HR09] gave sample-efficient robust estimators for various tasks (e.g., the Tukey median [Tuk75] for robust mean estimation), alas with runtimes exponential in the dimension. A recent line of work in computer science, starting with [Dia+16; LRV16], developed the first computationally efficient robust algorithms for several fundamental high-dimensional tasks. Since these early works, there has been significant progress in algorithmic aspects of robust high-dimensional statistics (see [DK19] and [DK23] for a comprehensive overview).

In this paper, we study the general problem of smooth (with Lipschitz gradient and Hessian) stochastic nonconvex optimization $\min_x \bar{f}(x)$ in the outlier-robust setting, where $\bar{f}(x) := \mathbb{E}_{A\sim\mathcal{G}} f(x, A)$ and $\mathcal{G}$ is a possibly unknown distribution of the random parameter A. We will focus on the following standard adversarial contamination model (see, e.g., [Dia+16]).

Definition 1.1 (Strong Contamination Model)**.** Given a parameter $0 < \epsilon < 1/2$ and an inlier distribution $\mathcal{G}$, an algorithm receives samples from $\mathcal{G}$ with ϵ-contamination as follows: The algorithm first specifies the number of samples n, and n samples are drawn independently from $\mathcal{G}$. An adversary is then allowed to inspect these samples, and replace an ϵ-fraction of the samples with arbitrary points. This modified set of n points is said to be ϵ-corrupted, which is then given to the algorithm.

The stochastic optimization problem we consider is computationally intractable in full generality — even without corruption — if the goal is to obtain globally optimal solutions. At a high level, an achievable goal is to design sample and computationally efficient robust algorithms for finding *locally* optimal solutions. Prior work [Pra+20; Dia+19] studied outlier-robust stochastic optimization and obtained efficient algorithms for finding approximate *first-order* stationary points. While first-order guarantees suffice for convex problems, it is known that in many tractable non-convex problems, first-order stationary points may be bad solutions, but all *second-order* stationary points (SOSPs) are globally optimal. This motivates us to study the following questions:

> *Can we develop a general framework for finding* ***second-order*** *stationary points in outlier-robust stochastic optimization?*
>
> *Can we obtain sample and computationally efficient algorithms for outlier-robust versions of tractable nonconvex problems using this framework?*

In this work, we answer both questions affirmatively. We introduce a framework for efficiently finding an approximate SOSP when ϵ-fraction of the functions are corrupted and then use our framework to solve the problem of outlier-robust low rank matrix sensing.

In addition to the gradient being zero, a SOSP requires the Hessian matrix to not have negative eigenvalues. The second-order optimality condition is important because it rules out suboptimal solutions such as strict saddle points. It is known that all SOSPs are globally optimal in nonconvex formulations of many important machine learning problems, such as matrix completion [GLM16], matrix sensing [BNS16], phase retrieval [SQW16], phase synchronization [BBV16], dictionary learning [SQW17], and tensor decomposition [Ge+15] (see also [WM22, Chapter 7]). However, the properties of SOSPs are highly sensitive to perturbation in the input data. For example, it is possible to create spurious SOSPs for nonconvex formulations of low rank matrix recovery problems, even for a semi-random adversary that can add additional sensing matrices but cannot corrupt the measurements in matrix sensing [GC23] or an adversary who can only reveal more entries of the ground-truth matrix in matrix completion [CG18]. Those spurious SOSPs correspond to highly suboptimal solutions.

Finding SOSPs in stochastic nonconvex optimization problems in the presence of arbitrary outliers was largely unaddressed prior to our work. Prior works [Pra+20; Dia+19] obtained efficient and robust algorithms for finding *first-order* stationary points with dimension-independent accuracy guarantees. These works relied on the following simple idea: Under certain smoothness assumptions, projected gradient descent with an *approximate* gradient oracle efficiently converges to an *approximate* first-order stationary point. Moreover, in the outlier-robust setting, approximating the gradient at a specific point amounts to a robust mean estimation problem (for the underlying distribution of the gradients), which can be solved by leveraging existing algorithms for robust mean estimation.

Our work is the first to find approximate SOSPs with dimension-independent errors in outlier-robust settings. Note that in standard non-robust settings, approximate SOSPs can be computed using first-order methods such as perturbed gradient descent [Jin+17; Jin+21]. This strategy might seem extendable to outlier-robust settings through perturbed approximate gradient descent, utilizing robust mean estimation algorithms to approximate gradients. The approach in [Yin+19] follows this idea, but unfortunately their second-order guarantees scale polynomially with dimension, even under very strong distributional assumptions (e.g., subgaussianity). Our lower bound result provides evidence that approximating SOSPs with dimension-independent error is as hard as approximating *full* Hessian, suggesting that solely approximating the gradients is not sufficient. On a different note, [IPL23] recently employed robust estimators for both gradient and Hessian in solving certain convex

stochastic optimization problems, which has a different focus than ours and does not provide SOSPs with the guarantees that we achieve.

1.1 Our Results and Contributions

The notation we use in this section is defined in Section 2. To state our results, we first formally define our generic nonconvex optimization problem. Suppose there is a true distribution over functions $f : \mathbb{R}^D \times \mathcal{A} \to \mathbb{R}$, where $f(x, A)$ takes an argument $x \in \mathbb{R}^D$ and is parameterized by a random variable $A \in \mathcal{A}$ drawn from a distribution $\mathcal{G}$. Our goal is to find an (ϵ_g, ϵ_H)-approximate SOSP of the function $\bar{f}(x) := \mathbb{E}_{A\sim\mathcal{G}} f(x, A)$.

Definition 1.2 (ϵ-Corrupted Stochastic Optimization)**.** The algorithm has access to n functions $(f_i)_{i=1}^n$ generated as follows. First n random variables $(A_i)_{i=1}^n$ are drawn independently from $\mathcal{G}$. Then an adversary arbitrarily corrupts an ϵ fraction of the A_i's. Finally, the ϵ-corrupted version of $f_i(\cdot) = f(\cdot, A_i)$ is sent to the algorithm as input. The task is to find an approximate SOSP of the ground-truth average function $\bar{f}(\cdot) := \mathbb{E}_{A\sim\mathcal{G}} f(\cdot, A)$.

Definition 1.3 (Approximate SOSPs)**.** A point x is an (ϵ_g, ϵ_H)-approximate second-order stationary point (SOSP) of $\bar{f}$ if $\left\|\nabla \bar{f}(x)\right\| \le \epsilon_g$ and $\lambda_{\min}\left(\nabla^2 \bar{f}(x)\right) \ge -\epsilon_H$.

We make the following additional assumptions on f and $\mathcal{G}$.

Assumption 1.4. *There exists a bounded region $\mathcal{B}$ such that the following conditions hold:*

(i) There exists a lower bound $f^* > -\infty$ such that for all $x \in \mathcal{B}$, $f(x, A) \ge f^*$ with probability 1.

(ii) There exist parameters L_{D_g}, L_{D_H}, B_{D_g}, and B_{D_H} such that, with high probability over the randomness in $A \sim \mathcal{G}$, letting $g(x) = f(x, A)$, we have that $g(x)$ is L_{D_g}-gradient Lipschitz and L_{D_H}-Hessian Lipschitz over $\mathcal{B}$, and $\|\nabla g(x)\| \le B_{D_g}$ and $\left\|\nabla^2 g(x)\right\|_F \le B_{D_H}$ for all $x \in \mathcal{B}$.

(iii) There exist parameters $\sigma_g, \sigma_H > 0$ such that for all $x \in \mathcal{B}$,

$$\left\|\mathrm{Cov}_{A\sim\mathcal{G}}(\nabla f(x, A))\right\|_{\mathrm{op}} \le \sigma_g^2 \text{ and } \left\|\mathrm{Cov}_{A\sim\mathcal{G}}(\mathrm{vec}(\nabla^2 f(x, A)))\right\|_{\mathrm{op}} \le \sigma_H^2.$$

Note that the radius of $\mathcal{B}$ and the parameters L_{D_g}, L_{D_H}, B_{D_g}, B_{D_H} are all allowed to depend polynomially on D and ϵ (but not on x and A).

Our main algorithmic result for ϵ-corrupted stochastic optimization is summarized in the following theorem. A formal version of this theorem is stated as Theorem 3.1 in Section 3.

Theorem 1.5 (Finding an Outlier-Robust SOSP, informal)**.** *Suppose f satisfies Assumption 1.4 in a region $\mathcal{B}$ with parameters σ_g and σ_H. Given an arbitrary initial point $x_0 \in \mathcal{B}$ and an ϵ-corrupted set of $n = \widetilde{\Omega}\left(D^2/\epsilon\right)$ functions where D is the ambient dimension, there exists a polynomial-time algorithm that with high probability outputs an $\left(O(\sigma_g\sqrt{\epsilon}), O(\sigma_H\sqrt{\epsilon})\right)$-approximate SOSP of $\bar{f}$, provided that all iterates of the algorithm stay inside $\mathcal{B}$.*

Although the bounded iterate condition in Theorem 1.5 appears restrictive, this assumption holds if the objective function satisfies a "dissipativity" property, which is a fairly general phenomenon [Hal10]. Moreover, adding an ℓ_2-regularization term enables any Lipschitz function to satisfy the dissipativity property [RRT17, Section 4]. As an illustrating example, a simple problem-specific analysis shows that this bounded iterate condition holds for outlier-robust matrix sensing by exploiting the fact that the matrix sensing objective satisfies the dissipativity property.

In this paper, we consider the problem of outlier-robust symmetric low rank matrix sensing, which we formally define below. We focus on the setting with Gaussian design.

Definition 1.6 (Outlier-Robust Matrix Sensing)**.** There is an unknown rank-r ground-truth matrix $M^* \in \mathbb{R}^{d\times d}$ that can be factored into $U^*U^{*\top}$ where $U^* \in \mathbb{R}^{d\times r}$. The (clean) sensing matrices $\{A_i\}_{i\in[n]}$ have i.i.d. standard Gaussian entries. The (clean) measurements y_i are obtained as $y_i = \langle A_i, M^*\rangle + \zeta_i$, where the noise $\zeta_i \sim \mathcal{N}(0, \sigma^2)$ is independent from all other randomness. We denote the (clean) data generation process by $(A_i, y_i) \sim \mathcal{G}_\sigma$. When $\sigma = 0$, we have $\zeta_i = 0$ and we write $\mathcal{G} := \mathcal{G}_0$ for this noiseless (measurement) setting. In outlier robust matrix sensing, an adversary can

arbitrarily change any ϵ-fraction of the sensing matrices and the corresponding measurements. This corrupted set of (A_i, y_i)'s is then given to the algorithm as input, where the goal is to recover M^*.

We highlight that in our setting, both the sensing matrices $A_i \in \mathbb{R}^{d\times d}$ and the measurements $y_i \in \mathbb{R}$ can be corrupted, presenting a substantially more challenging problem compared to prior works (e.g., [Li+20b; Li+20a]) that only allow corruption in y_i.

Let $\sigma_1^\star$ and $\sigma_r^\star$ denote the largest and the smallest nonzero singular value of M^* respectively. We assume $\sigma_r^\star$ and the rank r are given to the algorithm, and we assume that the algorithm knows a multiplicative upper bound Γ of $\sigma_1^\star$ such that $\Gamma \geq 36\sigma_1^\star$ (a standard assumption in matrix sensing even for non-robust settings [GJZ17; Jin+17]). Let $\kappa = \Gamma/\sigma_r^\star$.

Our main algorithmic result for the low rank matrix sensing problem is summarized in the following theorem. For a more detailed statement, see Theorems 3.2 and 3.3 in Section 3.

Theorem 1.7 (Our Algorithm for Outlier-Robust Matrix Sensing)**.** *Let $M^* \in \mathbb{R}^{d\times d}$ be the rank r ground-truth matrix with smallest nonzero singular value $\sigma_r^\star$. Let $\Gamma \geq 36\,\|M^*\|_{\mathrm{op}}$ and let $\kappa = \Gamma/\sigma_r^\star$. There exists an algorithm for outlier-robust matrix sensing, where an $\epsilon = O(1/(\kappa^3 r^3))$ fraction of samples from $\mathcal{G}_\sigma$ as in Definition 1.6 gets arbitrarily corrupted, that can output a rank-r matrix $\widehat{M}$ such that $\|\widehat{M} - M^*\|_F \leq \iota$ with probability at least $1-\xi$, where $\iota > 0$ is the error parameter:*

1) If $\sigma \geq r\Gamma$, then $\iota = O(\sigma\sqrt{\epsilon})$;

2) If $\sigma \leq r\Gamma$, then $\iota = O(\kappa\sigma\sqrt{\epsilon})$;

3) If $\sigma = 0$ (noiseless), then ι can be made arbitrarily small, achieving exact recovery.

The algorithm uses $n = \widetilde{O}\Big(\frac{d^2r^2 + dr\log(\Gamma/\xi)}{\epsilon}\Big)$ samples and runs in time $\mathrm{poly}(n, \kappa, \log(\sigma_r^\star/\iota))$.

Finally, we complement our algorithmic results for outlier-robust matrix sensing with a Statistical Query (SQ) lower bound, which provides strong evidence that quadratic dependence on d in the sample complexity is unavoidable for efficient algorithms. A detailed statement of this result is provided in Section 4.

1.2 Our Techniques

Outlier-robust nonconvex optimization. To obtain our algorithmic result in the general nonconvex setting, we leverage existing results on robust mean estimation [DKP20], which we use as a black box to robustly estimate the gradient and the (vectorized) Hessian. We use these robust estimates as a subroutine in a randomized nonconvex optimization algorithm (described in Appendix A.2), which can tolerate inexactness in both the gradient and the Hessian. With high probability, this algorithm outputs an (ϵ_g, ϵ_H)-approximate SOSP, where ϵ_g and ϵ_H depend on the inexactness of the gradient and Hessian oracles.

We remark that robust estimation of the Hessian is crucial to obtaining our *dimension-independent* approximation error result and is what causes D^2 dependence in the sample complexity (which is unavoidable for SQ algorithms as discussed below). Notably, the only prior work on approximating SOSPs in the outlier-robust setting [Yin+19] used robust mean estimation only on the gradients and had sample complexity scaling linearly with D; however, they can only output an order $(\sqrt{\epsilon}, (\epsilon D)^{1/5})$-SOSP, which is uninformative for many problems of interest, including the matrix sensing problem considered in this paper, due to the dimensional dependence in the approximation error.

Application to low rank matrix sensing. Our main contribution on the algorithmic side is showing that our outlier-robust nonconvex optimization framework can be applied to solve outlier-robust matrix sensing with *dimension-independent* approximation error, even achieving exact recovery when the measurements are noiseless. We obtain this result using the following geometric insights about the problem: We show that the norm of the covariance of the gradient and the Hessian can both be upper bounded by the sum of σ^2 and a function of the distance to the closest optimal solution. We further prove that all iterates stay inside a nice region using a "dissipativity" property [Hal10]), which says that the iterate aligns with the direction of the gradient when the iterate's norm is large.

This allows us to invoke Theorem 1.5 to obtain an approximate SOSP of the ground-truth objective function.

We show that this approximate SOSP must be close to a global optimal solution. Additionally, we establish a local regularity condition in a small region around globally optimal solutions (which is similar to strong convexity but holds only locally). This local regularity condition bounds below a measure of stationarity, which allows us to prove that gradient descent-type updates contract the distance to the closest global optimum under appropriate stepsize. We leverage this local regularity condition to prove that the iterates of the algorithm stay near a global optimum, so that the regularity condition continues to hold, and moreover, the distance between the current solution and the closest global optimum contracts, as long as it is larger than a function of σ. Consequently, the distance-dependent component of the gradient and Hessian covariance bound contracts as well, which allows us to obtain more accurate gradient and Hessian estimates. While such a statement may seem evident to readers familiar with linear convergence arguments, we note that proving it is quite challenging, due to the circular dependence between the distance from the current solution to global optima, the inexactness in the gradient and Hessian estimates, and the progress made by our algorithm.

The described distance-contracting argument allows us to control the covariance of the gradient and Hessian, which we utilize to recover M^* exactly when $\sigma = 0$, and recover M^* with error roughly $O(\sigma\sqrt{\epsilon})$ when $0 \neq \sigma \leq r\Gamma$. We note that the $\sigma\sqrt{\epsilon}$ appears unavoidable in the $\sigma \neq 0$ case, due to known limits of robust mean estimation algorithms [DKS17].

SQ lower bound. We exhibit a hard instance of low rank matrix sensing problem to show that quadratic dependence on the dimension in sample complexity is unavoidable for computationally efficient algorithms. Our SQ lower bound proceeds by constructing a family of distributions, corresponding to corruptions of low rank matrix sensing, that are nearly uncorrelated in a well-defined technical sense [Fel+17]. To achieve this, we follow the framework of [DKS17] which considered a family of distributions that are rotations of a carefully constructed one-dimensional distribution. The proof builds on [DKS19; Dia+21], using a new univariate moment-matching construction which yields a family of corrupted conditional distributions. These induce a family of joint distributions that are SQ-hard to learn.

1.3 Roadmap

Section 2 defines the necessary notation and discusses relevant building blocks of our algorithm and analysis. Section 3 introduces our framework for finding SOSPs in the outlier-robust setting. Section 3.1 presents how to extend and apply our framework to solve outlier-robust low rank matrix sensing. Section 4 proves that our sample complexity has optimal dimensional dependence for SQ algorithms. Most proofs are deferred to the supplementary material due to space limitations.

2 Preliminaries

For an integer n, we use $[n]$ to denote the ordered set $\{1, 2, \ldots, n\}$. We use $[a_i]_{i\in\mathcal{I}}$ to denote the matrix whose columns are vectors a_i, where $\mathcal{I}$ is an ordered set. We use $\mathbb{1}_E(x)$ to denote the indicator function that is equal to 1 if $x \in E$ and 0 otherwise. For two functions f and g, we say $f = \widetilde{O}(g)$ if $f = O(g \log^k(g))$ for some constant k, and we similarly define $\widetilde{\Omega}$.

For vectors x and y, we let $\langle x, y\rangle$ denote the inner product $x^\top y$ and $\|x\|$ denote the ℓ_2 norm of x. For $d \in \mathbb{Z}_+$, we use I_d to denote the identity matrix of size $d \times d$. For matrices A and B, we use $\|A\|_{\mathrm{op}}$ and $\|A\|_F$ to denote the spectral norm and Frobenius norm of A respectively. We use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the maximum and minimum eigenvalue of A respectively. We use $\mathrm{tr}(A)$ to denote the trace of a matrix A. We use $\langle A, B\rangle = \mathrm{tr}(A^\top B)$ to denote the entry-wise inner product of two matrices of the same dimension. We use $\mathrm{vec}(A) = [a_1^\top, a_2^\top, \ldots, a_d^\top]^\top$ to denote the canonical flattening of A into a vector, where $a_1, a_2, \ldots, a_d$ are columns of A.

Definition 2.1 (Lipschitz Continuity)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces. A function $h : \mathcal{X} \to \mathcal{Y}$ is ℓ-Lipschitz if $\|h(x_1) - h(x_2)\|_{\mathcal{Y}} \leq \ell \|x_1 - x_2\|_{\mathcal{X}}, \forall x_1, x_2$.

In this paper, when $\mathcal{Y}$ is a space of matrices, we take $\|\cdot\|_{\mathcal{Y}}$ to be the spectral norm $\|\cdot\|_{\mathrm{op}}$. When $\mathcal{X}$ is a space of matrices, we take $\|\cdot\|_{\mathcal{X}}$ to be the Frobenius norm $\|\cdot\|_F$; this essentially views the function

h as operating on the vectorized matrices endowed with the usual ℓ_2 norm. When $\mathcal{X}$ or $\mathcal{Y}$ is the Euclidean space, we take the corresponding norm to be the ℓ_2 norm.

A Randomized Algorithm with Inexact Gradients and Hessians. We now discuss how to solve the unconstrained nonconvex optimization problem $\min_{x\in\mathbb{R}^D} f(x)$, where $f(\cdot)$ is a smooth function with Lipschitz gradients and Lipschitz Hessians. The goal of this section is to find an approximate SOSP as defined in Definition 1.3.

Proposition 2.2 ([LW23]). *Suppose a function f is bounded below by $f^* > -\infty$, has L_g-Lipschitz gradient and L_H-Lipschitz Hessian, and its inexact gradient and Hessian computations $\widetilde{g}_t$ and $\widetilde{H}_t$ satisfy $\|\widetilde{g}_t - \nabla f(x_t)\| \le \frac{1}{3}\epsilon_g$ and $\|\widetilde{H}_t - \nabla^2 f(x_t)\|_{\mathrm{op}} \le \frac{2}{9}\epsilon_H$. Then there exists an algorithm (Algorithm A.1) with the following guarantees:*

1. *(Correctness) If Algorithm A.1 terminates and outputs x_n, then x_n is a $(\frac{4}{3}\epsilon_g, \frac{4}{3}\epsilon_H)$-approximate SOSP.*
2. *(Runtime) Algorithm A.1 terminates with probability 1. Let $C_\epsilon := \min\left(\frac{\epsilon_g^2}{6L_g}, \frac{2\epsilon_H^3}{9L_H^2}\right)$. With probability at least $1-\delta$, Algorithm A.1 terminates after k iterations for*

$$k = O\Big(\frac{f(x_0) - f^*}{C_\epsilon} + \frac{L_H^2 L_g^2 \epsilon_g^2}{\epsilon_H^6} \log\Big(\frac{1}{\delta}\Big)\Big)\,. \tag{1}$$

The constants $1/3$ and $2/9$ are chosen for ease of presentation. For all constructions of Hessian oracles in this paper, we take the straightforward relaxation $\|\widetilde{H}_t - \nabla^2 f(x_t)\|_{\mathrm{op}} \le \|\widetilde{H}_t - \nabla^2 f(x_t)\|_F$ and upper bound Hessian inexactness using Frobenius norm. Proof of a simplified version of Proposition 2.2 with a weaker high probability bound that is sufficient for our purposes is provided in Appendix A.2 for completeness.

Robust Mean Estimation. Recent algorithms in robust mean estimation give dimension-independent error in the presence of outliers under strong contamination model.

We use the following results, see, e.g., [DKP20], where the upper bound σ on the spectral norm of covariance matrix is unknown to the algorithm.

Proposition 2.3 (Robust Mean Estimation). *Fix any $0 < \xi < 1$. Let S be a multiset of $n = O((k\log k + \log(1/\xi))/\epsilon)$ i.i.d. samples from a distribution on $\mathbb{R}^k$ with mean μ_S and covariance Σ. Let $T \subset \mathbb{R}^k$ be an ϵ-corrupted version of S as in Definition 1.1. There exists an algorithm (Algorithm A.2) such that, with probability at least $1-\xi$, on input ϵ and T (but not $\|\Sigma\|_{\mathrm{op}}$) returns a vector $\widehat{\mu}$ in polynomial time so that $\|\mu_S - \widehat{\mu}\| = O(\sqrt{\|\Sigma\|_{\mathrm{op}}\,\epsilon})$.*

Algorithm A.2 is given in Appendix A.3. Proposition 2.3 states that for Algorithm A.2 to succeed with high probability, $\widetilde{O}(k/\epsilon)$ i.i.d. samples need to be drawn from a k-dimensional distribution of bounded covariance. State of the art algorithms for robust mean estimation can be implemented in near-linear time, requiring only a logarithmic number of passes on the data, see, e.g, [Dia+22; CDG19; DHL19]. Any of these faster algorithms could be used for our purposes.

With the above results, the remaining technical component for applying the robust estimation subroutine (Algorithm A.2) in this paper is handling the dependence across iterations.

Because we will run RobustMeanEstimation in each iteration of our optimization algorithm, the gradients $\{\nabla f_i(x_t)\}_{i=1}^n$ and Hessians $\{\nabla^2 f_i(x_t)\}_{i=1}^n$ can no longer be considered as independently drawn from a distribution after the first iteration. Although they are i.i.d. for fixed x, the dependence on previous iterations through x will break the independence assumption. Therefore, we will need a union bound over all x_t to handle dependence across different iterations t. We deal with this technicality in Appendix B.

3 General Robust Nonconvex Optimization

In this section, we establish a general result that uses Algorithm A.1 to obtain approximate SOSPs in the presence of outliers under strong contamination. The inexact gradient and inexact Hessian

oracles are constructed with the robust mean estimation subroutine (Algorithm A.2). We consider stochastic optimization tasks in Definition 1.2 satisfying Assumption 1.4. We construct the inexact gradient and Hessian oracle required by Algorithm A.1 as follows:

$$\widetilde{g}_t \leftarrow \mathbf{RobustMeanEstimation}(\{\nabla f_i(x_t)\}_{i=1}^n, 4\epsilon)$$
$$\widetilde{H}_t \leftarrow \mathbf{RobustMeanEstimation}(\{\nabla^2 f_i(x_t)\}_{i=1}^n, 4\epsilon)$$

Then we have the following guarantee:

Theorem 3.1. *Suppose we are given ϵ-corrupted set of functions $\{f_i\}_{i=1}^n$ for sample size n, generated according to Definition 1.2. Suppose Assumption 1.4 holds in a bounded region $\mathcal{B} \subset \mathbb{R}^D$ of diameter γ with gradient and Hessian covariance bound σ_g and σ_H respectively, and we have an arbitrary initialization $x_0 \in \mathcal{B}$. Algorithm A.1 initialized at x_0 outputs an (ϵ_g, ϵ_H)-approximate SOSP for a sufficiently large sample with probability at least $1-\xi$ if the following conditions hold:*

(I) All iterates x_t in Algorithm A.1 stay inside the bounded region $\mathcal{B}$.

(II) For an absolute constant $c > 0$, it holds that $\sigma_g\sqrt{\epsilon} \le c\epsilon_g$ and $\sigma_H\sqrt{\epsilon} \le c\epsilon_H$.

The algorithm uses $n = \widetilde{O}(D^2/\epsilon)$ samples, where $\widetilde{O}(\cdot)$ hides logarithmic dependence on $D, \epsilon, L_{D_g}, L_{D_H}, B_{D_g}, B_{D_H}, \gamma/\sigma_H, \gamma/\sigma_g$, and $1/\xi$. The algorithm runs in time polynomial in the above parameters.

Note that we are able to obtain dimension-independent errors ϵ_g and ϵ_H, provided that σ_g and σ_H are dimension-independent.

3.1 Low Rank Matrix Sensing Problems

In this section, we study the problem of outlier-robust low rank matrix sensing as formally defined in Definition 1.6. We first apply the above framework to obtain an approximate SOSP in Section 3.1.2. Then we make use of the approximate SOSP to obtain a solution that is close to the ground-truth matrix M^* in Section 3.1.3; this demonstrates the usefulness of approximate SOSPs.

3.1.1 Main results for Robust Low Rank Matrix Sensing

The following are the main results that we obtain in this section:

Theorem 3.2 (Main Theorem Under Noiseless Measurements)**.** *Consider the noiseless setting as in Theorem 1.7 with $\sigma = 0$. For some sample size $n = \widetilde{O}((d^2r^2 + dr\log(\Gamma/\xi))/\epsilon)$ and with probability at least $1-\xi$, there exists an algorithm that outputs a solution that is ι-close to M^* in Frobenius norm in $O(r^2\kappa^3\log(1/\xi) + \kappa\log(\sigma_r^\star/\iota))$ calls to the robust mean estimation subroutine (Algorithm A.2).*

This result achieves exact recovery of M^*, despite the strong contamination of samples. Each iteration involves a subroutine call to robust mean estimation. Algorithm A.2 presented here is one simple example of robust mean estimation; there are refinements [Dia+22; DHL19] that run in nearly linear time, so the total computation utilizing those more efficient algorithms indeed requires $\widetilde{O}\left(r^2\kappa^3\right)$ passes of data (computed gradients and Hessians).

Theorem 3.3 (Main Theorem Under Noisy Measurements)**.** *Consider the same setting as in Theorem 1.7 with $\sigma \neq 0$. There exists a sample size $n = \widetilde{O}((d^2r^2 + dr\log(\Gamma/\xi))/\epsilon)$ such that*

- *if $\sigma \le r\Gamma$, then with probability at least $1-\xi$, there exists an algorithm that outputs a solution $\widehat{M}$ in $\widetilde{O}(r^2\kappa^3)$ calls to robust mean estimation routine A.2, with error $\|\widehat{M} - M^*\|_F = O(\kappa\sigma\sqrt{\epsilon})$;*
- *if $\sigma \ge r\Gamma$, then with probability at least $1-\xi$, there exists a (different) algorithm that outputs a solution $\widehat{M}$ in one call to robust mean estimation routine A.2, with error $\|\widehat{M} - M^*\|_F = O(\sigma\sqrt{\epsilon})$.*

We prove Theorem 3.3 in Appendix C.4, and instead focus on the noiseless measurements with $\sigma = 0$ when we develop our algorithms in this section; the two share many common techniques. In the remaining part of Section 3.1, we use $\mathcal{G}_0$ in Definition 1.6 for the data generation process.

We now describe how we obtain the solution via nonconvex optimization. Consider the following objective function for (uncorrupted) matrix sensing:

$$\min_{\substack{M\in\mathbb{R}^{d\times d}\\ \mathrm{rank}(M)=r}} \frac{1}{2}\mathop{\mathbb{E}}_{(A_i,y_i)\sim\mathcal{G}_0}(\langle M, A_i\rangle - y_i)^2. \tag{2}$$

We can write $M = UU^\top$ for some $U \in \mathbb{R}^{d\times r}$ to reparameterize the objective function. Let

$$f_i(U) := \frac{1}{2}\left(\langle UU^\top, A_i\rangle - y_i\right)^2. \tag{3}$$

We can compute

$$\bar{f}(U) := \mathop{\mathbb{E}}_{(A_i,y_i)\sim\mathcal{G}_0} f_i(U) = \frac{1}{2}\mathrm{Var}\langle UU^\top - M^*, A_i\rangle = \frac{1}{2}\left\|UU^\top - M^*\right\|_F^2. \tag{4}$$

We seek to solve the following optimization problem under the corruption model in Definition 1.2:

$$\min_{U\in\mathbb{R}^{d\times r}} \bar{f}(U). \tag{5}$$

The gradient Lipschitz constant and Hessian Lipschitz constant of $\bar{f}$ are given by the following result.

Fact 3.4 ([Jin+17], Lemma 6)**.** *For any $\Gamma > \sigma_1^\star$, $\bar{f}(U)$ has gradient Lipschitz constant $L_g = 16\Gamma$ and Hessian Lipschitz constant $L_H = 24\Gamma^{\frac{1}{2}}$ inside the region $\{U : \|U\|_{\mathrm{op}}^2 < \Gamma\}$.*

3.1.2 Global Convergence to an Approximate SOSP

In this section, we apply our framework Theorem 3.1 to obtain global convergence from an arbitrary initialization to an approximate SOSP, by providing problem-specific analysis to guarantee that both Assumption 1.4 and algorithmic assumptions (I) and (II) required by Theorem 3.1 are satisfied.

Theorem 3.5 (Global Convergence to a SOSP)**.** *Consider the noiseless setting as in Theorem 1.7 with $\sigma = 0$ and $\epsilon = O(1/(\kappa^3 r^3))$. Assume we have an arbitrary initialization U_0 inside $\{U : \|U\|_{\mathrm{op}}^2 \le \Gamma\}$. There exists a sample size $n = \widetilde{O}\left((d^2r^2 + dr\log(\Gamma/\xi))/\epsilon\right)$ such that with probability at least $1-\xi$, Algorithm A.1 initialized at U_0 outputs a $(\frac{1}{24}\sigma_r^{\star 3/2}, \frac{1}{3}\sigma_r^\star)$-approximate SOSP using at most $O\left(r^2\kappa^3\log(1/\xi)\right)$ calls to robust mean estimation subroutine (Algorithm A.2).*

Proof of Theorem 3.5. To apply Theorem 3.1, we verify Assumption 1.4 first. To verify (i), for all U and A_i, $f_i(U) = \frac{1}{2}\left(\langle UU^\top, A_i\rangle - y_i\right)^2 \ge 0$, so $f^* = 0$ is a uniform lower bound. We verify (ii) in Appendix C.2: conceptually, by Fact 3.4, $\bar{f}$ is gradient and Hessian Lipschitz; both gradient and Hessian of f_i are sub-exponential and concentrate around those of $\bar{f}$. To check (iii), we calculate the gradients and Hessians of f_i in Appendix C.1.1 and bound their covariances from above in Appendix C.1.2 and C.1.3. The result is summarized in the following lemma. Note that the domain of the target function in Algorithm A.1 and Theorem 3.1 is the Euclidean space $\mathbb{R}^D$, so we vectorize U and let $D = dr$. The gradient becomes a vector in $\mathbb{R}^{dr}$ and the Hessian becomes a matrix in $\mathbb{R}^{dr\times dr}$.

Lemma 3.6 (Gradient and Hessian Covariance Bounds)**.** *For all $U \in \mathbb{R}^{d\times r}$ with $\|U\|_{\mathrm{op}}^2 \le \Gamma$ and f_i defined in Equation (3), it holds*

$$\left\|\mathrm{Cov}(\mathrm{vec}(\nabla f_i(U)))\right\|_{\mathrm{op}} \le 8\left\|UU^\top - M^*\right\|_F^2\|U\|_{\mathrm{op}}^2 \le 32r^2\Gamma^3 \tag{6}$$

$$\left\|\mathrm{Cov}(\mathrm{vec}(H_i))\right\|_{\mathrm{op}} \le 16r\left\|UU^\top - M^*\right\|_F^2 + 128\|U\|_{\mathrm{op}}^4 \le 192r^3\Gamma^2 \tag{7}$$

We proceed to verify the algorithmic assumptions in Theorem 3.1. For the assumption (I), we prove the following Lemma in Appendix C.2 to show that all iterates stay inside the bounded region in which we compute the covariance bounds.

Lemma 3.7. *All iterates of Algorithm A.1 stay inside the region $\{U : \|U\|_{\mathrm{op}}^2 \le \Gamma\}$.*

To verify Theorem 3.1 (II), we let $\epsilon_g = \frac{1}{32}\sigma_r^{\star 3/2}, \epsilon_H = \frac{1}{4}\sigma_r^\star$ and $\sigma_g = 8r\Gamma^{1.5}, \sigma_H = 16r^{1.5}\Gamma$. So if we assume $\epsilon = O(1/(\kappa^3 r^3))$, then for the absolute constant c in Theorem 3.1 it holds that

$$\sigma_g\sqrt{\epsilon} \le c\epsilon_g \qquad \sigma_H\sqrt{\epsilon} \le c\epsilon_H.$$

Hence, Theorem 3.1 applies and Algorithm A.1 outputs an (ϵ_g, ϵ_H)-approximate SOSP with high probability in polynomial time. To bound the runtime, since $\bar{f}(U_0) = 1/2\left\|U_0U_0^\top - M^*\right\|_F^2 = O(r^2\Gamma^2)$ for an arbitrary initialization U_0 with $\|U_0\|_{\text{op}}^2 < \Gamma$, the initial distance can be bounded by $O(r^2\Gamma^2)$. Setting $L_g = 16\Gamma, L_H = 24\Gamma^{1/2}, \bar{f}(U_0) = O(r^2\Gamma^2), f^* = 0$ and thus $C_\epsilon = O(\sigma_r^{\star 3}/\Gamma)$, Proposition 2.2 implies that Algorithm A.1 outputs a $(\frac{1}{24}\sigma_r^{\star 3/2}, \frac{1}{3}\sigma_r^\star)$-approximate second order stationary point U_{SOSP} in $O(r^2\kappa^3\log(1/\xi))$ steps with high probability. □

3.1.3 Local Linear Convergence

In this section, we describe a local search algorithm that takes a $(\frac{1}{24}\sigma_r^{\star 3/2}, \frac{1}{3}\sigma_r^\star)$-approximate second-order stationary point as its initialization and achieves exact recovery even in the presence of outliers.

Algorithm 3.1: Local Inexact Gradient Descent

Data: The initialization U_{SOSP} is a $(\frac{1}{24}\sigma_r^{\star 3/2}, \frac{1}{3}\sigma_r^\star)$-approximate SOSP, corruption fraction is ϵ, corrupted samples are $\{(A_i, y_i)\}_{i=1}^n$, target distance to optima is ι

Result: U that is ι-close in Frobenius norm to some global minimum

1 $\eta = 1/\Gamma, U_0 = U_{SOSP}$

2 **for** *t = 0, 1, …* **do**

3 $\quad \widetilde{g}_t := \mathbf{RobustMeanEstimation}(\{\nabla f_i(U_t)\}_{i=1}^n, 4\epsilon)$

4 $\quad U_{t+1} \leftarrow U_t - \eta\widetilde{g}_t$

Theorem 3.8 (Local Linear Convergence)**.** *Consider the same noiseless setting as in Theorem 1.7. Assume we already found a $(\frac{1}{24}\sigma_r^{\star 3/2}, \frac{1}{3}\sigma_r^\star)$-approximate SOSP U_{SOSP} of $\bar{f}$. Then there exists a sample size $n = \widetilde{O}(dr\log(1/\xi)/\epsilon)$ such that with probability at least $1-\xi$, Algorithm 3.1 initialized at U_{SOSP} outputs a solution that is ι-close to some global minimum in Frobenius norm after $O(\kappa\log(\sigma_r^\star/\iota))$ calls to robust mean estimation subroutine (Algorithm A.2). Moreover, all iterates U_t are $\frac{1}{3}\sigma_r^{\star 1/2}$-close to some global minimum in Frobenius norm.*

Proof Sketch. First we use known properties of $\bar{f} = \mathbb{E}_{(A_i,y_i)\sim\mathcal{G}_0} f_i$ from the literature [Jin+17] to show approximate SOSPs of $\bar{f}$ — in particular our initialization U_{SOSP} — are in a small neighborhood of the global minima of $\bar{f}$. In that neighborhood, it was also known that $\bar{f}$ satisfies some local regularity conditions that enable gradient descent's linear convergence.

However, the algorithm only has access to the *inexact* gradient from the robust mean estimation subroutine (line 3 in Algorithm 3.1) and, therefore, we need to establish linear convergence of *inexact gradient descent*. We achieve this with the following iterative argument: As the iterate gets closer to global optima, the covariance bound of sample gradient in Equation (6) gets closer to 0. Because the accuracy of the robust mean estimation scales with the covariance bound (see Proposition 2.3), a more accurate estimate of population gradient $\nabla\bar{f}$ can be obtained via the robust estimation subroutine (Algorithm A.2). This, in turn, allows for an improved inexact gradient descent step, driving the algorithm towards an iterate that is even closer to global optima.

See Appendix C.3 for the complete proof. □

4 Statistical Query Lower Bound for Low Rank Matrix Sensing

Our general algorithm (Theorem 3.1) leads to an efficient algorithm for robust low rank matrix sensing with sample complexity $O(d^2r^2/\epsilon)$ (Theorem 3.3). Interestingly, the sample complexity of the underlying robust estimation problem — ignoring computational considerations — is $\Theta(dr/\epsilon)$. The information-theoretic upper bound of $O(dr/\epsilon)$ can be achieved by an exponential (in the dimension)

time algorithm (generalizing the Tukey median to our regression setting); see, e.g., Theorem 3.5 in [Gao20].

Given this discrepancy, it is natural to ask whether the sample complexity achieved by our algorithm can be improved via a different computationally efficient method. In this section, we provide evidence that this may not be possible. In more detail, we establish a near-optimal information-computation tradeoff for the problem, within the class of Statistical Query (SQ) algorithms. To formally state our lower bound, we require basic background on SQ algorithms.

Basics on SQ Model. SQ algorithms are a class of algorithms that, instead of access to samples from some distribution $\mathcal{P}$, are allowed to query expectations of bounded functions over $\mathcal{P}$.

Definition 4.1 (SQ Algorithms and STAT Oracle [Kea98])**.** Let $\mathcal{P}$ be a distribution on $\mathbb{R}^{d^2+1}$. A Statistical Query (SQ) is a bounded function $q : \mathbb{R}^{d^2+1} \to [-1, 1]$. For $\tau > 0$, the $\mathsf{STAT}(\tau)$ oracle responds to the query q with a value v such that $|v - \mathbb{E}_{X\sim\mathcal{P}}[q(X)]| \leq \tau$. An SQ algorithm is an algorithm whose objective is to learn some information about an unknown distribution $\mathcal{P}$ by making adaptive calls to the corresponding $\mathsf{STAT}(\tau)$ oracle.

In this section, we consider $\mathcal{P}$ as the unknown corrupted distribution where (A_i, y_i) are drawn. The SQ algorithm tries to learn the ground truth matrix M^* from this corrupted distribution; the goal of the lower bound result is to show that this is hard.

The SQ model has the capability to implement a diverse set of algorithmic techniques in machine learning such as spectral techniques, moment and tensor methods, local search (e.g., Expectation Maximization), and several others [Fel+17]. A lower bound on the SQ complexity of a problem provides evidence of hardness for the problem. [Bre+21] established that (under certain assumptions) an SQ lower bound also implies a qualitatively similar lower bound in the low-degree polynomial testing model. This connection can be used to show a similar lower bound for low-degree polynomials.

Our main result here is a near-optimal SQ lower bound for robust low rank matrix sensing that applies even for rank $r = 1$, i.e., when the ground truth matrix is $M^* = uu^\top$ for some $u \in \mathbb{R}^d$. The choice of rank $r = 1$ yields the strongest possible lower bound in our setting because it is the easiest parameter regime: Recall that the sample complexity of our algorithm is $\widetilde{O}(d^2r^2)$ as in Theorems 3.2 and 3.3, and the main message of our SQ lower bound is to provide evidence that the d^2 factor is necessary for computationally efficient algorithms *even if* $r = 1$.

Theorem 4.2 (SQ Lower Bound for Robust Rank-One Matrix Sensing)**.** *Let $\epsilon \in (0, 1/2)$ be the fraction of corruptions and let $c \in (0, 1/2)$. Assume the dimension $d \in \mathbb{N}$ is sufficiently large. Consider the ϵ-corrupted rank-one matrix sensing problem with ground-truth matrix $M^* = uu^\top$ and noise $\sigma^2 = O(1)$. Any SQ algorithm that outputs $\widehat{u}$ with $\|\widehat{u} - u\| = O(\epsilon^{1/4})$ either requires $2^{\Omega(d^c)}/d^{2-4c}$ queries or makes at least one query to $\mathsf{STAT}\Big(e^{O\left(1/\sqrt{\epsilon}\right)}/O\big(d^{1-2c}\big)\Big)$.*

In other words, we show that, when provided with SQ access to an ϵ-corrupted distribution, approximating u is impossible unless employing a statistical query of higher precision than what can be achieved with a strictly sub-quadratic number (e.g., $d^{1.99}$) of samples. Note that the SQ oracle $\mathsf{STAT}(e^{O\left(1/\sqrt{\epsilon}\right)}/O(d^{1-2c}))$ can be simulated with $O(d^{2-4c})/e^{O(1/\epsilon)}$ samples, and this bound is tight in general. Informally speaking, this theorem implies that improving the sample complexity from d^2 to d^{2-4c} requires exponentially many queries. This result can be viewed as a near-optimal information-computation tradeoff for the problem, within the class of SQ algorithms.

The proof follows a similar analysis as in [DKS19; Dia+21], using one-dimensional moment matching to construct a family of corrupted conditional distributions, which induce a family of corrupted joint distributions that are SQ-hard to learn. We provide the details of the proof in Appendix D. Apart from the formal proof, in Appendix E we also informally discuss the intuition for why some simple algorithms that require $O(d)$ samples do not provide dimension-independent error guarantees.

Acknowledgements

Shuyao Li was supported in part by NSF Awards DMS-2023239, NSF Award CCF-2007757 and the U. S. Office of Naval Research under award number N00014-22-1-2348. Yu Cheng was supported in

part by NSF Award CCF-2307106. Ilias Diakonikolas was supported in part by NSF Medium Award CCF-2107079, NSF Award CCF-1652862 (CAREER), a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant. Jelena Diakonikolas was supported in part by NSF Award CCF-2007757 and by the U. S. Office of Naval Research under award number N00014-22-1-2348. Rong Ge was supported in part by NSF Award DMS-2031849, CCF-1845171 (CAREER) and a Sloan Research Fellowship. Stephen Wright was supported in part by NSF Awards DMS-2023239 and CCF-2224213 and AFOSR via subcontract UTA20-001224 from UT-Austin.

References

[Bar+10] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar. "The security of machine learning". In: *Machine Learning* 81.2 (2010), pp. 121–148.

[BBV16] A. S. Bandeira, N. Boumal, and V. Voroninski. "On the low-rank approach for semidefinite programs arising in synchronization and community detection". In: *Conference on learning theory*. PMLR. 2016, pp. 361–382.

[Ber+22] E. H. Bergou, Y. Diouane, V. Kunc, V. Kungurtsev, and C. W. Royer. "A subsampling line-search method with second-order results". In: *INFORMS Journal on Optimization* 4.4 (2022), pp. 403–425.

[BNL12] B. Biggio, B. Nelson, and P. Laskov. "Poisoning Attacks against Support Vector Machines". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress, 2012, pp. 1467–1474.

[BNS16] S. Bhojanapalli, B. Neyshabur, and N. Srebro. "Global optimality of local search for low rank matrix recovery". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3873–3881.

[Bre+21] M. S. Brennan, G. Bresler, S. Hopkins, J. Li, and T. Schramm. "Statistical Query Algorithms and Low Degree Tests Are Almost Equivalent". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 774–774.

[CDG19] Y. Cheng, I. Diakonikolas, and R. Ge. "High-Dimensional Robust Mean Estimation in Nearly-Linear Time". In: *Proceedings of the 30th ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2019, pp. 2755–2771.

[CG18] Y. Cheng and R. Ge. "Non-convex matrix completion against a semi-random adversary". In: *Conference On Learning Theory*. PMLR. 2018, pp. 1362–1394.

[DHL19] Y. Dong, S. Hopkins, and J. Li. "Quantum entropy scoring for fast robust mean estimation and improved outlier detection". In: *Advances in Neural Information Processing Systems* 32 (2019).

[Dia+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: *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*. IEEE Computer Soc., Los Alamitos, CA, 2016, pp. 655–664.

[Dia+17] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. "Being Robust (in High Dimensions) Can Be Practical". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 999–1008.

[Dia+19] I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and A. Stewart. "Sever: A Robust Meta-Algorithm for Stochastic Optimization". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 1596–1606.

[Dia+21] I. Diakonikolas, D. Kane, A. Pensia, T. Pittas, and A. Stewart. "Statistical Query Lower Bounds for List-Decodable Linear Regression". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 3191–3204.

[Dia+22] I. Diakonikolas, D. M. Kane, A. Pensia, and T. Pittas. "Streaming Algorithms for High-Dimensional Robust Statistics". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 5061–5117.

[DK19] I. Diakonikolas and D. M. Kane. "Recent advances in algorithmic high-dimensional robust statistics". In: *arXiv preprint arXiv:1911.05911* (2019).

[DK23] I. Diakonikolas and D. M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023.

[DKP20] I. Diakonikolas, D. M. Kane, and A. Pensia. "Outlier Robust Mean Estimation with Subgaussian Rates via Stability". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1830–1840.

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. "Statistical Query Lower Bounds for Robust Estimation of High-Dimensional Gaussians and Gaussian Mixtures". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. 2017, pp. 73–84.

[DKS19] I. Diakonikolas, W. Kong, and A. Stewart. "Efficient algorithms and lower bounds for robust linear regression". In: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, Philadelphia, PA, 2019, pp. 2745–2754.

[Dur19] R. Durrett. *Probability—theory and examples*. Vol. 49. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019.

[Fel+17] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. "Statistical Algorithms and a Lower Bound for Detecting Planted Cliques". In: *J. ACM* 64.2 (2017), 8:1–8:37.

[Gao20] C. Gao. "Robust regression via mutivariate regression depth". In: *Bernoulli* 26.2 (2020), pp. 1139–1170.

[GC23] X. Gao and Y. Cheng. "Robust Matrix Sensing in the Semi-Random Model". In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)* (2023).

[Ge+15] R. Ge, F. Huang, C. Jin, and Y. Yuan. "Escaping from saddle points—online stochastic gradient for tensor decomposition". In: *Conference on Learning Theory*. 2015, pp. 797–842.

[GJZ17] R. Ge, C. Jin, and Y. Zheng. "No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 1233–1242.

[GLM16] R. Ge, J. D. Lee, and T. Ma. "Matrix completion has no spurious local minimum". In: *Advances in Neural Information Processing Systems*. 2016, pp. 2973–2981.

[Hal10] J. K. Hale. *Asymptotic behavior of dissipative systems*. 25. American Mathematical Soc., 2010.

[Ham+86] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics. The approach based on influence functions*. Wiley New York, 1986.

[HR09] P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley New York, 2009.

[Hub64] P. J. Huber. "Robust estimation of a location parameter". In: *Annals of Mathematical Statistics* 35 (1964), pp. 73–101.

[IPL23] E. Ioannou, M. S. Pydi, and P.-L. Loh. "Robust empirical risk minimization via Newton's method". In: *arXiv preprint arXiv:2301.13192* (2023).

[Jin+17] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. "How to Escape Saddle Points Efficiently". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 1724–1732.

[Jin+21] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. "On nonconvex optimization for machine learning: gradients, stochasticity, and saddle points". In: *Journal of the ACM* 68.2 (2021), Art. 11, 29.

[Kea98] M. J. Kearns. "Efficient noise-tolerant Learning from Statistical Queries". In: *Journal of the ACM* 45.6 (1998), pp. 983–1006.

[Li+08] J. Li, D. Absher, H. Tang, A. Southwick, A. Casto, S. Ramachandran, H. Cann, G. Barsh, M. Feldman, L. Cavalli-Sforza, and R. Myers. "Worldwide human relationships inferred from genome-wide patterns of variation". In: *Science* 319 (2008), pp. 1100–1104.

[Li+20a] X. Li, Z. Zhu, A. Man-Cho So, and R. Vidal. "Nonconvex robust low-rank matrix recovery". In: *SIAM Journal on Optimization* 30.1 (2020), pp. 660–686.

[Li+20b] Y. Li, Y. Chi, H. Zhang, and Y. Liang. "Non-convex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent". In: *Information and Inference: A Journal of the IMA* 9.2 (2020), pp. 289–325.

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala. "Agnostic Estimation of Mean and Covariance". In: *focs2016*. 2016, pp. 665–674.

[LW23] S. Li and S. J. Wright. "A randomized algorithm for nonconvex minimization with inexact evaluations and complexity guarantees". In: *arXiv preprint arXiv:2310.18841* (2023).

[Pas+10] P. Paschou, J. Lewis, A. Javed, and P. Drineas. "Ancestry Informative Markers for Fine-Scale Individual Assignment to Worldwide Populations". In: *Journal of Medical Genetics* 47 (2010), pp. 835–847.

[Pra+20] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar. "Robust estimation via robust gradient estimation". In: *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 82.3 (2020), pp. 601–627.

[Ros+02] N. Rosenberg, J. Pritchard, J. Weber, H. Cann, K. Kidd, L. Zhivotovsky, and M. Feldman. "Genetic structure of human populations". In: *Science* 298 (2002), pp. 2381–2385.

[RRT17] M. Raginsky, A. Rakhlin, and M. Telgarsky. "Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by S. Kale and O. Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, July 2017, pp. 1674–1703.

[SKL17] J. Steinhardt, P. W. Koh, and P. S. Liang. "Certified Defenses for Data Poisoning Attacks". In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 3520–3532.

[SQW16] J. Sun, Q. Qu, and J. Wright. "A geometric analysis of phase retrieval". In: *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE. 2016, pp. 2379–2383.

[SQW17] J. Sun, Q. Qu, and J. Wright. "Complete Dictionary Recovery Over the Sphere I: Overview and the Geometric Picture". In: *IEEE Trans. Inf. Theor.* 63.2 (Feb. 2017), pp. 853–884.

[Tuk75] J. Tukey. "Mathematics and picturing of data". In: *Proceedings of ICM*. Vol. 6. 1975, pp. 523–531.

[WM22] J. Wright and Y. Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

[Yin+19] D. Yin, Y. Chen, R. Kannan, and P. Bartlett. "Defending Against Saddle Point Attack in Byzantine-Robust Distributed Learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 7074–7084.

Supplementary Material

Supplementary material is organized as follows. In Appendix A, we provide useful auxiliary facts and relevant technical results from previous works. Appendix B proves our result for general robust nonconvex optimization (Theorem 3.1). Appendix C provides omitted computation and proofs for robust low rank matrix sensing (Section 3.1). Appendix D proves our SQ lower bound (Section 4) for the sample complexity of efficient algorithms for the outlier-robust low rank matrix sensing problem, and Appendix E discusses the intuition why some simple algorithms that violate our SQ lower bound fail.

A Technical Preliminaries

A.1 Notation and Auxiliary Facts

In the appendix, we use $\widetilde{O}(\cdot)$ notation to suppress, for conciseness, logarithmic dependences on all defined quantities, even if they do not appear inside $\widetilde{O}(\cdot)$. More precise statements can be found in the corresponding main parts of the paper. For matrix space $\mathbb{R}^{d\times r}$ and a set $\mathcal{X}^\star \subset \mathbb{R}^{d\times r}$, we use $\mathcal{P}_{\mathcal{X}^\star}(\cdot)$ to denote the Frobenius projection onto $\mathcal{X}^\star$, i.e., $\mathcal{P}_{\mathcal{X}^\star}(U) = \arg\min_{Z\in\mathcal{X}^\star} \|U - Z\|_F^2$. We use $\mathrm{dist}(U, \mathcal{X}^\star)$ to denote $\|U - \mathcal{P}_{\mathcal{X}^\star}(U)\|_F$.

Fact A.1. *For matrices A, B with compatible dimensions,*

$$\|AB\|_F \le \|A\|_{\mathrm{op}} \|B\|_F$$
$$\|AB\|_F \le \|A\|_F \|B\|_{\mathrm{op}}$$

For two matrices A and B, let $A \otimes B$ denote the Kronecker product of A and B.

Fact A.2. *For matrices A, B, C with compatible dimensions,* $\mathrm{vec}(ABC) = \left(C^\top \otimes A\right)\mathrm{vec}(B)$.

Fact A.3. $\|A \otimes B\|_{\mathrm{op}} = \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}}$

We will frequently use the following fact about the mean and the variance of the quadratic form for zero-mean multivariate normal distributions.

Fact A.4. *Let $X \sim \mathcal{N}(0, \Sigma)$ be a k-dimensional random variable and let G be a $k \times k$ real matrix; then*

$$\mathbb{E}[X^\top G X] = \mathrm{tr}(G\Sigma) = \langle G, \Sigma\rangle$$
$$\mathrm{Var}[X^\top G X] = \mathrm{tr}(G\Sigma(G + G^\top)\Sigma)$$

A.2 Simplified Proof for the Optimization Algorithm with Inexact Gradients and Hessians

Algorithm A.1: Nonconvex minimization with inexact gradients and Hessians [LW23]

for $t = 1, 2, \ldots$ **do**
  Obtain $\widetilde{g}_t$ from the inexact gradient oracle
  if $\|\widetilde{g}_t\| > \epsilon_g$ **then**
    $x_{t+1} = x_t - \frac{1}{L}\widetilde{g}_t$
  else
    Obtain $\widetilde{H}_t$ from the inexact Hessian oracle
    Compute smallest eigenvalue and its corresponding eigenvector $(\widetilde{\lambda}_t, \widetilde{p}_t)$
    if $\widetilde{\lambda}_t < -\epsilon_H$ **then**
      $\sigma_t = \pm 1$ with probability $\frac{1}{2}$
      $x_{t+1} = x_t + \frac{2\epsilon_H}{L_H}\sigma_t \widetilde{p}_t$
    else
      return x_t

We stated the following guarantee (Proposition 2.2) in Section 2:

Proposition A.5. *Suppose a function f is bounded below by $f^* > -\infty$, has L_g-Lipschitz gradient and L_H-Lipschitz Hessian, and its inexact gradient and Hessian computation $\widetilde{g}_t$ and $\widetilde{H}_t$ satisfy $\|\widetilde{g}_t - \nabla f(x_t)\| \le \frac{1}{3}\epsilon_g$ and $\|\widetilde{H}_t - \nabla^2 f(x_t)\|_{\mathrm{op}} \le \frac{2}{9}\epsilon_H$. Then there exists an algorithm (Algorithm A.1) with the following guarantees:*

1. *(Correctness) If Algorithm A.1 terminates and outputs x_n, then x_n is a $(\frac{4}{3}\epsilon_g, \frac{4}{3}\epsilon_H)$-approximate second-order stationary point.*

2. *(Runtime) Algorithm A.1 terminates with probability 1. Let $C_\epsilon := \min\left(\frac{\epsilon_g^2}{6L_g}, \frac{2\epsilon_H^3}{9L_H^2}\right)$. With probability at least $1-\delta$, Algorithm A.1 terminates after k iterations for*

$$k = O\Big(\frac{f(x_0) - f^*}{C_\epsilon} + \frac{L_H^2 L_g^2 \epsilon_g^2}{\epsilon_H^6} \log\Big(\frac{1}{\delta}\Big)\Big)\,.$$

This section proves the correctness and a slightly weaker runtime guarantee in terms of the dependence on δ. We prove a $O(1/\delta)$ dependence instead of $O(\log(1/\delta))$, i.e., with probability at least $1-\delta$, Algorithm A.1 terminates after k iterations for

$$k = O\Big(\frac{f(x_0) - f^*}{\delta C_\epsilon}\Big)\,. \tag{8}$$

Proof of correctness and a weaker runtime [LW23]. We prove the correctness first. From the stopping criteria, we have $\|\widetilde{g}_k\| \le \epsilon_g$ and $\widetilde{\lambda}_k \ge -\epsilon_H$. Thus,

$$\|\nabla f(x_k)\| \le \|g_k\| + \frac{1}{3}\epsilon_g = \frac{4}{3}\epsilon_g.$$

To bound $\lambda_{\min}(\nabla^2 f(x_k))$, write $U_k = \widetilde{H}_k - \nabla^2 f(x_k)$. By Hessian inexactness condition, $\lambda_{\min}(-U_k) \ge -\|-U_k\|_{\mathrm{op}} \ge -\frac{2}{9}\epsilon_H$.

We use Weyl's theorem to conclude that

$$\lambda_{\min}(\nabla^2 f(x_k)) \ge \lambda_{\min}(\widetilde{H}_k) + \lambda_{\min}(-U_k) \ge -\epsilon_H - \frac{2}{9}\epsilon_H > -\frac{4}{3}\epsilon_H,$$

as required.

Now we analyze the runtime. We proceed by first establishing an expectation bound and then use Markov's inequality.

Write $v_t := \widetilde{g}_t - \nabla f(x_t)$. When the algorithm takes a gradient step, we have $\|\widetilde{g}_t\| > \epsilon_g$. By gradient inexactness condition, we have $\|v_t\| \le \frac{1}{3}\|\widetilde{g}_t\|$, so that $\|\nabla f(x_t)\| \ge \frac{2}{3}\|\widetilde{g}_t\|$. From Taylor's Theorem, we have

$$\begin{aligned}
f(x_{t+1}) &= f\left(x_t - \frac{1}{L_g}\widetilde{g}_t\right) \\
&\le f(x_t) - \frac{1}{L_g}\nabla f(x_t)^\top \widetilde{g}_t + \frac{L_g}{2}\cdot\frac{1}{L_g^2}\|\widetilde{g}_t\|^2 \\
&= f(x_t) - \frac{1}{2L_g}\left(\|\nabla f(x_t)\|^2 - \|v_t\|^2\right) \\
&\le f(x_t) - \frac{1}{2L_g}\left(\frac{4}{9}\|\widetilde{g}_t\|^2 - \frac{1}{9}\|\widetilde{g}_t\|^2\right) \\
&\le f(x_t) - \frac{1}{6L_g}\|\widetilde{g}_t\|^2 \\
&\le f(x_t) - \frac{1}{6L_g}\epsilon_g^2,
\end{aligned} \tag{9}$$

For the negative curvature step, recall that $\|\widetilde{H}_t - \nabla^2 f(x_t)\|_{\text{op}} \leq \frac{2}{9}\epsilon_H$. It follows from Taylor's theorem that

$$\begin{aligned} f(x_{t+1}) &= f(x_t + \frac{2\epsilon_H}{L_H}\sigma_t\widetilde{p}_t) \\ &\leq f(x_t) + 2\frac{\epsilon_H}{L_H}\nabla f(x_t)^\top \sigma_t\widetilde{p}_t + \frac{1}{2}\frac{4\epsilon_H^2}{L_H^2}\widetilde{p}_t^\top \nabla^2 f(x_t)\widetilde{p}_t + \frac{L_H}{6}\frac{8\epsilon_H^3}{L_H^3} \\ &= f(x_t) + \frac{2\epsilon_H^2}{L_H^2}\left(\widetilde{p}_t^\top \nabla^2 f(x_t)\widetilde{p}_t + \frac{2}{3}\epsilon_H\right) + 2\frac{\epsilon_H}{L_H}\nabla f(x_t)^\top \sigma_t\widetilde{p}_t \end{aligned}$$

When the algorithm takes a negative curvature step, we have $\widetilde{\lambda}_t < -\epsilon_H < 0$, so by Hessian inexactness condition, we have $\|\widetilde{H}_t - \nabla^2 f(x_t)\| \leq \frac{2}{9}\epsilon_H \leq \frac{2}{9}|\widetilde{\lambda}_t|$. It follows from the definition of operator norm and Cauchy-Schwarz that $|\widetilde{p}_t^\top \widetilde{H}_t\widetilde{p}_t - \widetilde{p}_t^\top \nabla^2 f(x_t)\widetilde{p}_t| \leq \|\widetilde{H}_t\widetilde{p}_t - \nabla^2 f(x_t)\widetilde{p}_t\| \leq \frac{2|\widetilde{\lambda}_t|}{9}$, so

$$\widetilde{p}_t^\top \nabla^2 f(x_t)\widetilde{p}_t \leq \frac{2}{9}|\widetilde{\lambda}_t| + p_t^\top \widetilde{H}_t\widetilde{p}_t = \frac{2}{9}|\widetilde{\lambda}_t| + (-|\widetilde{\lambda}_t|) = -\frac{7}{9}|\widetilde{\lambda}_t|.$$

We thus have

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \frac{2\epsilon_H^2}{L_H^2}\left(\widetilde{p}_t^\top \nabla^2 f(x_t)\widetilde{p}_t + \frac{2}{3}\epsilon_H\right) + 2\frac{\epsilon_H}{L_H}\nabla f(x_t)^\top \sigma_t\widetilde{p}_t \\ &\leq f(x_t) + \frac{2\epsilon_H^2}{L_H^2}\left(-\frac{7}{9}|\widetilde{\lambda}_t| + \frac{2}{3}\alpha_t\right) + 2\frac{\epsilon_H}{L_H}\nabla f(x_t)^\top \sigma_t\widetilde{p}_t \\ &\leq f(x_t) - \frac{2\epsilon_H^3}{9L_H^2} + 2\frac{\epsilon_H}{L_H}\nabla f(x_t)^\top \sigma_t\widetilde{p}_t. \end{aligned} \tag{10}$$

We have the following result for expected stopping time of Algorithm A.1. Here the expectation is taken with respect to the random variables σ_t used at the negative curvature iterations. For purposes of this and later results, we define

$$C_\epsilon := \min\left(\frac{\epsilon_g^2}{6L_g}, \frac{2\epsilon_H^3}{9L_H^2}\right). \tag{11}$$

Assuming Lemma A.6, the runtime guarantee in Proposition A.5 follows from Markov inequality. □

Lemma A.6. *Consider the same setting as in Proposition A.5. Let T denote the iteration at which Algorithm A.1 terminates. Then $T < \infty$ almost surely and*

$$\mathbb{E}\, T \leq \frac{f(x_0) - f^*}{C_\epsilon}, \tag{12}$$

where C_ϵ is defined in (11).

The proof of this result is given below. It constructs a supermartingale[1] based on the function value and uses a supermartingale convergence theorem and optional stopping theorem to obtain the final result. A similar proof technique is used in [Ber+22] but for a line-search algorithm. We collect several relevant facts about supermartingales before proving the result.

First, we need to ensure the relevant supermartingale is well defined even after the algorithm terminates, so that it is possible to let the index t of the supermartingale go to ∞.

Fact A.7 ([Dur19, Theorem 4.2.9]). *If T is a stopping time and X_t is a supermartingale, then $X_{\min(T,t)}$ is a supermartingale.*

The following supermatingale convergence theorem will be used to ensure the function value converges, so that the algorithm terminates with probability 1.

[1] A supermartingale with respect to filtration $\{\mathcal{G}_1, \mathcal{G}_2, \dots\}$ is a sequence of random variables $\{Y_1, Y_2, \dots\}$ such that for all $k \in \mathbb{Z}_+$, (i) $\mathbb{E}\,|Y_t| < \infty$, (ii) Y_t is $\mathcal{G}_t$-measurable, and (iii) $\mathbb{E}(Y_{t+1}\,|\mathcal{G}_t) \leq Y_t$.

Fact A.8 (Supermartingale Convergence Theorem, [Dur19, Theorem 4.2.12]). *If $X_t \geq 0$ is a supermartingale, then as $t \to \infty$, there exists a random variable X such that $X_t \to X$ a.s. and $\mathbb{E}\, X \leq \mathbb{E}\, X_0$.*

Finally, we will use the optional stopping theorem to derive the expected iteration complexity. Note that we use a version of the optional stopping theorem specific to nonnegative supermartingales that does not require uniform integrability.

Fact A.9 (Optional Stopping Theorem, [Dur19, Theorem 4.8.4]). *If X_n is a nonnegative supermartingale and $N \leq \infty$ is a stopping time, then $\mathbb{E}\, X_0 \geq \mathbb{E}\, X_N$.*

Proof of Lemma A.6. We first construct a supermartingale based on function values. Since $\mathbb{E}\,[\sigma_t] = 0$, linearity of expectation implies that $\mathbb{E}\left[2\frac{\alpha_t}{M}\nabla f(x_t)^\top \sigma_t \widetilde{p}_t \middle| x_t\right] = 0$. We therefore have from (10) that

$$\mathbb{E}\,[f(x_{t+1})|x_t] \leq f(x_t) - \frac{2\epsilon_H^3}{9L_H^2}.$$

By combining with the (deterministic) first-order decrease estimate (9), we have

$$\mathbb{E}\,[f(x_{t+1})|x_t] \leq f(x_t) - \min\left(\frac{\epsilon_g^2}{6L_g}, \frac{2\epsilon_H^3}{9L_H^2}\right) = f(x_t) - C_\epsilon.$$

Consider the stochastic process $M_t := f(x_t) + tC_\epsilon$. We have

$$\begin{aligned}\mathbb{E}\,[M_{t+1}|x_t] &= \mathbb{E}\,[f(x_{t+1}) + (t+1)C_\epsilon|x_t] \\ &\leq \mathbb{E}\,[f(x_t) - C_\epsilon + (t+1)C_\epsilon|x_t] \\ &= \mathbb{E}\,[f(x_t) + tC_\epsilon|x_t] = M_t.\end{aligned}$$

We need to select a filtration to define the supermartingale M_t. We view x_t as random variables defined with respect to $\sigma_i, i \leq k$. Since M_t is expressed as a function of x_t only, we define the filtration $\{\mathcal{G}_t\}$ to be the filtration generated by x_t, and it naturally holds that M_t is $\mathcal{G}_t$-measurable for all k and $\mathbb{E}\,[M_{t+1}|\mathcal{G}_t] = \mathbb{E}\,[M_{t+1}|x_t]$. Hence, $\{M_t\}$ is a supermartingale with respect to filtration $\{\mathcal{G}_t\}$.

Let T denote the iteration at which our algorithm stops. Since the decision to stop at iteration t depends only on x_t, we have $\{T = t\} \in \mathcal{G}_t$, which implies T is a stopping time.

We will use the supermartingale convergence theorem (Fact A.8) to show that $T < +\infty$ almost surely, since the function value cannot decrease indefinitely as it is bounded by f^* from below. To apply Fact A.8, we need to let $t \to \infty$, so we need to transform $\{M_t\}$ to obtain a supermartingale $\{Y_t\}$that is well defined even after the algorithm terminates.

It follows from Fact A.7 that $Y_t := M_{\min(t,T)}$ is also a supermartingale. Since $Y_t \geq f^*$, it follows from the supermartingale convergence theorem (Fact A.8) applied to $Y_t - f^*$ that $Y_t \to Y_\infty$ almost surely for some random variable Y_∞ with $\mathbb{E}\, Y_\infty \leq \mathbb{E}\, Y_0 = \mathbb{E}\, M_0 = f(x_0) < \infty$. Hence $\mathbb{P}[Y_\infty = +\infty] = 0$. On the other hand, as $t \to \infty$, we have $tC_\epsilon \to \infty$, so $T = +\infty \implies Y_t = M_t \geq f^* + tC_\epsilon \to \infty \implies Y_\infty = +\infty$. Therefore we have $\mathbb{P}[T < +\infty] = 1$.

We can then apply the optional stopping theorem (Fact A.9) to $Y_t - f^*$. It follows that

$$f^* + \mathbb{E}\, T \cdot C_\epsilon \leq \mathbb{E}\, f(x_T) + \mathbb{E}\, T \cdot C_\epsilon = \mathbb{E}\,[M_T] = \mathbb{E}\,[Y_T] \leq \mathbb{E}[Y_0] = \mathbb{E}\,[M_0] = f(x_0),$$

where the first equality uses $T < +\infty$ almost surely and the last inequality is Fact A.9. By reorganizing this bound, we obtain $\mathbb{E}\, T \leq \frac{f(x_0) - f^*}{C_\epsilon}$, as desired. ∎

A.3 Robust Mean Estimation—Omitted background

A class of algorithms that robustly estimate quantities in the presence of outliers under strong contamination model (Definition 1.1) are based on the stability of samples:

Definition A.10 (Definition of Stability [Dia+17]). Fix $0 < \epsilon < 1/2$ and $\delta > \epsilon$. A finite set $S \subset \mathbb{R}^k$ is (ϵ, δ)-stable with respect to mean $\mu \in \mathbb{R}^k$ and σ^2 if for every $S' \subset S$ with $|S'| \geq (1-\epsilon)|S|$, the following conditions hold: (i) $\|\mu_{S'} - \mu\| \leq \sigma\delta$, and (ii) $\left\|\bar{\Sigma}_{S'} - \sigma^2 I\right\|_{\text{op}} \leq \sigma^2\delta^2\epsilon$, where $\mu_{S'}$ and $\bar{\Sigma}_{S'}$ denote the empirical mean and empirical covariance over the set S' respectively.

Algorithm A.2: RobustMeanEstimation with unknown covariance bound

Data: $0 < \epsilon < 1/2$ and T is an ϵ-corrupted set

Result: $\widehat{\mu}$ with $\|\widehat{\mu} - \mu_S\| = O(\sqrt{\|\Sigma\|_{\mathrm{op}}\, \epsilon})$

Initialize a weight function $w : T \to \mathbb{R}_{\geq 0}$ with $w(x) = 1/|T|$ for all $x \in T$;
while $\|w\|_1 \geq 1 - 2\epsilon$ **do**
  $\mu(w) := \frac{1}{\|w\|_1} \sum_{x \in T} w(x) x$;
  $\Sigma(w) := \frac{1}{\|w\|_1} \sum_{x \in T} w(x)(x - \mu(w))(x - \mu(w))^\top$;
  Compute the largest eigenvector v of $\Sigma(w)$;
  $g(x) := |v^\top (x - \mu(w))|^2$;
  Find the largest threshold t such that $\Sigma_{x \in T : g(x) \geq t} w(x) \geq \epsilon$;
  $f(x) := g(x) \mathbb{1}\{g(x) \geq t\}$;
  $w(x) \leftarrow w(x) \left(1 - \frac{f(x)}{\max_{y \in T : w(y) \neq 0} f(y)}\right)$;
return $\mu(w)$

On input the corrupted version of a stable set, Algorithm A.2 has the following guarantee.

Proposition A.11 (Robust Mean Estimation with Stability [DKP20, Theorem A.3])**.** *Let $T \subset \mathbb{R}^k$ be an ϵ-corrupted version of a set S, where S is $(C\epsilon, \delta)$-stable with respect to μ_S and σ^2, and where $C > 0$ is a sufficiently large constant. Algorithm A.2 on input ϵ and T (but not σ or δ) returns (deterministically) a vector $\widehat{\mu}$ in polynomial time so that $\|\mu_S - \widehat{\mu}\| = O(\sigma\delta)$.*

Proposition A.11 requires the uncorrupted samples to be stable. With a large enough sample size, most independent and identically distributed samples from a bounded covariance distribution are stable. The remaining samples can be treated as corruptions.

Proposition A.12 (Sample Complexity for Stability [DKP20, Theorem 1.4])**.** *Fix any $0 < \xi < 1$. Let S be a multiset of n independent and identically distributed samples from a distribution on $\mathbb{R}^k$ with mean μ and covariance Σ. Then, with probability at least $1 - \xi$, there exists a sample size $n = O\left(\frac{k \log k + \log(1/\xi)}{\epsilon}\right)$ and a subset $S' \subseteq S$ such that $|S'| \geq (1 - \epsilon)\, n$ and S' is $(2\epsilon, \delta)$-stable with respect to μ and $\|\Sigma\|_{\mathrm{op}}$, where $\delta = O(\sqrt{\epsilon})$.*

Proof of Proposition 2.3. Proposition A.12 implies that for i.i.d. samples from a k-dimensional bounded covariance distribution to contain a stable set of more than $(1 - \epsilon)$-fraction of samples with high probability, we need $\widetilde{O}\left(\frac{k}{\epsilon}\right)$ samples. We refer to the remaining ϵ-fraction of samples as unstable samples. Since the adversary corrupts an ϵ-fraction of clean samples S, the input set T can be considered as a 2ϵ-corrupted version of a stable set, if we view the unstable samples as corruptions. Therefore, Proposition A.11 applies and gives the desired error guarantee. ☐

B Omitted Proofs in General Robust Nonconvex Optimization

Sample Size for Successful Robust Mean Estimation For each iteration t in Algorithm A.1 and Algorithm 3.1, we would like to robustly estimate the mean of a set of corrupted points $\{\mathrm{vec}\,(\nabla f(x_t))\}_{i=1}^n$ and/or $\{\mathrm{vec}\,(\nabla^2 f(x_t))\}_{i=1}^n$ where the inliers are drawn from a distribution with bounded covariance $\Sigma \preceq \sigma^2 I$. For fixed x_t, inliers are drawn independently and identically distributed (i.i.d.), but the dependence across all iterations t is allowed.

Theorem B.1. *Let $\mathcal{X} \subset \mathbb{R}^l$ be a closed and bounded set with radius at most γ (i.e., $\|x\| \leq \gamma$, $\forall x \in \mathcal{X}$). Let p^* be a distribution over functions $h : \mathcal{X} \to \mathbb{R}^k$. Let $\xi \in (0, 1)$ be the failure probability. Suppose h is L-Lipschitz and uniformly bounded, i.e., there exists $B > 0$ such that $\|h(x)\| \leq B$ $\forall x \in \mathcal{X}$ almost surely. Assume further that for each $x \in \mathcal{X}$ we have $\|\mathrm{Cov}_{h \sim p^*}(h(x))\|_{\mathrm{op}} \leq \sigma^2$. Let S be a multiset of n i.i.d. samples $\{h_i\}_{i=1}^n$ from p^*. Then there exists a sample size*

$$n = O\left(\frac{k \log k + l \log\left(\frac{\gamma L B}{\sigma \epsilon \xi}\right)}{\epsilon}\right)$$

such that with probability $1-\xi$, it holds that for each $x \in \mathcal{X}$, there exists a subset $S' \subset S$ with $|S'| \geq (1-2\epsilon)n$ (potentially different subsets S' for different x) such that $\{h_i(x)\}_{i \in S'}$ is $(2\epsilon, \delta)$-stable with respect to μ and σ^2, where $\delta = O(\sqrt{\epsilon})$.

To prove this theorem, we work with an easier version of stability condition specialized to samples from a bounded covariance distribution.

Claim B.2 (Claim 2.1 in [DKP20])**.** *(Stability for bounded covariance) Let $R \subset \mathbb{R}^k$ be a finite multiset such that $\|\mu_R - \mu\| \leq \sigma\delta$, and $\left\|\bar{\Sigma}_R - \sigma^2 I\right\|_{\text{op}} \leq \sigma^2\delta^2/\epsilon$ for some $0 \leq \epsilon \leq \delta$. Then R is $(\Theta(\epsilon), \delta')$ stable with respect to μ and σ^2, where $\delta' = O(\delta + \sqrt{\epsilon})$.*

Proof of Theorem B.1. Given Claim B.2, it suffices to show that with probability $1-\xi$, for each $x \in \mathcal{X}$ there exists a subset $S' \subset S$ with $|S'| \geq (1-2\epsilon)n$ such that

$$\left\| \mathop{\mathbb{E}}_{i \in S'} h_i(x) - \mathop{\mathbb{E}}_{h \sim p^*} h(x) \right\| \leq O(\sigma\delta) \tag{13}$$

$$\left\| \mathop{\mathbb{E}}_{i \in S'} (h_i(x) - \mathop{\mathbb{E}}_{h \sim p^*} h(x))(h_i(x) - \mathop{\mathbb{E}}_{h \sim p^*} h(x))^\top - \sigma^2 I \right\|_{\text{op}} \leq O\big(\sigma^2\delta^2/\epsilon\big) \tag{14}$$

By Proposition A.12, for each $x \in \mathcal{X}$, with probability $1 - \left(\frac{\xi\gamma BL}{\sigma\sqrt{\epsilon}}\right)^l \geq 1 - \xi/\left(\frac{\gamma BL}{\sigma\sqrt{\epsilon}}\right)^l$, there exists a subset $S' \subset S$ with $|S'| \geq (1-2\epsilon)n$ such that (13) and (14) hold.

We proceed with a net argument. Up to a multiplicative error that can be suppressed by $O(\cdot)$, if Equation (13) holds for some $x \in \mathcal{X}$, it also holds for all other x' in a ball of radius $\sigma\delta/L$ because $h(\cdot)$ is L-Lipschitz. Similarly, (14) is equivalent to

$$\left| \mathop{\mathbb{E}}_{i \in S'} ((h_i(x) - \mathop{\mathbb{E}}_{h \sim p^*} h(x))^\top v)^2 - \sigma^2 v^\top v \right| \leq O\big(\sigma^2\delta^2/\epsilon\big) \qquad \forall v \in \mathbb{R}^k.$$

Since $h(\cdot)$ is L-Lipschitz and uniformly bounded by B, if (14) holds for some $x \in \mathcal{X}$, it also holds for all x' in a ball of radius $\frac{\sigma^2\delta^2}{\epsilon BL}$. Therefore, it suffices for Equation (13) and (14) to hold for a τ-net of $\mathcal{X}$, where for $\delta = O(\sqrt{\epsilon})$ we have $\tau = \min\left(\frac{\sigma\delta}{L}, \frac{\sigma^2\delta^2}{\epsilon BL}\right) = \Omega\left(\frac{\sigma\sqrt{\epsilon}}{BL}\right)$. An $\Omega\left(\frac{\sigma\sqrt{\epsilon}}{BL}\right)$-net of γ-radius ball in $\mathbb{R}^l$ has size $O\left(\left(\frac{\gamma BL}{\sigma\sqrt{\epsilon}}\right)^l\right)$. Taking a union bound over this net completes the proof. ☐

Proof of Main Theorem We establish a slightly enhanced version of Theorem 3.1 by including the last sentence as an additional component to the original theorem.

Theorem B.3. *Suppose we are given ϵ-corrupted set of functions $\{f_i\}_{i=1}^n$ for sample size n, generated according to Definition 1.2. Suppose Assumption 1.4 holds in a bounded region $\mathcal{B} \subset \mathbb{R}^D$ of radius Γ with gradient and Hessian covariance bound σ_g and σ_H respectively, and we have an arbitrary initialization $x_0 \in \mathcal{B}$. For some $n = \widetilde{O}(D^2/\epsilon)$, Algorithm A.1 initialized at x_0 outputs an (ϵ_g, ϵ_H)-approximate SOSP in polynomial time with high probability if the following conditions hold:*

- *(I) All iterates x_t in Algorithm A.1 stay inside the bounded region $\mathcal{B}$.*
- *(II) For an absolute constant $c > 0$, it holds that $\sigma_g\sqrt{\epsilon} \leq c\epsilon_g$ and $\sigma_H\sqrt{\epsilon} \leq c\epsilon_H$.*

Moreover, there exists an absolute constant C_{est} such that for each iteration t, the gradient oracle $\widetilde{g}_t$ and Hessian oracle $\widetilde{H}_t$ satisfy $\|\widetilde{g}_t - \nabla f(x_t)\| \leq C_{est}\sigma_g\sqrt{\epsilon}$ and $\|\widetilde{H}_t - \nabla^2 f(x_t)\|_F \leq C_{est}\sigma_H\sqrt{\epsilon}$.

And we recall that the we construct the gradient and Hessian oracles in Algorithm A.1 in the following way:

$$\widetilde{g}_t \leftarrow \textbf{RobustMeanEstimation}(\{\nabla f_i(x_t)\}_{i=1}^n, 4\epsilon)$$
$$\widetilde{H}_t \leftarrow \textbf{RobustMeanEstimation}(\{\nabla^2 f_i(x_t)\}_{i=1}^n, 4\epsilon)$$

If we ignore that the dependence between $\{\nabla f_i(x_t)\}$ introduced via x_t, this theorem follows directly from Proposition 2.3. Here we fix this technicality via a union bound over all x_t with Theorem B.1.

Proof. By Proposition 2.2, it suffices to check that gradient and Hessian inexactness condition is satisfied. That is, for all iterations t, it holds that

$$\|\widetilde{g}_t - \nabla f(x_t)\| \le \frac{1}{3}\epsilon_g, \tag{15}$$

$$\|\mathrm{vec}(\widetilde{H}_t) - \mathrm{vec}(\nabla^2 f(x_t))\| \le \frac{2}{9}\epsilon_H. \tag{16}$$

Here we use the fact that $\|\mathrm{vec}(\widetilde{H}_t) - \mathrm{vec}(\nabla^2 f(x_t))\| = \|\widetilde{H}_t - \nabla^2 f(x_t)\|_F \ge \|\widetilde{H}_t - \nabla^2 f(x_t)\|_{\mathrm{op}}$.

We proceed to apply Theorem B.1 and consider the function $h_g(x, A) = \mathrm{vec}(\nabla_x f(x, A))$ and $h_H(x) = \mathrm{vec}(\nabla^2_{xx} f(x, A))$. We consider $h_H : \mathbb{R}^D \to \mathbb{R}^{D^2}$ first. By assumption (I), all iterates never leave the bounded region $\mathcal{B}$ with high probability; we condition on this event in the remaining analysis and it follows that $\left\|\mathrm{Cov}_{A\sim\mathcal{G}}(\mathrm{vec}(\nabla^2 f(x_t, A)))\right\|_{\mathrm{op}} \le \sigma_H^2$ for all iterations t.

Assumption 1.4 (ii) posits that with high probability $h_H(\cdot, A)$ is L_{D_H}-Lipschitz and its ℓ_2-norm is bounded above by B_{D_H}. We further condition on this event. Let $\xi \in (0, 1)$. Since $\|\mathrm{Cov}_{A\in\mathcal{G}}(h_H(x, A))\|_{\mathrm{op}} \le \sigma_H$ by Assumption 1.4 (iii), there exists a sample size

$$n = O\left(\frac{D^2 \log D + D \log\left(\frac{\gamma L_{D_H} B_{D_H}}{\sigma_H \epsilon \xi}\right)}{\epsilon}\right),$$

such that with probability $1 - \xi$, it holds that for each $x \in \mathcal{X}$, there exists a $(2\epsilon, O(\sqrt{\epsilon}))$-stable subset of size at least $(1 - 2\epsilon)n$ in the sense of Definition A.10 (potentially different subsets S' for different x). Therefore, conditioning on the event that for all x there exists a stable subset, Proposition A.11 implies that there exists an absolute constant C_{est} such that $\|\mathrm{vec}(\widetilde{H}_t) - \mathrm{vec}(\nabla^2 f(x_t))\| \le C_{est}\sigma_H\sqrt{\epsilon}$ for all t. By assumption (II) with a sufficiently small constant c, we have $C_{est}\sigma_H\sqrt{\epsilon} \le 2\epsilon_H/9$, and therefore $\|\mathrm{vec}(\widetilde{H}_t) - \mathrm{vec}(\nabla^2 f(x_t))\| \le \frac{2}{9}\epsilon_H$.

Now we established that for Equation (16) to hold for all iterations t with high probability, the sample complexity is required to be at least $n = \widetilde{O}(D^2/\epsilon)$. A similar argument implies that $\widetilde{O}(D/\epsilon)$ samples are needed for Equation (15) to hold for all iterations t with high probability, which is dominated by $\widetilde{O}(D^2/\epsilon)$. □

C Omitted Proofs in Low Rank Matrix Sensing

This section provides omitted computation and proofs for robust low rank matrix sensing (Section 3.1), establishing Theorem 1.7. Appendix C.1 computes the gradient and Hessian of f_i defined in Equation (3) and their covariance upper bounds under Definition 1.6 with noiseless measurements ($\sigma = 0$). Appendix C.2 proves the global convergence of our general robust nonconvex optimization framework (Theorem 3.1) applied to the noiseless robust low rank matrix sensing problem, which leads to an approximate SOSP, and Appendix C.3 proves the local linear convergence to a global minimum from this approximate SOSP. Finally, Appendix C.4 discusses the case of noisy measurements ($\sigma \neq 0$) under Definition 1.6.

C.1 Omitted Computation in Low Rank Matrix Sensing

C.1.1 Sample Gradient and Sample Hessian

Summary of Results Consider the noiseless setting as in Theorem 1.7 with $\sigma = 0$. We now compute the gradient and the Hessian of $f_i : \mathbb{R}^{d\times r} \to \mathbb{R}$. We firstly summarize the results. The gradient of f_i at point U, denoted by $\nabla f_i(U)$, is a matrix in $\mathbb{R}^{d\times r}$ given by

$$\nabla f_i(U) = \langle UU^\top - M^*, A_i\rangle(A_i + A_i^\top)U.$$

The Hessian of f_i at point U, denoted by $(H_i)_U$, is a linear operator $(H_i)_U : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$. The operator $(H_i)_U$ acting on a matrix $Y \in \mathbb{R}^{d\times r}$ gives:

$$(H_i)_U(Y) = \langle UU^\top - M^*, A_i\rangle(A_i + A_i^\top)Y + \langle Y, (A_i + A_i^\top)U\rangle(A_i + A_i^\top)U.$$

We want to identify this linear map with a matrix $H_i(U)$ of dimension $\mathbb{R}^{dr\times dr}$. We vectorize both the domain and the codomain of $(H_i)_U$ so that it can be represented by a matrix.

Let I_r and I_d denote the identity matrices of dimension $r \times r$ and $d \times d$ respectively.

$$H_i(U) = \langle UU^\top - M^*, A_i\rangle I_r \otimes (A_i + A_i^\top) + \operatorname{vec}\left((A_i + A_i^\top)U\right) \operatorname{vec}\left((A_i + A_i^\top)U\right)^\top.$$

In this paper, we sometimes abuse the notation and use $\nabla^2 f_i(U)$ to refer to either $H_i(U)$ or $(H_i)_U$, but its precise meaning should be clear from its domain.

Computation of Gradients and Hessians Let $(D^k f_i)_U$ denote the k-th order derivative of f_i at point $U \in \mathbb{R}^{d\times r}$ and $\mathcal{L}(X, Y)$ denote the space of all linear mappings from X to Y. We consider higher-order derivatives as linear maps:

$$(Df_i)_U : \mathbb{R}^{d\times r} \to \mathbb{R}$$
$$(D^2 f_i)_U : \mathbb{R}^{d\times r} \to \mathcal{L}(\mathbb{R}^{d\times r}, \mathbb{R})$$

We identify $(Df_i)_U$ with the matrix $\nabla f_i(U)$ such that

$$(Df_i)_U(Z) = \langle \nabla f_i(U), Z\rangle$$

and identify $(D^2 f_i)_U$ with a linear operator $(H_i)_U : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ such that

$$(D^2 f_i)_U(Y)(Z) = \langle (H_i)_U(Y), Z\rangle.$$

Since f_i is differentiable, applying its derivative at U to a matrix Z gives the corresponding directional derivative at U for the direction Z.

$$\begin{aligned}
(Df_i)_U(Z) &= \frac{\mathrm{d}}{\mathrm{d}t} f_i(U + tZ)\Big|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}(\langle (U + tZ)(U + tZ)^\top, A_i\rangle - y_i)^2\Big|_{t=0} \\
&= \left(\langle UU^\top, A_i\rangle - y_i\right)\left(\frac{\mathrm{d}}{\mathrm{d}t}\langle UU^\top + t^2 ZZ^\top + t(ZU^\top + UZ^\top), A_i\rangle\right)\Big|_{t=0} \\
&= \left(\langle UU^\top, A_i\rangle - y_i\right)\langle \frac{\mathrm{d}}{\mathrm{d}t}\left(t^2 ZZ^\top + t(ZU^\top + UZ^\top)\right), A_i\rangle\Big|_{t=0} \\
&\overset{(a)}{=} \left(\langle UU^\top, A_i\rangle - y_i\right)\langle ZU^\top + UZ^\top, A_i\rangle \\
&= \left(\langle UU^\top, A_i\rangle - y_i\right)\langle A_i, ZU^\top\rangle + \langle A_i, UZ^\top\rangle \\
&= \left(\langle UU^\top, A_i\rangle - y_i\right)\langle A_i U, Z\rangle + \langle U^\top A_i, Z^\top\rangle \\
&\overset{(b)}{=} \left(\langle UU^\top, A_i\rangle - y_i\right)\langle (A_i + A_i^\top)U, Z\rangle.
\end{aligned}$$

From (a) we conclude $(Df_i)_U$ is the functional $Z \mapsto \left(\langle UU^\top, A_i\rangle - y_i\right)\langle ZU^\top + UZ^\top, A_i\rangle$; it would be easier to calculate the second derivative from this form. From (b), using $y_i = \langle A_i, M^*\rangle$, we obtain the following closed-form expression for the gradient $\nabla f_i(U)$:

$$\nabla f_i(U) = \langle UU^\top - M^*, A_i\rangle (A_i + A_i^\top)U.$$

To calculate the second derivative of f_i at point U, we study the variation of its first derivative $U \mapsto (Df_i)_U$ at direction Y. Recall that the second derivative lives in the space of linear functionals

$\mathcal{L}(\mathbb{R}^{d\times r}, \mathbb{R})$ and we use $Z \in \mathbb{R}^{d\times r}$ to denote the input of this linear functional.

$$
\begin{aligned}
(D^2 f_i)_U(Y) &= \frac{\mathrm{d}}{\mathrm{d}t}(Df_i)_{U+tY}\Big|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\left\{Z \mapsto \left(\langle (U+tY)(U+tY)^\top, A_i\rangle - y_i\right)\langle Z(U+tY)^\top + (U+tY)Z^\top, A_i\rangle\right\}\Big|_{t=0} \\
&= \left(\langle UU^\top, A_i\rangle - y_i\right)\frac{\mathrm{d}}{\mathrm{d}t}\left\{Z \mapsto \langle Z(U+tY)^\top + (U+tY)Z^\top, A_i\rangle\right\}\Big|_{t=0} + \\
&\qquad \frac{\mathrm{d}}{\mathrm{d}t}\left\{\langle (U+tY)(U+tY)^\top, A_i\rangle - y_i\right\}\Big|_{t=0}\left\{Z \mapsto \langle ZU^\top + UZ^\top, A_i\rangle\right\} \\
&= \left(\langle UU^\top, A_i\rangle - y_i\right)\left\{Z \mapsto \langle ZY^\top + YZ^\top, A_i\rangle\right\} + \\
&\qquad \langle UY^\top + YU^\top, A_i\rangle\left\{Z \mapsto \langle ZU^\top + UZ^\top, A_i\rangle\right\} \\
&= Z \mapsto \left(\langle UU^\top, A_i\rangle - y_i\right)\langle ZY^\top + YZ^\top, A_i\rangle + \langle UY^\top + YU^\top, A_i\rangle\langle ZU^\top + UZ^\top, A_i\rangle \\
&= Z \mapsto (\langle UU^\top, A_i\rangle - y_i)\langle (A_i + A_i^\top)Y, Z\rangle + \langle Y, (A_i + A_i^\top)U\rangle\langle (A_i + A_i^\top)U, Z\rangle.
\end{aligned}
$$

Recall that we identify $(D^2 f_i)_U$ with a linear operator $(H_i)_U : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ such that

$$(D^2 f_i)_U(Y)(Z) = \langle (H_i)_U(Y), Z\rangle.$$

Note that $(H_i)_U : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ viewed as a fourth-order tensor is difficult to write down in a closed form, so we vectorize both the domain and the codomain of $(H_i)_U$ so that it can be represented by a matrix.

The operator $(H_i)_U$ acting on matrix $Y \in \mathbb{R}^{d\times r}$ gives

$$(H_i)_U(Y) = \langle UU^\top - M^*, A_i\rangle(A_i + A_i^\top)Y + \langle Y, (A_i + A_i^\top)U\rangle(A_i + A_i^\top)U.$$

We identify this linear map with a matrix $H_i(U)$ of dimension $\mathbb{R}^{dr\times dr}$, acting on the vectorized version of Y. As a shorthand, write $y = \mathrm{vec}(Y)$ and $B_i = A_i + A_i^\top$.

Let I_r and I_d denote the identity matrix of dimension $r \times r$ and $d \times d$ respectively. Then

$$
\begin{aligned}
H_i(U)\, y &= \mathrm{vec}\left(\langle UU^\top - M^*, A_i\rangle(A_i + A_i^\top)Y + \langle Y, (A_i + A_i^\top)U\rangle(A_i + A_i^\top)U\right) \\
&= \langle UU^\top - M^*, A_i\rangle\, \mathrm{vec}\left(B_i Y I_r\right) + \mathrm{vec}\left(B_i U\right)\mathrm{vec}(B_i U)^\top\, \mathrm{vec}(Y) \\
&= \langle UU^\top - M^*, A_i\rangle I_r \otimes B_i\, \mathrm{vec}\left(Y\right) + \mathrm{vec}\left(B_i U\right)\mathrm{vec}(B_i U)^\top\, \mathrm{vec}(Y) \\
&= \langle UU^\top - M^*, A_i\rangle(I_r \otimes B_i)y + \mathrm{vec}\left(B_i U\right)\mathrm{vec}(B_i U)^\top y.
\end{aligned}
$$

Hence

$$H_i(U) = \langle UU^\top - M^*, A_i\rangle I_r \otimes B_i + \mathrm{vec}\left(B_i U\right)\mathrm{vec}(B_i U)^\top.$$

C.1.2 Gradient Covariance Bound

Lemma C.1 (Equation (6)). *Consider the noiseless setting as in Theorem 1.1 with $\sigma = 0$. For all $U \in \mathbb{R}^{d\times r}$ with $\|U\|_{\mathrm{op}}^2 \le \Gamma$, it holds that*

$$\left\|\mathrm{Cov}(\mathrm{vec}(\nabla f_i(U)))\right\|_{\mathrm{op}} \le 8\left\|UU^\top - M^*\right\|_F^2 \|U\|_{\mathrm{op}}^2 \le 32 r^2 \Gamma^3. \tag{17}$$

Recall that $\Gamma \ge 36\sigma_1^\star$, so we have the following trivial bound that is used frequently

$$\|M^*\|_{\mathrm{op}} = \sigma_1^\star \le \Gamma/36 < \Gamma.$$

Proof of Equation (6). Recall that

$$\nabla f_i(U) = \langle UU^\top - M^*, A_i\rangle(A_i + A_i^\top)U.$$

We frequently encounter the following calculations.

Lemma C.2. *Let $P, Q \in \mathbb{R}^{d\times d}$ be given and let A_i have i.i.d. standard Gaussian entries. Then*

$$\mathbb{E}[\langle P, A_i\rangle\langle Q, A_i\rangle] = \langle P, Q\rangle, \tag{18}$$

$$\mathrm{Var}[\langle P, A_i\rangle\langle Q, A_i\rangle] \le 2\|P\|_F^2 \|Q\|_F^2. \tag{19}$$

Proof. By Definition 1.6, $X := \mathrm{vec}(A_i)$ is a standard Gaussian vector with identity covariance $\Sigma = I_{d^2}$. Let $a = \mathrm{vec}(P), b = \mathrm{vec}(Q)$, and $G = ab^\top$. Fact A.4 implies that

$$\begin{aligned}
\mathbb{E}\langle P, A_i\rangle\langle Q, A_i\rangle &= \langle ab^\top, I_{dr}\rangle = \langle a, b\rangle = \langle P, Q\rangle \\
\mathrm{Var}[\langle P, A_i\rangle\langle Q, A_i\rangle] &= \mathrm{tr}(G\Sigma(G + G^\top)\Sigma) = \mathrm{tr}(G(G + G^\top)) \\
&= \mathrm{tr}(ab^\top(ab^\top + ba^\top)) \\
&= b^\top a\,\mathrm{tr}(ab^\top) + b^\top b\,\mathrm{tr}(aa^\top) \\
&= \langle a, b\rangle^2 + \|a\|^2\|b\|^2 \\
&\le 2\|a\|^2\|b\|^2 \qquad \text{(Cauchy-Schwarz)} \\
&= 2\|P\|_F^2\|Q\|_F^2.
\end{aligned}$$

□

To compute the mean and the variance of the gradient $\nabla f_i(U)$, it is more convenient to work with $\langle \nabla f_i(U), Z\rangle$ for some $Z \in \mathbb{R}^{d\times r}$ with $\|Z\|_F = 1$.

$$\begin{aligned}
\langle \nabla f_i(U), Z\rangle &= \langle UU^\top - M^*, A_i\rangle\langle (A_i + A_i^\top)U, Z\rangle \\
&= \langle UU^\top - M^*, A_i\rangle\langle A_i + A_i^\top, ZU^\top\rangle \\
&= \langle UU^\top - M^*, A_i\rangle\langle ZU^\top + UZ^\top, A_i\rangle.
\end{aligned}$$

Lemma C.2 implies that the expectation

$$\begin{aligned}
\mathbb{E}\langle \nabla f_i(U), Z\rangle &= \langle UU^\top - M^*, ZU^\top + UZ^\top\rangle \\
&= \langle \left((UU^\top - M^*) + (UU^\top - M^*)^\top\right)U, Z\rangle \\
&= \langle 2\left(UU^\top - M^*\right)U, Z\rangle.
\end{aligned}$$

By linearity of expectation, we conclude

$$\mathbb{E}\,\nabla f_i(U) = 2(UU^\top - M^*)U. \tag{20}$$

The variance bound can also be obtained via Lemma C.2

$$\mathrm{Var}\langle \nabla f_i(U), Z\rangle \le 2\left\|UU^\top - M^*\right\|_F^2 \left\|ZU^\top + UZ^\top\right\|_F^2 \le 8\|UU^\top - M^*\|_F^2\|U\|_{\mathrm{op}}^2\|Z\|_F^2,$$

where the last inequality comes from Fact A.1.

Since $\mathrm{Var}\langle \nabla f_i(U), Z\rangle = \mathrm{Var}[\mathrm{vec}(\nabla f_i(U))^\top \mathrm{vec}(Z)] = \mathrm{vec}(Z)^\top \mathrm{Cov}(\mathrm{vec}(\nabla f_i(U)))\,\mathrm{vec}(Z)$, we have

$$\begin{aligned}
\|\mathrm{Cov}(\mathrm{vec}(\nabla f_i(U)))\|_{\mathrm{op}} &\le 8\left\|UU^\top - M^*\right\|_F^2\|U\|_{\mathrm{op}}^2 \\
&\le 8\Gamma\left(\left\|UU^\top\right\|_F + \|M^*\|_F\right)^2 \\
&\le 8\Gamma(r\Gamma + r\sigma_1^\star)^2 = 32r^2\Gamma^3.
\end{aligned}$$

□

C.1.3 Hessian Covariance Bound

Lemma C.3 (Equation (7))**.** *Consider the noiseless setting as in Theorem 1.7 with $\sigma = 0$. For all $U \in \mathbb{R}^{d\times r}$ with $\|U\|_{\mathrm{op}}^2 \le \Gamma$, it holds that*

$$\|\mathrm{Cov}(\mathrm{vec}(H_i))\|_{\mathrm{op}} \le 16r\left\|UU^\top - M^*\right\|_F^2 + 128\|U\|_{\mathrm{op}}^4 \le 192r^3\Gamma^2. \tag{21}$$

Proof. As a shorthand, we write $B_i = A_i + A_i^\top$. Recall that

$$H_i(U) = \langle UU^\top - M^*, A_i\rangle I_r \otimes B_i + \operatorname{vec}(B_iU)\operatorname{vec}(B_iU)^\top.$$

Let W be a $dr \times dr$ matrix that has Frobenius norm 1. We work with $\langle H_i(U), W\rangle$ as we did in the gradient covariance bound calculation.

By Young's inequality for L^2 random variables (i.e., $\|\mathcal{A}+\mathcal{B}\|_{L^2}^2 \le 2\|\mathcal{A}\|_{L^2}^2 + 2\|\mathcal{B}\|_{L^2}^2$),

$$\operatorname{Var}\langle H_i(U), W\rangle \le 2\operatorname{Var}\underbrace{[\langle UU^\top - M^*, A_i\rangle\langle I_r\otimes B_i, W\rangle]}_{\mathcal{A}} + 2\operatorname{Var}\underbrace{\left[\langle \operatorname{vec}(B_iU)\operatorname{vec}(B_iU)^\top, W\rangle\right]}_{\mathcal{B}}.$$

We first bound the term $\mathcal{A}$. Note $I_r \otimes B_i$ consists of r copies of B_i in the diagnal

$$I_r \otimes B_i = \begin{bmatrix} B_i & \cdots & O \\ \vdots & \ddots & \vdots \\ O & \cdots & B_i \end{bmatrix}.$$

We partition W into r^2 submatrices of dimension $d \times d$.

$$W = \begin{bmatrix} W_{11} & \cdots & W_{1r} \\ \vdots & \ddots & \vdots \\ W_{r1} & \cdots & W_{rr} \end{bmatrix}.$$

Then

$$\begin{aligned}
&\operatorname{Var}[\langle UU^\top - M^*, A_i\rangle\langle I_r\otimes B_i, W\rangle] \\
&= \operatorname{Var}[\langle UU^\top - M^*, A_i\rangle\langle B_i, \sum_{i=1}^r W_{ii}\rangle] \\
&= \operatorname{Var}[\langle UU^\top - M^*, A_i\rangle\langle A_i, \sum_{i=1}^r (W_{ii}+W_{ii}^\top)\rangle] \\
&\le 2\left\|UU^\top - M^*\right\|_F^2 \left\|\sum_{i=1}^r (W_{ii}+W_{ii}^\top)\right\|_F^2,
\end{aligned}$$

where the last inequality uses Lemma C.2.

Write $a_i = \operatorname{vec}(W_{ii}) \in \mathbb{R}^{d^2}, b_i = \operatorname{vec}(W_{ii}^\top), c_i = a_i + b_i$. Then $\|a_i\| = \|b_i\| = \|W_{ii}\|_F$ and

$$\begin{aligned}
\left\|\sum_{i=1}^r (W_{ii}+W_{ii}^\top)\right\|_F^2 &= \left\|\sum_{i=1}^r c_i\right\|^2 \\
&= \sum_{j=1}^{d^2}\left(\sum_{i=1}^r c_{ij}\right)^2 \le \sum_{j=1}^{d^2} r\sum_{i=1}^r c_{ij}^2 \qquad \text{(Cauchy-Schwarz)} \\
&= r\sum_{i=1}^r \|c_i\|^2 = r\sum_{i=1}^r \|a_i+b_i\|^2 \\
&\le 2r\left(\sum_{i=1}^r (\|a_i\|^2 + \|b_i\|^2)\right) \\
&= 4r\sum_{i=1}^r \|W_{ii}\|_F^2 \le 4r\|W\|_F^2 = 4r.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\operatorname{Var}[\langle UU^\top - M^*, A_i\rangle\langle I_r\otimes B_i, W\rangle] &\le 2\left\|UU^\top - M^*\right\|_F^2 \left\|\sum_{i=1}^r (W_{ii}+W_{ii}^\top)\right\|_F^2 \\
&\le 8r\left\|UU^\top - M^*\right\|_F^2 \|W\|_F^2 = 8r\left\|UU^\top - M^*\right\|_F^2.
\end{aligned}$$

To bound the term $\mathcal{B}$, we apply Fact A.4 with $\Sigma = \mathrm{Cov}(\mathrm{vec}(B_i U)) \in \mathbb{R}^{dr \times dr}$:

$$\begin{aligned}
&\mathrm{Var}\left[\langle \mathrm{vec}\left(B_i U\right) \mathrm{vec}(B_i U)^\top, W\rangle\right] \\
&= \mathrm{Var}\left[\mathrm{vec}(B_i U)^\top W \,\mathrm{vec}(B_i U)\right] \\
&= \mathrm{tr}(W\Sigma(W + W^\top)\Sigma) \\
&= \langle W\Sigma, \Sigma W\rangle + \langle W\Sigma, \Sigma W^\top\rangle \\
&\le \|W\Sigma\|_F \|\Sigma W\|_F + \|W\Sigma\|_F \left\|\Sigma W^\top\right\|_F \\
&\le \|W\|_F \|\Sigma\|_{\mathrm{op}} \|\Sigma\|_{\mathrm{op}} \|W\|_F + \|W\|_F \|\Sigma\|_{\mathrm{op}} \|\Sigma\|_{\mathrm{op}} \|W\|_F \\
&= 2\|W\|_F^2 \|\Sigma\|_{\mathrm{op}}^2 = 2\|\Sigma\|_{\mathrm{op}}^2,
\end{aligned}$$

where the last inequality comes from Fact A.1.

To bound $\|\Sigma\|_{\mathrm{op}}^2$, we use Fact A.2 again. Since $\mathrm{vec}(B_i U) = \mathrm{vec}(I_d B_i U) = (U^\top \otimes I_d)\,\mathrm{vec}(B_i)$, we have

$$\mathrm{Cov}(\mathrm{vec}\left(B_i U\right)) = (U^\top \otimes I_d)\,\mathrm{Cov}(\mathrm{vec}(B_i))(U \otimes I_d). \tag{22}$$

Let $P \in \mathbb{R}^{d^2 \times d^2}$ be the permutation that maps $\mathrm{vec}(A_i)$ to $\mathrm{vec}(A_i^\top)$. Then

$$\begin{aligned}
\mathrm{Cov}(\mathrm{vec}(B_i)) &= \mathrm{Cov}(\mathrm{vec}(A_i) + \mathrm{vec}(A_i^\top)) \\
&\le 2I_{d^2} + 2\,\mathrm{Cov}(\mathrm{vec}(A_i), \mathrm{vec}(A_i^\top)) && \text{(Young's inequality)} \\
&= 2I_{d^2} + 2\,\mathrm{Cov}(\mathrm{vec}(A_i), P\,\mathrm{vec}(A_i)) \\
&= 2I_{d^2} + 2P\,\mathrm{Cov}(\mathrm{vec}(A_i), \mathrm{vec}(A_i)) \\
&= 2I_{d^2} + 2P.
\end{aligned}$$

Since P is a permutation, we have $\|\mathrm{Cov}(\mathrm{vec}(B_i))\|_{\mathrm{op}} \le 4$.

It follows from Equation (22) and Fact A.3 that $\|\mathrm{Cov}(\mathrm{vec}(B_i U))\|_{\mathrm{op}} \le 4\|U\|_{\mathrm{op}}^2$ and therefore

$$\mathrm{Var}\left[\langle \mathrm{vec}\left(B_i U\right) \mathrm{vec}(B_i U)^\top, W\rangle\right] \le 64\|U\|_{\mathrm{op}}^4 \|W\|_F^2 = 64\|U\|_{\mathrm{op}}^4.$$

We conclude that for all $W \in \mathbb{R}^{dr \times dr}$,

$$\mathrm{Var}\langle H_i, W\rangle \le 16r\left\|UU^\top - M^*\right\|_F^2 + 128\|U\|_{\mathrm{op}}^4,$$

hence

$$\|\mathrm{Cov}(\mathrm{vec}(H_i))\|_{\mathrm{op}} \le 16r\left\|UU^\top - M^*\right\|_F^2 + 128\|U\|_{\mathrm{op}}^4 \le 192r^3\Gamma^2,$$

where the last inequality uses $r \ge 1$ and $\|U\|_{\mathrm{op}}^2 \le \Gamma$. □

C.2 Global Convergence in Low Rank Matrix Sensing—Omitted Proofs

We firstly verify Assumption 1.4 (ii). We recall the entire assumption below:

Assumption C.4 (Assumption 1.4)**.** *There exists a bounded region $\mathcal{B}$ such that the function f satisfies:*

(i) There exists a lower bound $f^* > -\infty$ such that for all $x \in \mathcal{B}$, $f(x, A) \ge f^*$ with probability 1.

(ii) There exist parameters $L_{D_g}, L_{D_H}, B_{D_g}, B_{D_H}$ such that with high probability over the randomness in A, $f(\cdot, A)$ is L_{D_g}-gradient Lipschitz and L_{D_H}-Hessian Lipschitz, and the ℓ_2-norm of gradient and the Frobenius norm of Hessian of $f(\cdot, A)$ are upper bounded by B_{D_g} and B_{D_H} respectively.

(iii) There exist parameters $\sigma_g, \sigma_H > 0$ such that

$$\|\mathrm{Cov}_{A\sim\mathcal{G}}(\nabla f(x, A))\|_{\mathrm{op}} \le \sigma_g \text{ and } \left\|\mathrm{Cov}_{A\sim\mathcal{G}}(\mathrm{vec}(\nabla^2 f(x, A)))\right\|_{\mathrm{op}} \le \sigma_H.$$

Verifying Assumption 1.4(ii). We take the bounded region $\mathcal{B}$ to be the region $\{U : \|U\|_{\text{op}}^2 \le \Gamma\}$. Let $f_i(U) = \frac{1}{2}\left(\langle UU^\top, A_i\rangle - y_i\right)^2$ be the cost function corresponding to clean samples. We need to verify that $\nabla f_i(U)$ and $\nabla^2 f_i(U)$ are $\text{poly}(dr\Gamma/\epsilon)$-Lipschitz and bounded within $\mathcal{B}$ with high probability so that the violated samples constitute at most an ϵ-fraction.

We discuss gradient-Lipschitzness as an example, focusing on the constant L_{D_g}. The rest of conditions can be checked similarly. We drop the subscript $L := L_{D_g}$ in the following analysis for conciseness of the notation.

Since $\nabla f_i(X) - \nabla f_i(Y) = \langle XX^\top - M^*, A_i\rangle(A_i + A_i^\top)(X-Y) + \langle XX^\top - YY^\top, A_i\rangle(A_i + A_i^\top)Y$, it suffices that $\|A_i\|_F^2 \le 2d^2 + 3/\epsilon$ for $\nabla f_i(\cdot)$ to be $\text{poly}(dr\Gamma/\epsilon)$-Lipschitz. Since $\|A_i\|_F^2$ follows chi-square distribution with degree of freedom d^2, by Laurent-Massart bound, we have

$$\mathbb{P}\{\|A_i\|_F^2 - d^2 \ge d^2 + 3/\epsilon\} \le \exp(-1/\epsilon).$$

This completes the verification of Assumption 1.4(ii) for L_{D_g}. We proceed to discuss why this high probability result implies that at most an ϵ-fraction of clean samples violates gradient Lipschitzness. By Chernoff's inequality, the probability that more than ϵ-fraction of uncorrupted A_i's fail to satisfy $\|A_i\|_F^2 \le 2d^2 + 3/\epsilon$ is less than

$$\exp\left(-n\left((1-\epsilon)\log\frac{1-\epsilon}{1-\exp(-1/\epsilon)} + \epsilon\log\frac{\epsilon}{\exp(-1/\epsilon)}\right)\right) = O\left(\frac{1}{\epsilon}\exp(-n)\right),$$

which is a small if $n = \widetilde{O}\left(\frac{1}{\epsilon}\right)$ is sufficiently large. □

Corollary C.5. *Consider the noiseless setting as in Theorem 1.7 with $\sigma = 0$. There exists a sample size $n = \widetilde{O}\left(\frac{d^2r^2}{\epsilon}\right)$ such that with n samples, all subroutine calls to* RobustMeanEstimation *(Algorithm A.2) to estimate population gradients $\nabla\bar{f}(U_t)$ and population Hessians $\nabla^2\bar{f}(U_t)$ succeed with high probability. In light of Equations (6) and (7), this means that there exists a large enough constant C_{est} such that with high probability, for all iterations t, we have*

$$\left\|\widetilde{g}_t - \nabla\bar{f}(U_t)\right\|_F \le 2C_{est}\left\|U_tU_t^\top - M^*\right\|_F\|U_t\|_{\text{op}} \tag{23}$$

$$\left\|\widetilde{H}_t - \nabla^2\bar{f}(U_t)\right\|_F \le 4r^{1/2}\left\|U_tU_t^\top - M^*\right\|_F + 16\|U_t\|_{\text{op}}^2. \tag{24}$$

Proof. This follows directly from the last sentence of Theorem B.3, with $D = d^2r^2$ and σ_g, σ_H given by Equations (6) and (7). □

We next prove Lemma 3.7, establishing that all iterates stay inside a bounded region in which the covariance bounds are valid. The analysis utilizes a "dissipativity" property [Hal10]), which says that the iterate aligns with the direction of the gradient when the iterate's norm is large. For the gradient step, the gradient will reduce the norm of the iterate. For the negative curvature step, dissipativity property provides a lower bound on the gradient when the iterate's norm is large, but we show that Algorithm A.1 only takes a negative curvature step when this lower bound is violated, therefore the iterate's norm must be small and negative curvature steps with a fixed and small stepsize cannot increase the iterate's norm by too much.

Proof of Lemma 3.7. Recall Algorithm A.1 consists of inexact gradient descent steps of size $1/L_g = \frac{1}{16\Gamma}$ and randomized inexact negative curvature steps of size

$$\frac{2\epsilon_H}{L_H} = \frac{\sigma_r^\star}{36\Gamma} \le \frac{\Gamma^{1/2}}{36}. \tag{25}$$

We proceed to use induction to prove the following for Algorithm A.1:

1. Suppose at step τ we run the negative curvature step to update the iterate from U_τ to $U_{\tau+1}$, then $\|U_\tau\|_{\text{op}} \le \frac{1}{2}\Gamma^{1/2}$
2. $\|U_t\|_{\text{op}} \le \Gamma^{1/2}$ for all $t \ge 0$.

First we consider the inexact gradient steps. We denote the gradient inexactness by $e_t = \nabla \bar{f}(U_t) - \widetilde{g}_t$. Recall that

$$\|e_t\|_F \le 4rC_{est}\sigma_1^\star \|U_t\|_{\mathrm{op}} \sqrt{\epsilon}$$

according to Equation (23) in Lemma C.5 whenever $\|U_t\|_{\mathrm{op}} \le \Gamma^{1/2}$, which is true according to the induction hypothesis. Therefore, we have

$$\begin{aligned}
\|U_{t+1}\|_{\mathrm{op}} &= \left\| U_t - \frac{1}{16\Gamma}\widetilde{g}_t \right\|_{\mathrm{op}} = \left\| U_t - \frac{1}{16\Gamma}(2(U_tU_t - M^*)U_t - e_t) \right\|_{\mathrm{op}} \\
&\le \left\| U_t - \frac{1}{8\Gamma}U_tU_t^\top U_t \right\|_{\mathrm{op}} + \frac{1}{8\Gamma}\left\| M^*U_t + \frac{1}{2}e_t \right\|_{\mathrm{op}} \\
&\le \max_i\{\sigma_i(U_t) - \frac{1}{8\Gamma}\sigma_i^3(U_t)\} + \frac{1}{8\Gamma}(\sigma_1^\star + 2rC_{est}\sigma_1^\star\sqrt{\epsilon}) \|U_t\|_{\mathrm{op}}
\end{aligned}$$

Since the function $t \mapsto t - \frac{1}{8\Gamma}t^3$ is increasing in $[0, \sqrt{\frac{8\Gamma}{3}}]$ and $\|U_t\|_{\mathrm{op}} \le \Gamma^{1/2} \le \sqrt{\frac{8\Gamma}{3}}$ by induction hypothesis, the maximum is taken when $i = 1$, hence

$$\begin{aligned}
\|U_{t+1}\|_{\mathrm{op}} &\le \|U_t\|_{\mathrm{op}} - \frac{1}{8\Gamma}\|U_t\|_{\mathrm{op}}^3 + \frac{1}{8\Gamma}(\sigma_1^\star + 2rC_{est}\sigma_1^\star\sqrt{\epsilon}) \|U_t\|_{\mathrm{op}} \\
&\le \|U_t\|_{\mathrm{op}} - \frac{1}{8\Gamma}(\|U_t\|_{\mathrm{op}}^2 - \sigma_1^\star - 2rC_{est}\sigma_1^\star\sqrt{\epsilon}) \|U_t\|_{\mathrm{op}} \\
&\le \|U_t\|_{\mathrm{op}} - \frac{1}{8\Gamma}(\|U_t\|_{\mathrm{op}}^2 - 2\sigma_1^\star) \|U_t\|_{\mathrm{op}}, \qquad (26)
\end{aligned}$$

where the last inequality uses $\epsilon = O(\frac{1}{\kappa^3 r^3}) = O(\frac{1}{r^2})$. We split into two cases now:

Case $\|U_t\|_{\mathrm{op}} > \frac{1}{2}\Gamma^{1/2}$. Recall that $\Gamma \ge 36\sigma_1^\star$, hence

$$\begin{aligned}
\|U_{t+1}\|_{\mathrm{op}} &\le \|U_t\|_{\mathrm{op}} - \frac{1}{8\Gamma}(\|U_t\|_{\mathrm{op}}^2 - 2\sigma_1^\star) \|U_t\|_{\mathrm{op}} \\
&\le \|U_t\|_{\mathrm{op}} - \frac{1}{8\Gamma}(\frac{\Gamma}{4} - 2\sigma_1^\star)\frac{\Gamma^{1/2}}{2} \\
&\le \|U_t\|_{\mathrm{op}} - \frac{1}{8\Gamma}(\frac{\Gamma}{4} - \frac{\Gamma}{18})\frac{\Gamma^{1/2}}{2} \\
&\le \|U_t\|_{\mathrm{op}} - \frac{1}{96}\Gamma^{1/2},
\end{aligned}$$

where the second inequality is because $t \mapsto (t^2 - 2\sigma_1^\star)t$ is increasing on $\left[\sqrt{\frac{2\sigma_1^\star}{3}}, +\infty\right]$ and $\|U\|_{\mathrm{op}} \ge \frac{1}{2}\Gamma^{1/2} \ge 3\sigma_1^\star$. This case captures the dissipativity condition satisfied by the matrix sensing problem, which says that the iterate aligns with the direction of the gradient when the iterate's norm is large, so descending along the gradient decreases the norm of the iterate.

Case $\|U_t\|_{\mathrm{op}} \le \frac{1}{2}\Gamma^{1/2}$. From Equation (26) we know that if $\|U_t\|_{\mathrm{op}}^2 \ge 2\|M^*\|_{\mathrm{op}}$, it always holds that $\|U_{t+1}\|_{\mathrm{op}} \le \|U_t\|_{\mathrm{op}} \le \frac{1}{2}\Gamma^{1/2}$. For $\|U_t\|_{\mathrm{op}}^2 \le 2\|M^*\|_{\mathrm{op}}$, we have $\|U_t\|_{\mathrm{op}} \le \sqrt{2}{\sigma_1^\star}^{1/2}$, and therefore

$$\begin{aligned}
\|U_{t+1}\|_{\mathrm{op}} &\le \|U_t\|_{\mathrm{op}} - \frac{1}{8\Gamma}(\|U_t\|_{\mathrm{op}}^2 - 2\sigma_1^\star) \|U_t\|_{\mathrm{op}} \\
&\le \|U_t\|_{\mathrm{op}} + \frac{1}{8\Gamma}(\sqrt{2}{\sigma_1^\star}^{1/2} + \|U_t\|_{\mathrm{op}})((\sqrt{2}{\sigma_1^\star}^{1/2} - \|U_t\|_{\mathrm{op}})) \|U_t\|_{\mathrm{op}} \\
&\le \|U_t\|_{\mathrm{op}} + \frac{2\sqrt{2}{\sigma_1^\star}^{1/2}}{8\Gamma}(\sqrt{2}{\sigma_1^\star}^{1/2} - \|U_t\|_{\mathrm{op}})\sqrt{2}{\sigma_1^\star}^{1/2} \\
&= \|U_t\|_{\mathrm{op}} + \frac{\sigma_1^\star}{2\Gamma}(\sqrt{2}\sigma_1^\star - \|U_t\|_{\mathrm{op}}) \\
&\le \sqrt{2}{\sigma_1^\star}^{1/2} \qquad \text{(because the above expression is increasing in } \|U_t\|_{\mathrm{op}}) \\
&\le \frac{\sqrt{2}}{6}\Gamma^{1/2} < \Gamma^{1/2},
\end{aligned}$$

where in the second last inequality we use $\sigma_1^\star \le \Gamma/36$.

In summary, for all $t \in \mathbb{N}$,

$$\|U_t\|_{\mathrm{op}} > \frac{1}{2}\Gamma^{1/2} \implies \left\|U_t - \frac{1}{16\Gamma}\widetilde{g}_t\right\|_{\mathrm{op}} \le \|U_t\|_{\mathrm{op}} - \frac{1}{96}\Gamma^{1/2} \tag{27}$$

$$\|U_t\|_{\mathrm{op}} \le \frac{1}{2}\Gamma^{1/2} \implies \left\|U_t - \frac{1}{16\Gamma}\widetilde{g}_t\right\|_{\mathrm{op}} \le \frac{1}{2}\Gamma^{1/2} \tag{28}$$

In particular, for gradient step regime $U_{t+1} := U_t - \frac{1}{16\Gamma}\widetilde{g}_t$, we know the iterate U_{t+1} stays inside the region $\{U : \|U\|_{\mathrm{op}}^2 \le \Gamma\}$ provided that U_t is inside the region.

We proceed to analyze negative curvature steps, which only happen if the inexact gradient is small $\|\widetilde{g}_t\|_F \le \epsilon_g = \frac{1}{32}\sigma_r^{\star 3/2}$. Note that it follows from Equation (27) that

$$\begin{aligned}
\|\widetilde{g}_t\|_F \le \frac{1}{32}\sigma_r^{\star 3/2} &\implies \|\widetilde{g}_t\|_{\mathrm{op}} \le \frac{1}{32}\sigma_r^{\star 3/2} \le \frac{1}{32}\Gamma^{3/2} \\
&\implies \left\|U_t - \frac{1}{16\Gamma}\widetilde{g}_t\right\|_{\mathrm{op}} \ge \|U_t\|_{\mathrm{op}} - \frac{1}{512}\Gamma^{1/2} > \|U_t\|_{\mathrm{op}} - \frac{1}{96}\Gamma^{1/2} \\
&\implies \|U_t\|_{\mathrm{op}} \le \frac{1}{2}\Gamma^{1/2}.
\end{aligned}$$

Therefore, if at step τ we run the inexact negative curvature step to update the iterate from U_τ to $U_{\tau+1}$, then $\|U_\tau\|_{\mathrm{op}} \le \frac{1}{2}\Gamma^{1/2}$. Recall in Equation (25) the negative curvature stepsize $\frac{2\epsilon_H}{L_H} \le \frac{\Gamma^{1/2}}{36} < \frac{\Gamma^{1/2}}{2}$, so for the negative curvature update

$$U_{\tau+1} = U_\tau + \frac{2\epsilon_H}{L_H}\sigma_\tau \widetilde{p}_\tau$$

where $\sigma_\tau = \pm 1$ with probability 1/2 and p_τ is a vector with $\|p_\tau\|_F = 1$, we have

$$\begin{aligned}
\|U_{\tau+1}\|_{\mathrm{op}} &\le \|U_\tau\|_{\mathrm{op}} + \left\|\frac{2\epsilon_H}{L_H}\sigma_\tau\widetilde{p}_\tau\right\|_{\mathrm{op}} \\
&\le \|U_\tau\|_{\mathrm{op}} + \left\|\frac{2\epsilon_H}{L_H}\sigma_\tau\widetilde{p}_\tau\right\|_F < \frac{1}{2}\Gamma^{1/2} + \frac{1}{2}\Gamma^{1/2} = \Gamma^{1/2}.
\end{aligned}$$

Finally, the initialization satisfies $\|U_0\|_{\mathrm{op}}^2 \le \Gamma$ by assumption, so the induction is complete. □

C.3 Local Linear Convergence in Low Rank Matrix Sensing—Omitted Proofs

Recall that Theorem 3.5 gives us a $(\frac{1}{24}\sigma_r^{\star 3/2}, \frac{1}{3}\sigma_r^\star)$-approximate SOSP as the initialization for Algorithm 3.1. We now explain why approximate SOSP is useful for the local search in the low rank matrix sensing problems.

Definition C.6 (Strict Saddle Property)**.** Function $f(\cdot)$ is a $(\epsilon_g, \epsilon_H, \zeta)$-strict saddle function if for any $U \in \mathbb{R}^{r\times d}$, at least one of following properties holds:

(a) $\|\nabla f(U)\| > \epsilon_g$.

(b) $\lambda_{\min}(\nabla^2 f(U)) < -\epsilon_H$.

(c) U is ζ-close to $\mathcal{X}^\star$ — the set of local minima; namely, $\mathrm{dist}(U, \mathcal{X}^\star) \le \zeta$.

Approximate SOSPs are close to the set of local minima for strict saddle functions:

Proposition C.7. *Let f be a $(\epsilon_g, \epsilon_H, \zeta)$-strict saddle function and $\mathcal{X}^\star$ be the set of its local minima. If U_{SOSP} is a (ϵ_g, ϵ_H)-approximate SOSP of f, then* $\mathrm{dist}(U_{SOSP}, \mathcal{X}^\star) \le \zeta$.

Proof. Since U_{SOSP} is a (ϵ_g, ϵ_H)-approximate SOSP of f, we have $\|\nabla f(U)\| \le \epsilon_g$ and $\lambda_{\min}(\nabla^2 f(U)) \ge -\epsilon_H$, i.e., (a) and (b) in Definition C.6 fail to hold. Therefore, (c) holds, i.e., $\mathrm{dist}(U, \mathcal{X}^\star) \le \zeta$. □

Some regularity conditions only hold in a local neighborhood of $\mathcal{X}^\star$. We focus on the following definition.

Definition C.8 (Local Regularity Condition)**.** In a ζ-neighborhood of the set of local minima $\mathcal{X}^\star$ (that is, $\operatorname{dist}(U, \mathcal{X}^\star) \le \zeta$), the function $f(\cdot)$ satisfies an (α, β)-regularity condition if for any U in this neighborhood:

$$\langle \nabla f(U), U - \mathcal{P}_{\mathcal{X}^\star}(U) \rangle \ge \frac{\alpha}{2} \left\| U - \mathcal{P}_{\mathcal{X}^\star}(U) \right\|^2 + \frac{1}{2\beta} \left\| \nabla f(U) \right\|^2 . \tag{29}$$

Local regularity condition is a weaker condition than strong convexity, but both conditions would allow a first-order algorithm to obtain local linear convergence.

Recall that $\bar{f}(U) = \frac{1}{2} \left\| UU^\top - M^* \right\|_F^2$. The global minima of $\bar{f}$ solve the so-called symmetric low rank matrix factorization problem and the corresponding function value is zero. Let $TDT^\top$ be the singular value decomposition of the real symmetric matrix M^*. We observe $U^* = TD^{1/2}$ is a global optimum of $\bar{f}$.

The function $\bar{f}$ satisfies the local regularity condition and the strict saddle property, and all of its local minima are global minima [Jin+17].

Fact C.9 ([Jin+17], Lemma 7)**.** *For $\bar{f}$ defined in (4), all local minima are global minima. The set of global minima is characterized by $\mathcal{X}^\star = \{U^* R | RR^\top = R^\top R = I\}$. Furthermore, $\bar{f}(U)$ satisfies:*

1. *$(\frac{1}{24}(\sigma_r^\star)^{3/2}, \frac{1}{3}\sigma_r^\star, \frac{1}{3}(\sigma_r^\star)^{1/2})$-strict saddle property, and*
2. *$(\frac{2}{3}\sigma_r^\star, 10\sigma_1^\star)$-regularity condition in the $\frac{1}{3}(\sigma_r^\star)^{1/2}$-neighborhood of $\mathcal{X}^\star$ in Frobenius norm.*

Similar to Corollary C.5, we need a theorem to say that all calls to RobustMeanEstimation (Algorithm A.2) are successful. For the local search (Algorithm 3.1), only inexact gradient oracles are required. Since gradients have dimension dr, the sample complexity needed for all robust estimates of gradient to be accurate with high probability is $\widetilde{O}\left(\frac{dr}{\epsilon}\right)$.

Corollary C.10. *Consider the noiseless setting as in Theorem 1.7 with $\sigma = 0$. There exists a sample size $n = \widetilde{O}\left(\frac{dr}{\epsilon}\right)$ such that with n samples, all subroutine calls to* RobustMeanEstimation *(Algorithm A.2) to estimate gradients $\nabla \bar{f}(U_t)$ succeed with high probability. In light of Equation (6), this means that there exists a large enough constant C_{est} such that with high probability which is suppresed by $\widetilde{O}(\cdot)$, for all iterations t, we have*

$$\left\| \widetilde{g}_t - \nabla \bar{f}(U_t) \right\|_F \le 2C_{est} \left\| U_t U_t^\top - M^* \right\|_F \left\| U_t \right\|_{\mathrm{op}} . \tag{30}$$

Proof. The proof follows exactly the same line of argument as Corollary C.5 and Theorem B.3, which discusses the sample complexity required for robust mean estimation of both gradients and Hessians to be successful. Since the Hessian has dimension d^2r^2, the sample complexity in Corollary C.5 is dominated by $\widetilde{\Omega}\left(d^2r^2/\epsilon\right)$. Here the gradient has dimension dr, so we have the required sample complexity $\widetilde{\Omega}(dr/\epsilon)$. □

Proof of Theorem 3.8. Let $e_t = \widetilde{g}_t - \nabla \bar{f}(U_t)$ and we rewrite the inexact gradient descent update as

$$U_{t+1} = U_t - \eta(\nabla \bar{f}(U_t) + e_t).$$

It follows from Corollary C.10 that, with high probability and for some large constant C_{est}, all calls to **RobustMeanEstimation** are successful, i.e., for all iterations t,

$$\left\| e_t \right\|_F \le 2C_{est}\Gamma^{1/2} \left\| U_t U_t^\top - M^* \right\|_F \sqrt{\epsilon}. \tag{31}$$

We condition on this event for the rest of the analysis.

It is straightforward to check that $\frac{1}{3}\sigma_r^{\star 1/2}$-neighborhood of $\mathcal{X}^\star$ in Frobenius norm is inside the region $\{U : \|U\|_{\mathrm{op}}^2 \le \Gamma\}$. By strict saddle property in Fact C.9, we know from Proposition C.7 that $U_0 = U_{SOSP}$ is $\frac{1}{3}\sigma_r^{\star 1/2}$-close to $\mathcal{X}^\star$ in Frobenius norm. We assume for the sake of induction that U_t is $\frac{1}{3}\sigma_r^{\star 1/2}$-close to $\mathcal{X}^\star$.

Let $\mathcal{P}_{\mathcal{X}^\star}(U)$ be the Frobenius projection of $U \in \mathbb{R}^{d\times r}$ onto $\mathcal{X}^\star$. By Fact C.9, for a given $U \in \mathbb{R}^{d\times r}$, there exists a rotation R_U so that $\mathcal{P}_{\mathcal{X}^\star}(U) = U^* R_U$. Therefore, $M^* = U^* U^{*\top} = U^* R_U R_U^\top U^{*\top} = \mathcal{P}_{\mathcal{X}^\star}(U)\mathcal{P}_{\mathcal{X}^\star}(U)^\top$. We have

$$\begin{aligned} UU^\top - M^* &= UU^\top - U\mathcal{P}_{\mathcal{X}^\star}(U)^\top + U\mathcal{P}_{\mathcal{X}^\star}(U)^\top - \mathcal{P}_{\mathcal{X}^\star}(U)\mathcal{P}_{\mathcal{X}^\star}(U)^\top \\ &= U(U - \mathcal{P}_{\mathcal{X}^\star}(U))^\top + (U - \mathcal{P}_{\mathcal{X}^\star}(U))\mathcal{P}_{\mathcal{X}^\star}(U)^\top. \end{aligned}$$

Since $\|U_t\|_{\mathrm{op}} \le \Gamma^{1/2}$ and $\|\mathcal{P}_{\mathcal{X}^\star}(U)\|_{\mathrm{op}} \le \Gamma^{1/2}$, it follows from (31) that

$$\|e_t\|_F \le 2C_{est}\Gamma^{1/2}\sqrt{\epsilon}\left\|U_tU_t^\top - M^*\right\|_F \le 4C_{est}\Gamma\sqrt{\epsilon}\left\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\right\|_F.$$

We proceed to derive the exponential decrease of the distance to global minima.

$$\begin{aligned} &\|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_{t+1})\|_F^2 \\ &\le \|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 \\ &= \left\|U_t - \eta(\nabla\bar{f}(U_t) + e_t) - \mathcal{P}_{\mathcal{X}^\star}(U_t)\right\|_F^2 \\ &= \|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + \eta^2\left\|\nabla\bar{f}(U_t) + e_t\right\|_F^2 - 2\eta\langle\nabla\bar{f}(U_t) + e_t, U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\rangle \\ &\le (1-\eta\alpha)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + \eta^2\left\|\nabla\bar{f}(U_t) + e_t\right\|_F^2 - \frac{\eta}{\beta}\left\|\nabla\bar{f}(U_t)\right\|_F^2 - 2\eta\langle e_t, U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\rangle, \end{aligned}$$

where the last inequality comes from the regularity condition in Fact C.9 with $\alpha = \frac{2}{3}\sigma_r^\star, \beta = 10\sigma_1^\star$.

The cross term can be bounded by

$$-2\eta\langle e_t, U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\rangle \le 2\eta\|e_t\|_F\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F \le 8C_{est}\eta\Gamma\sqrt{\epsilon}\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2,$$

where the last inequality comes from Equation 31.

By Young's inequality,

$$\begin{aligned} \eta^2\left\|\nabla\bar{f}(U_t) + e_t\right\|_F^2 &\le 2\eta^2\left\|\nabla\bar{f}(U_t)\right\|_F^2 + 2\eta^2\|e_t\|_F^2 \\ &\le 2\eta^2\left\|\nabla\bar{f}(U_t)\right\|_F^2 + 32\eta^2C_{est}^2\Gamma^2\epsilon\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2. \end{aligned}$$

Recall that $\Gamma \ge 36\sigma_1^\star$. Choosing $\eta = \frac{1}{\Gamma} \le \frac{1}{36\sigma_1^\star} < \frac{1}{20\sigma_1^\star} = \frac{1}{2\beta}$, it follows that

$$\begin{aligned} &\|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_{t+1})\|_F^2 \\ &\le (1-\eta\alpha)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + \eta^2\left\|\nabla\bar{f}(U_t) + e_t\right\|_F^2 - \frac{\eta}{\beta}\left\|\nabla\bar{f}(U_t)\right\|_F^2 - 2\eta\langle e_t, U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\rangle \\ &\le (1 - \eta\alpha + 8C_{est}\eta\Gamma\sqrt{\epsilon} + 32\eta^2C_{est}^2\Gamma^2\epsilon)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + (2\eta^2 - \frac{\eta}{\beta})\left\|\nabla\bar{f}(U_t)\right\|_F^2 \\ &\le \left(1 - \frac{\sigma_r^\star}{30\Gamma} + \frac{2}{5}C_{est}\sqrt{\epsilon} + 32(C_{est}/20)^2\epsilon\right)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 \\ &\le \left(1 - \left(O\left(\frac{1}{\kappa}\right) - O(\sqrt{\epsilon})\right)\right)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2. \end{aligned} \tag{32}$$

One consequence of this calculation is $\|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_{t+1})\|_F \le \|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F$, so U_{t+1} is also $\frac{1}{3}\sigma_r^{\star 1/2}$-close to $\mathcal{X}^\star$ in Frobenius norm, completing the induction. The other consequence is local linear convergence of U_t to $\mathcal{X}^\star$ with rate $\left(1 - O\left(\frac{1}{\kappa}\right)\right)$, assuming $\epsilon\kappa^2$ is sufficiently small. Since the initial distance is bounded by $O(\sigma_r^{\star 1/2})$, converging to a point that is ι-close to $\mathcal{X}^\star$ in Frobenius norm requires the following number of iterations:

$$O\left(\frac{\log(\iota/\sigma_r^{\star 1/2})}{\log(1 - O(1/\kappa))}\right) = O\left(\kappa\log\left(\frac{\sigma_r^\star}{\iota}\right)\right). \qquad \square$$

C.4 Low Rank Matrix Sensing with Noise

In Section 3.1, we focused on the case that $\sigma = 0$ as in Theorem 3.2. Now we consider the case that $\sigma \neq 0$ and prove Theorem 3.3.

Recall that $y_i \sim \mathcal{N}(\langle A_i, M^* \rangle, \sigma^2)$ in Definition 1.6. Since $\langle A_i, M^* \rangle$ follows Gaussian distribution with mean 0 and variance $\|M^*\|_F^2$, we optionally assume $\sigma = O(\|M^*\|_F) = O(r\Gamma)$ to keep the signal-to-noise ratio in constant order. For the ease of presentation, we make the following assumption:

Assumption C.11. *Assume $\sigma \leq r\Gamma$.*

We present different algorithms depending on whether Assumption C.11 holds.

As in the noiseless case, we define $f_i(U) = \frac{1}{2}\left(\langle UU^\top, A_i \rangle - y_i\right)^2$ and

$$\bar{f}(U) = \mathop{\mathbb{E}}_{(A_i, y_i) \sim \mathcal{G}_\sigma} f_i(U) = \frac{1}{2}\left\|UU^\top - M^*\right\|_F^2 + \frac{1}{2}\sigma^2.$$

Note that the $\frac{1}{2}\sigma^2$ term in $\bar{f}(U)$ has no effect on its minimum, gradients, or Hessians, and Fact C.9 and 3.4 still apply in verbatim.

We prove the following result when Assumption C.11 holds:

Theorem C.12. *Consider the same setting as in Theorem 1.7 with $0 \neq \sigma \leq r\Gamma$ (Assumption C.11 holds). There exists a sample size $n = \widetilde{O}(d^2 r^2/\epsilon)$ such that with high probability, there exists an algorithm that outputs a solution $\widehat{M}$ in $\widetilde{O}(r^2\kappa^3)$ calls to robust mean estimation routine A.2 with error $\|\widehat{M} - M^*\|_F = O(\kappa\sigma\sqrt{\epsilon})$.*

Proof. We compute the gradient and Hessian of f_i:

$$\begin{aligned}
\nabla f_i(U) &= \left(\langle UU^\top - M^*, A_i \rangle - \zeta_i\right)(A_i + A_i^\top)U \\
H_i(U) &= \left(\langle UU^\top - M^*, A_i \rangle - \zeta_i\right) I_r \otimes B_i + \operatorname{vec}(B_i U)\operatorname{vec}(B_i U)^\top.
\end{aligned}$$

We also need to bound the covariance of sample gradients and sample Hessians in the bounded region $\{U \in \mathbb{R}^{d\times r} : \|U\|_{\mathrm{op}}^2 \leq \Gamma\}$, similar to Equations (6) and (7).

$$\left\|\operatorname{Cov}(\operatorname{vec}(\nabla f_i(U)))\right\|_{\mathrm{op}} \leq 4\left(\left\|UU^\top - M^*\right\|_F^2 + \sigma^2\right)\|U\|_{\mathrm{op}}^2 = O\left(r^2\Gamma^3\right). \tag{33}$$

$$\left\|\operatorname{Cov}(\operatorname{vec}(H_i))\right\|_{\mathrm{op}} \leq 16r\left(\left\|UU^\top - M^*\right\|_F^2 + \sigma^2\right) + 128\|U\|_{\mathrm{op}}^4 = O\left(r^3\Gamma^2\right). \tag{34}$$

Since global convergence analysis in Section 3.1.2 only requires $\left\|\operatorname{Cov}(\operatorname{vec}(\nabla f_i(U)))\right\|_{\mathrm{op}} = O\left(r^2\Gamma^3\right)$ and $\left\|\operatorname{Cov}(\operatorname{vec}(H_i))\right\|_{\mathrm{op}} = O\left(r^3\Gamma^2\right)$, Theorem 3.5 also holds in the noisy setting (up to a constant factor). This means that we could obtain a $(\frac{1}{24}\sigma_r^{\star 3/2}, \frac{1}{3}\sigma_r^\star)$-approximate SOSP of $\bar{f}$, which is in $\frac{1}{3}\sigma_r^{\star 1/2}$-neighborhood of $\mathcal{X}^\star$ in Frobenius norm.

We now focus on the local convergence analysis. We use the same algorithm (Algorithm 3.1), which uses RobustMeanEstimation subroutine in each iteration t to obtain an inexact gradient $\widetilde{g}_t$ for $\bar{f}(U)$. We rewrite the inexact gradient descent update as

$$U_{t+1} = U_t - \eta(\nabla\bar{f}(U_t) + e_t).$$

We condition the remaining analysis on the event that all calls to **RobustMeanEstimation** are successful as in Corollary C.10, which happens with high probability and implies

$$\begin{aligned}
\|e_t\|_F &\leq 2C_{est}\left(\left\|U_t U_t^\top - M^*\right\|_F + \sigma\right)\|U_t\|_{\mathrm{op}}\sqrt{\epsilon} \\
&\leq 2C_{est}\left(\left\|U_t U_t^\top - M^*\right\|_F + \sigma\right)\Gamma^{1/2}\sqrt{\epsilon}.
\end{aligned}$$

It is straightforward to check that $\frac{1}{3}\sigma_r^{\star 1/2}$-neighborhood of $\mathcal{X}^\star$ in Frobenius norm is inside the region $\{U : \|U\|_{\mathrm{op}}^2 \leq \Gamma\}$. By strict saddle property in Fact C.9, we know $U_0 = U_{SOSP}$ is $\frac{1}{3}\sigma_r^{\star 1/2}$-close

to $\mathcal{X}^\star$ in Frobenius norm. We assume for the sake of induction that U_t is $\frac{1}{3}\sigma_r^{\star 1/2}$-close to $\mathcal{X}^\star$. We will show that $\|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_{t+1})\|_F$ either shrinks or is at the order of $O\big(\frac{\sigma\epsilon}{\Gamma^{1/2}}\big)$, which is also $\frac{1}{3}\sigma_r^{\star 1/2}$-close to $\mathcal{X}^\star$ in Frobenius norm for sufficiently small $\epsilon = O\big(\frac{1}{r\kappa}\big)$.

As before, we define $\mathcal{P}_{\mathcal{X}^\star}(\cdot)$ to be the projection onto $\mathcal{X}^\star$ and bound the Frobenius distance to $\mathcal{X}^\star$ as follows:

$$
\begin{aligned}
&\|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_{t+1})\|_F^2 \\
&\le \|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 \\
&= \left\|U_t - \eta(\nabla \bar{f}(U_t) + e_t) - \mathcal{P}_{\mathcal{X}^\star}(U_t)\right\|_F^2 \\
&= \|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + \eta^2 \left\|\nabla \bar{f}(U_t) + e_t\right\|_F^2 - 2\eta\langle \nabla \bar{f}(U_t) + e_t, U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\rangle \\
&\le (1-\eta\alpha)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + \eta^2 \left\|\nabla \bar{f}(U_t) + e_t\right\|_F^2 - \frac{\eta}{\beta}\left\|\nabla \bar{f}(U_t)\right\|_F^2 - 2\eta\langle e_t, U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\rangle,
\end{aligned}
$$

where the last inequality comes from the regularity condition in Fact C.9 with $\alpha = \frac{2}{3}\sigma_r^\star, \beta = 10\sigma_1^\star$.

Using Cauchy-Schwarz inequality to bound $\left\|\nabla \bar{f}(U_t) + e_t\right\|_F^2$ and $\langle e_t, U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\rangle$ and choosing $\eta = \frac{1}{20\Gamma} \le \frac{1}{20\sigma_1^\star} = \frac{1}{2\beta}$ to kill the $\left\|\nabla \bar{f}(U_t)\right\|_F^2$ term, we obtain

$$
\begin{aligned}
&\|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_{t+1})\|_F^2 \\
&\le (1-\eta\alpha)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + 2\eta^2\left\|\nabla \bar{f}(U_t)\right\|_F^2 + 2\|e_t\|_F^2 - \frac{\eta}{\beta}\left\|\nabla \bar{f}(U_t)\right\|_F^2 - 2\eta\langle e_t, U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\rangle \\
&\le (1-\eta\alpha)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + 2\eta^2\|e_t\|_F^2 - 2\eta\langle e_t, U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\rangle + (2\eta^2 - \frac{\eta}{\beta})\left\|\nabla \bar{f}(U_t)\right\|_F^2 \\
&\le (1-\eta\alpha)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + 2\eta^2\|e_t\|_F^2 + 2\eta\|e_t\|_F\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F.
\end{aligned}
$$

Now we consider the following two possible cases:

Case 1: $\left\|U_tU_t^\top - M^*\right\|_F > \sigma$. Then

$$
\begin{aligned}
\|e_t\|_F &\le 2C_{est}\left(\left\|U_tU_t^\top - M^*\right\|_F + \sigma\right)\Gamma^{1/2}\sqrt{\epsilon} \\
&< 4C_{est}\left\|U_tU_t^\top - M^*\right\|_F\Gamma^{1/2}\sqrt{\epsilon}.
\end{aligned}
$$

And it follows similar to (32) that

$$
\begin{aligned}
&\|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_{t+1})\|_F^2 \\
&\le (1 - \eta\alpha + 8C_{est}\eta\Gamma\sqrt{\epsilon} + 32\eta^2C_{est}^2\Gamma^2\epsilon)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 \\
&\le \left(1 - \frac{\sigma_r^\star}{30\Gamma} + \frac{2}{5}C_{est}\sqrt{\epsilon} + 32(C_{est}/20)^2\epsilon\right)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 \\
&\le \left(1 - \left(O\left(\frac{1}{\kappa}\right) - O(\sqrt{\epsilon})\right)\right)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2.
\end{aligned}
$$

We conclude that the distance to $\mathcal{X}^\star$ decreases geometrically in this regime until $\left\|U_tU_t^\top - M^*\right\|_F \le \sigma$, which takes a total of $O\Big(\kappa\log\left(\frac{\sigma_r^\star}{\sigma}\right)\Big)$ iterations, which is very few if $\sigma = \Theta(r\Gamma)$.

Case 2: $\left\|U_tU_t^\top - M^*\right\|_F \le \sigma$. Then $\|e_t\|_F \le 4C_{est}\sigma\Gamma^{1/2}\sqrt{\epsilon}$. It follows that

$$
\begin{aligned}
&\|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_{t+1})\|_F^2 \\
&\le (1-\eta\alpha)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + 2\eta^2\|e_t\|_F^2 + 2\eta\|e_t\|_F\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F \\
&\le (1-\eta\alpha)\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F^2 + 32\eta^2C_{est}^2\sigma^2\Gamma\epsilon + 8\eta C_{est}\sigma\Gamma^{1/2}\sqrt{\epsilon}\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F.
\end{aligned}
$$

Write $\mathcal{C} = 4\sqrt{2}\eta C_{est}\sigma\Gamma^{1/2}\sqrt{\epsilon}$, $\mathcal{B}_t = \|U_{t+1} - \mathcal{P}_{\mathcal{X}^\star}(U_{t+1})\|_F/\mathcal{C}$, we have

$$
\mathcal{B}_{t+1} \le \sqrt{(1-\eta\alpha)\mathcal{B}_t^2 + \mathcal{B}_t + 1}.
$$

It is straightforward to compute that the function $b \mapsto \sqrt{(1-\eta\alpha)b^2 + b + 1}$ has Lipschitz constant

$$\rho = \begin{cases} \frac{1}{2} & \text{if } 0 < 1-\eta\alpha \le \frac{1}{4} \\ \sqrt{1-\eta\alpha} & \text{if } \frac{1}{4} \le 1-\eta\alpha < 1. \end{cases}$$

Since $\rho < 1$, Banach fixed-point theorem implies that $\mathcal{B}_t$ converges to some $\mathcal{B}^*$ with rate

$$|\mathcal{B}_{t+1} - \mathcal{B}^*| \le \rho \, |\mathcal{B}_t - \mathcal{B}^*| \, .$$

We can compute an upper bound on the fixed point $\mathcal{B}^*$ by solving

$$\mathcal{B}^* = \sqrt{(1-\eta\alpha)\mathcal{B}^{*2} + \mathcal{B}^* + 1},$$

which gives $\mathcal{B}^* = O\Big(\frac{1}{\eta\alpha}\Big) = O(\kappa)$.

Denote the first iteration of Case 2 regime ($\left\|U_t U_t^\top - M^*\right\|_F \le \sigma$) by t_0. By [GJZ17, Lemma 6], we have $\|U_{t_0} - \mathcal{P}_{\mathcal{X}^\star} U_{t_0}\|_F \le \sigma\sqrt{\frac{2}{\sigma_r^\star}}$ and therefore $\mathcal{B}_{t_0} = \sqrt{\kappa/\epsilon}$. Therefore, Case 2 takes a total of $O(\kappa \log 1/\epsilon)$ iterations to reach the error

$$\|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F \le O\Big(\kappa\Gamma^{-1/2}\sigma\sqrt{\epsilon}\Big) \, .$$

This translates to the error in the matrix space

$$\left\|U_t U_t^\top - M^*\right\|_F \le \Gamma^{1/2} \|U_t - \mathcal{P}_{\mathcal{X}^\star}(U_t)\|_F = O\big(\kappa\sigma\sqrt{\epsilon}\big) \, .$$

□

Assumption C.11 might fail to hold; a notable example is our construction in Section 4, where $\|M^*\|_F = O\big(\epsilon^{1/2}\big)$ but $\sigma = O(1)$. In this case, we consider a simple spectral method, relying on the observation that with clean samples $\mathbb{E}_{(A_i,y_i)\in\mathcal{G}_\sigma}[y_i A_i] = M^*$. So we run RobustMeanEstimation on $\{y_i A_i\}_{i=1}^n$ to get $\widetilde{M}$ with its singular decomposition $\widetilde{M} = \sum_{i=1}^n s_i u_i v_i^\top$, where singular values $s_1 \ge s_2 \ge \ldots \ge s_n$ are descending. We then form a new matrix $\widehat{M}$ from the r leading singular vectors, i.e., $\widehat{M} = \sum_{i=1}^r s_i u_i v_i^\top$.

This simple algorithm has the following guarantee:

Theorem C.13. *Consider the same setting as in Theorem I.7 with $\sigma > r\Gamma$, i.e., Assumption C.11 fails. There exists a sample size $n = \widetilde{O}(d^2/\epsilon)$ such that with high probability, there exists an algorithm that outputs a solution $\widehat{M}$ in one call to robust mean estimation routine A.2 with error $\|\widehat{M} - M^*\|_F = O(\sigma\sqrt{\epsilon})$.*

Proof. We compute the mean and variance of $\langle y_i A_i, Z\rangle$ for some unit Frobenius norm Z.

$$\begin{aligned} \mathbb{E}\langle y_i A_i, Z\rangle &= \mathbb{E}\langle A_i, M^*\rangle\langle A_i, Z\rangle + \mathbb{E}\,\zeta\langle A_i, Z\rangle \\ &= \mathbb{E}\langle A_i, M^*\rangle\langle A_i, Z\rangle = \langle M^*, Z\rangle. \end{aligned}$$

Hence $\mathbb{E}[y_i A_i] = M^*$.

$$\begin{aligned} &\mathrm{Var}\left(\langle A_i, M^*\rangle\langle A_i, Z\rangle + \zeta\langle A_i, Z\rangle\right) \\ &\le 2\,\mathrm{Var}\langle A_i, M^*\rangle\langle A_i, Z\rangle + 2\,\mathrm{Var}(\zeta\langle A_i, Z\rangle) \\ &\le 4\,\|M^*\|_F^2 + 4\sigma^2 \le 8\sigma^2, \end{aligned}$$

where the last inequality is because $\|M^*\|_F \le r\Gamma$. Hence running robust mean estimation algorithm on $\{y_i A_i\}$ outputs $\widetilde{M}$ with $\left\|\widetilde{M} - M^*\right\|_F = O(\sigma\sqrt{\epsilon})$.

Since $\widehat{M}$ is formed by the r leading singular vectors of $\widetilde{M}$, it is the best rank r approximation to $\widetilde{M}$ by Eckart–Young–Mirsky's theorem, i.e., $\left\|\widehat{M} - \widetilde{M}\right\|_F \le \left\|M^* - \widetilde{M}\right\|_F = O(\sigma\sqrt{\epsilon})$. We conclude that $\left\|\widehat{M} - M^*\right\|_F \le \left\|\widehat{M} - \widetilde{M}\right\|_F + \left\|\widetilde{M} - M^*\right\|_F = O(\sigma\sqrt{\epsilon})$ by triangle inequality. □

D SQ Lower bound—Omitted Proofs

We consider a weaker corruption model, known as Huber's ϵ-contamination model [Hub64], than Definition 1.1 in the sense that any corruptions in Huber's contamination model can be emulated by the adversary in Definition 1.1. We show a lower bound of the sample complexity in the presence of Huber's ϵ-contamination, and the sample complexity under Definition 1.1 can therefore be no better.

Assumption D.1 (Huber's Contamination Model)**.** *Let Q be the joint distribution of (A, y) with $\mathrm{vec}(A) \sim \mathcal{N}(0, I_{d^2})$ and $y|A \sim \mathcal{N}(\langle M^*, uu^\top\rangle, \sigma^2)$. The input samples are contaminated according to Huber's ϵ-contamination model in the following way: we observe samples from a mixture Q' of the clean joint distribution Q and an arbitrary noise distribution N, i.e., $Q' = (1-\epsilon)Q + \epsilon N$.*

The adversary in the ϵ-strong contamination model can emulate the Huber's adversary: on input clean samples, the ϵ-strong contamination adversary removes a randomly chosen ϵ-fraction of samples and replaces them with samples drawn i.i.d. from N.

Recall that SQ algorithms are a class of algorithms that, instead of access to samples, are allowed to query expectations of bounded functions of the distribution.

Definition D.2 (SQ Algorithms and STAT Oracle [Kea98])**.** Let $\mathcal{G}$ be a distribution on $\mathbb{R}^{d^2+1}$. A Statistical Query (SQ) is a bounded function $q : \mathbb{R}^{d^2+1} \to [-1, 1]$. For $\tau > 0$, the $\mathsf{STAT}(\tau)$ oracle responds to the query q with a value v such that $|v - \mathbb{E}_{X\sim D}[q(X)]| \le \tau$. An SQ algorithm is an algorithm whose objective is to learn some information about an unknown distribution $\mathcal{G}$ by making adaptive calls to the corresponding $\mathsf{STAT}(\tau)$ oracle.

A statistical query with accuracy τ can be implemented with error probability δ by taking $O(\log(1/\delta)/\tau^2)$ samples and evaluating the empirical mean of the query function $q(\cdot)$ evaluated at those samples [DK23, Chapter 8.2.1]. This bound is tight in general for a single query.

Lemma D.3 (Clean marginal and conditional distribution)**.** *Let $A \in \mathbb{R}^{d\times d}$ with $\mathrm{vec}(A) \sim \mathcal{N}(0, I_{d^2})$ and $y|A \sim \mathcal{N}(\langle uu^\top, M^*\rangle, \sigma^2)$. Then*

$$y \sim \mathcal{N}(0, \sigma^2 + \|u\|^4)$$
$$\mathrm{vec}(A)|y \sim \mathcal{N}\left(\frac{\mathrm{vec}(uu^\top)y}{\sigma^2 + \|u\|^4}, I_{d^2} - \frac{\mathrm{vec}(uu^\top)\,\mathrm{vec}(uu^\top)^\top}{\sigma^2 + \|u\|^4}\right).$$

Proof. Let Q be the joint distribution of $(\mathrm{vec}(A), y)$. We can write down the covariance of Q:

$$\begin{bmatrix} I_{d^2} & \mathrm{vec}(uu^\top) \\ \mathrm{vec}(uu^\top)^\top & \sigma^2 + \|u\|^4 \end{bmatrix}.$$

The mean and variance of $\mathrm{vec}(A)|y$ follow from the conditional Gaussian formula. □

We will construct a family of contaminated joint distributions of (A, y) that are hard to learn with SQ access. We first construct a family of clean distributions and then construct the contaminations.

Lemma D.4 (Defining Clean Distributions)**.** *Let v be a unit vector and set $u = c_1^{1/2}\epsilon^{1/4}v$ for some small constant c_1. Let $\mathrm{vec}(A) \sim \mathcal{N}(0, I_{d^2})$, and choose $\sigma = 1 - c_1^2\epsilon$. Then we have the marginal distribution $y \sim \mathcal{N}(0, 1)$ and*

$$\mathrm{vec}(A)|y \sim \mathcal{N}\left(c_1\sqrt{\epsilon}\,\mathrm{vec}(vv^\top)y, I_{d^2} - c_1^2\epsilon\,\mathrm{vec}(vv^\top)\,\mathrm{vec}(vv^\top)^\top\right).$$

Let Q_v denote the joint distribution from the construction in Lemma D.4. The family of contaminated joint distributions that are hard to learn will be denoted by Q'_v, and we proceed to construct them. The technique for the construction follows a standard non-Gaussian component analysis argument, which starts from mixing a one-dimensional distribution to match moments with the standard Gaussian. We note that the conditional distribution $\mathrm{vec}(A)|y$ in all directions perpendicular to $\mathrm{vec}(vv^\top)$ is already standard Gaussian, so we proceed to consider the one-dimensional distribution along $\mathrm{vec}(vv^\top)$ direction, which is

$$\mathcal{N}\left(c_1\sqrt{\epsilon}y, 1 - c_1^2\epsilon\right).$$

Let $\mu(y) = c_1\sqrt{\epsilon}y$ be a shortand for the mean of this one dimensional distribution. We proceed to mix it with some noise distribution B_μ. It is important to note here that the fraction of the noise depends on μ and thus on y—the larger μ is, the higher the noise level is needed.

Definition D.5. Let P, Q be two distributions with probability density functions $P(x), Q(x)$; we obscure the line between a distribution and its density function. The χ^2-divergence of P and Q is

$$\chi^2(P,Q) = \int \frac{(P(x)-Q(x))^2}{Q(x)}dx = \int \frac{P^2(x)}{Q(x)}dx - 1.$$

Let N be a distribution whose support contains the supports of P and Q. We define

$$\chi_N(P,Q) = \int \frac{P(x)Q(x)}{N(x)}dx.$$

Then $|\chi_N(P,Q)-1| = \int \frac{(P(x)-N(x))(Q(x)-N(x))}{N(x)}dx$ defines an inner product of P and Q.

Lemma D.6 (One-Dimensional Moment Matching)**.** *For any $\epsilon > 0, \mu \in \mathbb{R}$, there is a distribution D_μ such that D_μ agrees with the mean of $\mathcal{N}(0,1)$ and*

$$D_\mu = (1-\epsilon_\mu)\mathcal{N}(\mu, 1-c_1^2\epsilon) + \epsilon_\mu B_\mu,$$

for some distribution B_μ and ϵ_μ satisfying

1. *If $|\mu| < c_1^2\sqrt{\epsilon}$, then $\epsilon_\mu = \epsilon$ and $\chi^2(D_\mu, \mathcal{N}(0,1)) = e^{O(1/\epsilon)}$.*
2. *If $|\mu| \geq c_1^2\sqrt{\epsilon}$, then we take ϵ_μ such that $\epsilon_\mu/(1-\epsilon_\mu) = \mu^2$, which simplies to*

$$\epsilon_\mu = \frac{\mu^2}{1+\mu^2}, \tag{35}$$

and $\chi^2(D_\mu, \mathcal{N}(0,1)) = e^{O(\max(1/\mu^2, \mu^2))}$.

Proof. It is straightforward to verify that $D_\mu = (1-\epsilon_\mu)\mathcal{N}(\mu, 1-c_1^2\epsilon) + \epsilon_\mu \mathcal{N}(a,b)$ has mean 0 and second moment 1, where a and b are defined by

$$a = -\frac{\mu(1-\epsilon_\mu)}{\epsilon_\mu},\ b = 1.$$

Facts D.11 and D.12 imply that if $\sigma_1 = O(1), \sigma_2 = O(1)$, then

$$\chi^2(\epsilon_\mu \mathcal{N}(\mu_1, \sigma_1^2) + (1-\epsilon_\mu)\mathcal{N}(\mu_2, \sigma_2^2),\ \mathcal{N}(0,1)) = e^{O\left(\mu_1^2+\mu_2^2\right)}.$$

- **Case 1**: $|\mu| < c_1^2\sqrt{\epsilon}$. Then taking $\epsilon_\mu = \epsilon$ implies that $|a| \leq \mu/\epsilon < c_1/\sqrt{\epsilon}$, and $\chi^2(D_\mu, \mathcal{N}(0,1)) = e^{O(1/\epsilon)}$.
- **Case 2**: $|\mu| \geq c_1^2\sqrt{\epsilon}$. Then taking $\epsilon_\mu/(1-\epsilon_\mu) = \mu^2$ implies that $|a| = \frac{1}{|\mu|}$, and $\chi^2(D_\mu, \mathcal{N}(0,1)) = e^{O(\max(1/\mu^2, \mu^2))}$. □

We follow the non-Gaussian component analysis construction and define the conditional distribution of the corrupted $\mathrm{vec}(A)|y$ to have a hidden direction v.

Definition D.7 (High-Dimensional Hidden Direction Distribution [DK23, Definition 8.9])**.** For a one-dimensional distribution D that admits density $D(x)$ and a unit vector $v \in \mathbb{R}^{d^2}$, the distribution $\mathbf{P}_v^A$ is defined to be the product distribution of $D(x)$ in the v-direction and standard normal distribution in the subspace perpendicular to v, i.e.,

$$\mathbf{P}_v^D(a) = \frac{1}{(2\pi)^{\frac{d^2-1}{2}}} D(v^\top a) \exp\left(-\frac{1}{2}\left\|a - (v^\top a)v\right\|^2\right).$$

Lemma D.8 (Construction of Corrupted Conditional Distribution)**.** *Define $P_{\mu(y),v} = \mathbf{P}_{\mathrm{vec}(vv^\top)}^{D_{\mu(y)}}$. Then for all unit vectors $v, v' \in \mathbb{R}^d$, it holds that*

$$\left|\chi_{\mathcal{N}(0,I)}(P_{\mu(y),v}, P_{\mu(y),v'}) - 1\right| \leq \left(v^\top v'\right)^4 \chi^2(D_{\mu(y)}, \mathcal{N}(0,1)). \tag{36}$$

Proof. By Lemma 8.12 in the textbook [DK23], we have

$$\begin{aligned}\left|\chi_{\mathcal{N}(0,I)}(P_{\mu(y),v}, P_{\mu(y),v'}) - 1\right| &= \left|\chi_{\mathcal{N}(0,I)}(\mathbf{P}^{D_{\mu(y)}}_{\mathrm{vec}(vv^\top)}, \mathbf{P}^{D_{\mu(y)}}_{\mathrm{vec}(v'v'^\top)}) - 1\right| \\ &\le (\mathrm{vec}(vv^\top)^\top \mathrm{vec}(v'v'^\top))^2 \, \chi^2(D_{\mu(y)}, \mathcal{N}(0,1)) \\ &= (v^\top v')^4 \, \chi^2(D_{\mu(y)}, \mathcal{N}(0,1)),\end{aligned}$$

where the inequality is a direct application of [DK23, Lemma 8.12], and the last equality is because

$$\mathrm{vec}(vv^\top)^\top \mathrm{vec}(v'v'^\top) = \langle vv^\top, v'v'^\top \rangle = \mathrm{tr}(vv^\top v'v'^\top) = v^\top v' \, \mathrm{tr}(vv'^\top) = (v^\top v')^2.$$

□

Now we define $Q'_v(A, y)$ to be a contaminated version of the joint distribution of A and y such that $\mathrm{vec}(A)|y \sim P_{\mu(y),v}$ in the following lemma. Recall that, according to Lemma D.6, different level of noise $\epsilon_{\mu(y)}$ is needed for different y, and the implication of the following lemma is that the total noise is less than ϵ. This is done via integration with respect to y.

Lemma D.9 (Construction of Corrupted Joint Distribution: Controlling Large $\epsilon_{\mu(y)}$)**.** *Recall $\mu(y) = c_1\sqrt{\epsilon}y$ and Q_v is the joint distribution of A and y for clean samples. Define $Q'_v(A, y) = P_{\mu(y),v}(A)R(y)$, where*

$$R(y) = \frac{G(y)}{(1-\epsilon_{\mu(y)}) \int G(y')/(1-\epsilon_{\mu(y')})dy'},$$

$G(y)$ is the marginal distribution of y under Q_v, which is standard Gaussian.

Then $Q'_v(A, y)$ is the contaminated joint distribution of A, y i.e., $Q'_v = (1-\epsilon)Q_v + \epsilon N_v$ for some noise distribution N_v. Under Q'_v, it holds that $A|y \sim P_{\mu(y),v}$.

Proof. We proceed to bound $\int G(y')/(1-\epsilon_{\mu(y')})dy' \le 1/(1-\epsilon)$. This shows that $R(y)$ is well-defined, and we will use that to show $Q'_v \ge (1-\epsilon)Q_v$ (recall that we obscure the line between the distribution and its probability density function).

$$\begin{aligned}\int G(y')/(1-\epsilon_{\mu(y')})dy' &= \int G(y') \left(1 + \frac{\epsilon_{\mu(y')}}{1-\epsilon_{\mu(y')}}\right) dy' \\ &\le 1 + \int_{|y|\ge c_1} G(y') \left(\frac{\epsilon_{\mu(y')}}{1-\epsilon_{\mu(y')}}\right) dy' + \int_{|y|<c_1} G(y') \frac{\epsilon}{1-\epsilon} dy' \\ &\le 1 + 2c_1\epsilon + \int G(y')\mu(y')^2 dy' \\ &\le 1 + 2c_1\epsilon + \int G(y')c_1^2\epsilon y'^2 dy' \\ &\le 1 + 2c_1\epsilon + c_1^2\epsilon \\ &\le 1/(1-\epsilon),\end{aligned}$$

where the last step holds by choosing c_1 to be sufficiently small.

Since for any y and A, $D_{\mu(y)} \ge (1-\epsilon_{\mu(y)})\mathcal{N}(\mu(y), 1 - c_2\epsilon^2)$ by the construction of D_μ, so $P_{y,v} \ge (1-\epsilon_{\mu(y)})\mathcal{N}(\mu(y)\,\mathrm{vec}(vv^\top), I - c_1^2\epsilon^2\,\mathrm{vec}(vv^\top)\,\mathrm{vec}(vv^\top)^\top)$, since $P_{y,v}$ agrees with Gaussian in all directions except $\mathrm{vec}(vv^\top)$. We proved that $R(y) \ge (1-\epsilon)G(y)/(1-\epsilon_{\mu(y)})$, so $Q'_v = P_{\mu(y),v}R(y) \ge (1-\epsilon)G(y)\mathcal{N}(\mu(y)\,\mathrm{vec}(vv^\top), I - c_1^2\epsilon^2\,\mathrm{vec}(vv^\top)\,\mathrm{vec}(vv^\top)^\top) = (1-\epsilon)Q_v$. □

Next we show that Q'_v are near orthogonal to each other for distinct v, if we view $\chi_S(\cdot,\cdot) - 1$ as an inner product on the space of d^2+1-dimensional distributions.

Lemma D.10 (Near Orthogonality of Corrupted Joint Distributions)**.** *Let S be the joint distribution of A and y when they are independent and entries of A are i.i.d. standard Gaussian and $y \sim R$. Then we have*

$$\chi_S(Q'_v, Q'_{v'}) - 1 = e^{O(1/\epsilon^3)}(v^\top v')^4.$$

Proof. By the definition of $\chi_S(Q'_v, Q'_{v'})$, we have

$$\begin{aligned}
\chi_S(Q'_v, Q'_{v'}) &= \int Q'_v(a,y)Q'_{v'}(a,y)/S(x,y)dxdy \\
&= \int P_{\mu(y),v}(a)P_{\mu(y),v'}(a)R(y)^2/(G(a)R(y))dxdy \\
&= \int \chi_{\mathcal{N}(0,I_{d^2})}(P_{\mu(y),v}, P_{\mu(y),v'})R(y)dy \\
&\le 1 + O\big((v^\top v')^4\big) \int \chi^2(D_{\mu(y),\epsilon}, \mathcal{N}(0,1))R(y)dy,
\end{aligned}$$

where the last inequality follows from Lemma D.8.

Recall $\mu(y) = c_1\sqrt{\epsilon}y$. The main technical part is to bound the following quantity:

$$\begin{aligned}
\int \chi^2(D_{\mu(y),\epsilon}, \mathcal{N}(0,1))R(y)dy &\le e^{O(1/\epsilon)} + \int_{|y|\ge c_1} e^{O\left(\max\{1/\mu(y)^2, \mu(y)^2\}\right)} R(y)dy \\
&\le e^{O(1/\epsilon)} + \int_{|y|\ge c_1} e^{O\left(\max\{1/(c_1^4\epsilon), \mu(y)^2\}\right)} R(y)dy \\
&\le e^{O(1/\epsilon)} + e^{1/(c_1^4\epsilon)} \int_{|y|\ge c_1} e^{c_1^2\epsilon y^2} R(y)dy \\
&\le e^{O(1/\epsilon)} + e^{1/(c_1^4\epsilon)} \int_{|y|\ge c_1} e^{c_1^2\epsilon y^2}(1-\epsilon)G(y)/(1-\epsilon_{\mu(y)})dy \\
&\le e^{O(1/\epsilon)} + e^{1/(c_1^4\epsilon)} \int_{|y|\ge c_1} e^{c_1^2\epsilon y^2} c_1^2 y^2 G(y)dy \\
&= e^{O(1/\epsilon)},
\end{aligned}$$

where the first inequality splits into two cases as in Lemma D.6, and the last inequality follows by choosing c_1 to be a sufficiently small constant so that $c_1^2\epsilon \le 1/4$ and $\int e^{c_1^2\epsilon y^2} y^2 G(y)dy$ is bounded above by a constant. □

Proof of Theorem 4.2. The proof follows a standard non-Gaussian component analysis argument. For any $0 < c < 1/2$, there is a set S of at least 2^{d^c} unit vectors in $\mathbb{R}^d$ such that for each pair of distinct $v, v' \in S$, $\left|v^\top v'\right| = O(d^{c-1/2})$. By Lemma D.10, we have

$$\chi_S(Q'_v, Q'_{v'}) - 1 = \begin{cases} e^{O(1/\epsilon)}O(d^{4c-2}) =: \gamma & \text{if } v \ne v' \\ e^{O(1/\epsilon)} =: \beta & \text{if } v = v' \end{cases}.$$

Therefore, by Lemma 8.8 in [DK23], any SQ algoritm requires at least $2^{d^c}\gamma/\beta = 2^{d^c}/d^{2-4c}$ calls to $\mathsf{STAT}(e^{1/\sqrt{\epsilon}}/d^{1-2c})$ to find v and therefore u within better than $O\big(\epsilon^{1/4}\big)$. □

We state the auxiliary facts that we used in the above analysis:

Fact D.11 ([Dia+21, Lemma G.3])**.** *Let $k \in \mathbb{Z}_+$, distributions P_i and $\lambda_i \ge 0$, for $i \in [k]$ such that $\sum_{i=1}^k \lambda_i = 1$. We have that $\chi^2\left(\sum_{i=1}^k \lambda_i P_i, D\right) = \sum_{i=1}^k \sum_{j=1}^k \lambda_i\lambda_j\chi_D(P_i, P_j)$.*

Fact D.12 ([DKS19, Lemma F.8])**.**

$$\chi_{\mathcal{N}(0,1)}\left(\mathcal{N}(\mu_1,\sigma_1^2), \mathcal{N}(\mu_2,\sigma_2^2)\right) = \frac{\exp\left(-\frac{\mu_1^2(\sigma_2^2-1)+2\mu_1\mu_2+\mu_2^2(\sigma_1^2-1)}{2\sigma_1^2(\sigma_2^2-1)-2\sigma_2^2}\right)}{\sqrt{\sigma_1^2+\sigma_2^2-\sigma_1^2\sigma_2^2}}.$$

E A Counterexample: Why Simple Algorithms Do Not Work

While Section D formally showed that a broad class of algorithms with sample complexity proportional to the dimension d cannot achieve $O(\epsilon^{1/4})$ error without an exponential amount of computation, we informally discuss a concrete construction of adversarial corruptions where some straightforward algorithms fail. This discussion is intended only to provide intuition; the SQ lower bound

provides more rigorous evidence that improving sample complexity beyond quadratic is impossible. The $O(\cdot)$ notation in this section only displays dependence on the dimension d.

Let $(A_i, y_i) \sim \mathcal{G}$ be sampled according to the noiseless Gaussian design in Definition 1.6, i.e. entries of A_i are i.i.d. standard Gaussians and $y_i = \langle M^*, A_i \rangle$.

- The adversary first inspects all n samples and computes M^* up to some small error with out-of-shelf low rank matrix sensing algorithms. In the following discussion, we simply assume the adversary finds M^* exactly.
- The adversary discards a random ϵ-fraction of samples and let $\mathcal{S}$ denote the set of remaining samples. The adversary computes $P = -\frac{1}{n}\sum_{i\in\mathcal{S}} y_i A_i$.
- For $j = 1, \ldots, \epsilon n$, the adversary independently samples z_j uniformly between the spheres of radius $\|M^*\|_F / 2$ and $2\|M^*\|_F$ and computes $E_j = \frac{P}{\epsilon z_j} + A'_j$, where entries of A'_j are i.i.d. sampled from standard Gaussian.
- The adversary adds (z_j, E_j) as corruptions. Let $\mathcal{B}$ denote the points added by the adversary.

Recall that with the objective function f_i defined in Equation (3) and $B_i = A_i + A_i^T$, we have the gradient and Hessian

$$\nabla f_i(U) = (\langle UU^\top, A_i\rangle - y_i)(A_i + A_i^\top)U$$
$$\nabla^2 f_i(U) = (\langle UU^\top, A_i\rangle - y_i) I_r \otimes B_i + \operatorname{vec}(B_i U)\operatorname{vec}(B_i U)^\top.$$

Prior works on first-order robust stochastic optimization [Pra+20; Dia+19] required $O(d)$ samples, because they rely on the robust estimation of the gradients. They would accept 0 as the solution, because when $U = 0$, $\nabla f_i(U) = 0$ for all i, i.e., all sample gradients are zero at $U = 0$ whether or not they are corruptions, so gradient-based filtering does not work. On the other hand, $\nabla^2 \bar{f}(0) = -\mathbb{E}[\langle M^*, A_i\rangle I_r \otimes B_i] = -I_r \otimes \mathbb{E}[\langle M^*, A_i\rangle (A_i + A_i^T)] = -2I_r \otimes M^*$, where the last equality follows from Lemma C.2. Hence the error measured by the negative curvature of the solution is $\lambda_{\min}(\nabla^2 \bar{f}(0)) = 2\sigma_1^\star$, which does not even scale with ϵ.

This adversary can also help illustrate why spectral initialization similar to the algorithm discussed in Theorem C.13 cannot succeed with $O(d)$ samples. Spectral methods rely on the observation that with clean samples, $\mathbb{E}_{(A_i,y_i)\sim\mathcal{G}}[y_i A_i] = M^*$. The hope is that the matrix formed by the leading r singular vectors of some robust mean estimation of $\{y_i A_i\}$ is close enough to M^*. If so, it can be a good initialization to Algorithm 3.1, which only requires robust estimation of sample gradients and $O(d)$ samples suffice.

However, with $O(d)$ samples, it is unclear how to estimate the mean of $\{y_i A_i\}$. Robust mean estimation algorithms (e.g. Algorithm A.2) discussed in Proposition 2.3 do not work because they require $O(d^2)$ samples. Here we show that the naive filter based on the norm of $\{y_i A_i\}$ does not work, either.

Observe that $P \approx -M^*$ by construction. For each j,

$$\|E_j\|_F \le \left\|A'_j\right\|_F + \|P\|_F / (z_i\epsilon) \approx \left\|A'_j\right\|_F + \|M^*\|_F / (z_i\epsilon) \le \left\|A'_j\right\|_F + 2/\epsilon.$$

Since $\left\|A'_j\right\|_F$ scales with the dimension d but ϵ is a fixed constant, as we move to a high-dimensional regime, $\left\|A'_j\right\|_F$ dominates and the effects of the outliers diminish if we only look at the norm of data.

Without an effective filter, the mean of $\{y_i A_i\}$ across both clean samples and outliers becomes

$$\frac{1}{n}\sum_{i\in\mathcal{S}} y_i A_i + \frac{1}{n}\sum_{j\in\mathcal{B}} z_j E_j = -P + \frac{1}{n}\sum_{j\in\mathcal{B}} P/\epsilon + A'_j = \frac{1}{n}\sum_{j\in\mathcal{B}} A'_j,$$

which contains no signals — only noise — and is close to 0.

In summary, we constructed an adversary under which robust first-order methods and spectral methods with naive filter fail. Note that this construction relies on the ability of the adversary to corrupt the input matrices A_i; if the adversary could only corrupt y_i, then the results in [Li+20b; Li+20a] would apply and $\widetilde{O}(d)$ samples would suffice.