Optimal Robust Estimation under Local and Global Corruptions: Stronger Adversary and Smaller Error

Thanasis Pittas PITTAS@WISC.EDU
University of Wisconsin-Madison

Ankit Pensia ANKITP@BERKELEY.EDU
Simons Institute for the Theory of Computing

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Algorithmic robust statistics has traditionally focused on the contamination model where a small fraction of the samples are arbitrarily corrupted. We consider a recent contamination model that combines two kinds of corruptions: (i) small fraction of arbitrary outliers, as in classical robust statistics, and (ii) local perturbations, where samples may undergo bounded shifts on average. While each noise model is well understood individually, the combined contamination model poses new algorithmic challenges, with only partial results known. Existing efficient algorithms are limited in two ways: (i) they work only for a weak notion of local perturbations, and (ii) they obtain suboptimal error for isotropic subgaussian distributions (among others). The latter limitation led Nietert et al. (2024) to hypothesize that improving the error might, in fact, be computationally hard. Perhaps surprisingly, we show that information theoretically optimal error can indeed be achieved in polynomial time, under an even *stronger* local perturbation model (the sliced-Wasserstein metric as opposed to the Wasserstein metric). Notably, our analysis reveals that the entire family of *stability-based* robust mean estimators continues to work optimally in a black-box manner for the combined contamination model. This generalization is particularly useful in real-world scenarios where the specific form of data corruption is not known in advance. We also present efficient algorithms for distribution learning and principal component analysis in the combined contamination model.

Keywords: robust statistics, stability, mean estimation, distribution learning, Wasserstein

1. Introduction

We study high-dimensional parameter estimation and distribution learning in settings where data deviate from the traditional i.i.d. assumption. This may arise from (i) outliers due to data poisoning attacks (Barreno et al., 2010; Biggio et al., 2012; Steinhardt et al., 2017; Tran et al., 2018; Hayase et al., 2021), system errors in geometric perception (Yang and Carlone, 2023), and biological anomalies (Rosenberg et al., 2002; Paschou et al., 2010; Li et al., 2008), or (ii) local distribution shifts (Chao and Dobriban, 2023; Yang et al., 2024) caused by sensor biases. In high dimensions, such corruptions can severely impact standard estimators designed for the i.i.d. setting.

The field of robust statistics was initiated in the 1960s (Tukey, 1960; Huber, 1964) to develop estimators resilient to a small fraction of outliers, formalized below.

Global Contamination Model Let $\epsilon \in (0, 1/2)$. Let S be a multiset of n points in $\mathbb{R}^d$. Consider all n-sized sets in $\mathbb{R}^d$ that differ in at most ϵ-fraction of points, i.e.,

$$\mathcal{O}(S, \epsilon) := \left\{ S' \subset \mathbb{R}^d : |S'| = n \text{ and } |S' \cap S| \geq (1 - \epsilon)n \right\}. \tag{1}$$

The adversary can return any set $T \in \mathcal{O}(S, \epsilon)$. We call the points in S inliers and the points in $T \setminus S$ outliers.

We term these outliers "global" as they can be arbitrary in magnitude. Early work developed minimax optimal estimators for various robust estimation tasks, albeit with runtimes exponential in the dimension. Over the past decade (Diakonikolas et al., 2016; Lai et al., 2016; Diakonikolas and Kane, 2023), poly-time algorithms with optimal error have been developed for key distribution families.[1]

A limitation in Global Contamination Model is that the inliers remain *unchanged*, making the contamination *sparse*. However, in practice, *each point* could be perturbed, e.g., from miscalibrated sensors; These perturbations are *dense* but *local*. To model them, recent works have used Wasserstein distance for (Zhu et al., 2022b, 2019; Liu and Loh, 2022; Chao and Dobriban, 2023; Nietert et al., 2024), which measures perturbations using the Euclidean norm. In contrast, we allow stronger perturbation by measuring them in a directional manner as follows:

StrongLC (Strong Local Contamination) Model Let $\rho \geq 0$ and a set $S_0 = \{x_i\}_{i \in [n]}$. Define $\mathcal{W}^{\text{strong}}(S_0, \rho) := \{S = \{\widetilde{x}_1, \dots, \widetilde{x}_n\} \subset \mathbb{R}^d : \sup_{v \in \mathbb{R}^d : \|v\|_2 = 1} \frac{1}{n} \sum_{i \in [n]} |v^\top (\widetilde{x}_i - x_i)| \leq \rho\}$. The adversary returns an arbitrary set $S \in \mathcal{W}^{\text{strong}}(S_0, \rho)$ after possibly reordering the points.

While this model can cause significant perturbations of each point in Euclidean norm (up to $\rho\sqrt{d}$), it remains quite benign since the sample mean has optimal error of $O(\rho)$ (any estimator has error $\Omega(\rho)$ since all points can shift by ρ in the same direction). Moreover, mean estimation under Global Contamination Model is well understood, yet their combination surprisingly introduces new algorithmic challenges.

Global+StrongLC (Global plus Strong Local Contamination) Model Let $\epsilon \in (0, 1/2)$ and $\rho > 0$. Let $S_0 = \{x_1, \dots, x_n\}$ be a set of n points in $\mathbb{R}^d$. The adversary can return an arbitrary set T such that $T \in \mathcal{O}(S, \epsilon)$ for some $S \in \mathcal{W}^{\text{strong}}(S_0, \rho)$.

The combined contamination model is more realistic, as data can be perturbed in unforeseen ways (Zhu et al., 2022b; Nietert et al., 2024). However, the algorithms above face challenges: (i) the sample mean is sensitive to outliers, and (ii) Outlier-robust algorithms developed over the last decade may fail because they rely on moment structure, which can be disrupted by the local perturbations.

The remainder of this introduction is organized as follows: Section 1.1 and Section 1.2 present the key research questions and our results. Additional related work is deferred to Appendix A.

1.1. Motivating Questions

For brevity, we focus only on the task of mean estimation in this section. If $\mathcal{P}$ is a distribution family with $f_{\mathcal{P}}^{\text{robust}}(\epsilon)$ being its optimal asymptotic error for Global Contamination Model, the optimal asymptotic error under Global+StrongLC Model is $\Theta(f_{\mathcal{P}}^{\text{robust}}(\epsilon) + \rho)$. This is achieved by the multivariate trimmed mean, which is optimal in each model individually (Lugosi and Mendelson, 2021). However, the trimmed mean is computationally inefficient in high dimensions, raising the question:

[1] These families include bounded covariance distributions, isotropic subgaussians, and isotropic distributions with bounded k-th moments. We refer the reader to Appendix (Corollary 21) for precise statements.

Question 1 *Can we perform robust estimation (mean estimation and distribution learning) under Global+StrongLC Model in a computationally-efficient manner?*

Existing results for Question 1 achieve suboptimal error, both in terms of the global contamination parameter ϵ and the local contamination parameter ρ. The most relevant work is the recent paper Nietert et al. (2024), however it uses a weaker definition for the local corruptions:

WeakLC (Weak Local Contamination) Model (Nietert et al., 2024) Let $\rho > 0$ and $S_0 = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$. Define $\mathcal{W}^{\text{weak}}(S_0, \rho) := \{S = \{\widetilde{x}_1, \ldots, \widetilde{x}_n\} \subset \mathbb{R}^d : \frac{1}{n}\sum_{i\in[n]} \|\widetilde{x}_i - x_i\|_2 \leq \rho\}$. The adversary can return any $S \in \mathcal{W}^{\text{weak}}(S_0, \rho)$ after possibly reordering the samples.

Global+WeakLC (Global plus Weak Local Contamination) Model (Nietert et al., 2024) Let $\epsilon{<}1/2$, $\rho > 0$, and $S_0 = \{x_i\}_{i=1}^n$. The adversary can return any set $T \in \mathcal{O}(S, \epsilon)$ for any $S \in \mathcal{W}^{\text{weak}}(S_0, \rho)$.

Under Global+WeakLC Model and assuming bounded covariance data, Nietert et al. (2024) provided a polynomial-time algorithm with error on the order of $\sqrt{\epsilon} + \rho_{\text{weak}}$, where here we have renamed the radious parameter in Global+WeakLC Model as ρ_{weak} for clarity. WeakLC Model is up to a $\sqrt{d}$ factor weaker than StrongLC Model, in the sense that $\mathcal{W}^{\text{weak}}(S_0, \rho) \subset \mathcal{W}^{\text{strong}}(S_0, \rho) \subseteq \mathcal{W}^{\text{weak}}(S_0, \rho\sqrt{d})$, hence, applying the results of Nietert et al. (2024) to our contamination model (Global+StrongLC Model) leads to an error that has an extraneous $\sqrt{d}$ in front of ρ (the parameter used in StrongLC Model), which is undesirable in large dimensions. The fact that the trimmed mean analysis depends only on boundedness in each direction (rather than the Euclidean norm) inspired us to consider Global+StrongLC Model. However, the analysis in Nietert et al. (2024) critically relies on small Euclidean norms for local perturbations, leading to the question:

Question 2 *Do weak and strong local contaminations lead to different computational landscapes, when combined with global contamination? In particular, does the dependence on ρ in the computationally-efficiently achievable error differ between Global+StrongLC Model and Global+WeakLC Model?*

Regarding the error dependence on ϵ, even for the weaker Global+WeakLC Model, Nietert et al. (2024) achieves only partial results. While it achieves optimal dependence for bounded covariance distributions, it is suboptimal for isotropic subgaussians (by a $\widetilde{\Theta}(\epsilon^{-1/2})$ factor) and distributions with bounded k-th moments (by a $\epsilon^{-1/2+1/k}$ factor).[2] In fact, Nietert et al. (2024) conjectured that no polynomial-time algorithm achieves optimal error for these families, stating "*We suspect that there may be similar obstacles [(computational hardness)] as those known for robust mean estimation with stable but non-isotropic distributions* (Hopkins and Li, 2019)". Thus the following generalizes an open problem in Nietert et al. (2024):

Question 3 *Do local corruptions (either weak or strong) induce new information-computation gaps for robust estimation? In particular, does the dependence on ϵ in the computationally-efficiently achievable error change in the presence of local contamination?*

[2] In fact, their algorithm has suboptimal ϵ-dependence even in one dimension.

1.2. Our Results

We fully resolve Questions 1 to 3, with techniques that uniformly extend to mean estimation, distribution learning, and principal component analysis. This section presents efficient algorithms for mean estimation and distribution learning, while robust PCA results are deferred to Appendix H.

Mean estimation We show that mean estimation with optimal guarantees is possible under two well-studied and commonly used conditions in the literature: *stability* and *sum-of-squares-certifiable moments*. Starting with the former, the stability condition for a set of samples S is the following (where we use the notation $\mu_S = \frac{1}{|S|}\sum_{x\in S} x$ and $\overline{\mathbf{\Sigma}}_S = \frac{1}{|S|}\sum_{x\in S}(x-\mu)(x-\mu)^\top$):

Definition 1 (Stability, see, e.g., Diakonikolas and Kane (2023)) *Let $\epsilon \in (0, 1/2)$ and $\delta \in [\epsilon, \infty)$. A finite multiset $S \subset \mathbb{R}^d$ is called (ϵ, δ)-stable with respect to $\mu \in \mathbb{R}^d$ if for every $S' \subseteq S$ with $|S'| \geq (1-\epsilon)|S|$, the following hold: (i) $\|\mu_{S'} - \mu\|_2 \leq \delta$, and (ii) $\left\|\overline{\mathbf{\Sigma}}_{S'} - \mathbf{I}\right\|_{\text{op}} \leq \delta^2/\epsilon$.*

Definition 2 (Stability-based algorithms) *Let S be an (ϵ, δ)-stable set with respect to an (unknown) $\mu \in \mathbb{R}^d$. Let T be any set such that $T \in \mathcal{O}(S, \epsilon)$ (cf. Global Contamination Model). We call an algorithm $\mathcal{A}(T, \epsilon, \delta)$ stability-based if it takes as an input T, ϵ, and δ, and outputs an estimate $\widehat{\mu}$ in polynomial time such that, with high probability (over its internal randomness), $\|\widehat{\mu} - \mu\|_2 \lesssim \delta$.*

Over time, a growing body of work has introduced algorithms, ranging from convex programming to gradient descent, that rely solely on stability and have been highly optimized for runtime, sample complexity, and memory (see Appendix A for references). Importantly, stability holds with high-probability for well-behaved distribution families: Appendix C provides concrete known bounds, as well as an improvement developed in this paper. Our main result below, allows us to seamlessly apply this huge repertoire of algorithms to the combination of global and local contamination.

Theorem 3 (Mean estimation) *Let C be a sufficiently large constant, $c = 1/C$. Let $\epsilon \in (0, c)$ and $\rho > 0$ be parameters. Let $S_0 \subset \mathbb{R}^d$ be an (ϵ, δ)-stable set with respect to an unknown $\mu \in \mathbb{R}^d$, where $\delta > \epsilon$. Let T be a corrupted dataset with an ϵ-fraction of global outliers and ρ-strong local corruptions (Global+StrongLC Model). Then, any stability-based algorithm $\mathcal{A}$ on input T, ϵ and $\widetilde{\delta} := C \cdot (\delta + \rho)$, outputs $\widehat{\mu} \in \mathbb{R}^d$ such that, with high probability, $\|\widehat{\mu} - \mu\|_2 = O(\delta + \rho)$.*

Note that the dependence on ρ is optimal. Some additional comments are in order:

- The algorithms work for all contamination models (local, global, and combined) without modification, which is beneficial since in practice we do not know the types of corruption in the dataset.[3]
- The success of stability-based algorithms may seem paradoxical because: (i) stability requires bounded covariance, and (ii) the set S after local contamination can have significantly larger covariance (cf. Section 1.3). To overcome this we prove that despite S not being stable due to large covariance, it contains a large stable subset, which allows stability-based algorithms to work.
- Combining Theorem 3 with the fact that stability holds with high probability for well-behaved distributions, we immediately obtain robust mean estimators with (near)-optimal rates for these families. For such families, the stability parameter δ as a function of $\epsilon, \rho, d, n, \tau$ has been well-studied and almost optimal bounds (up to a $\sqrt{\log d}$ factor) are known. As an additional contribution, in this paper, we further tighten these bounds by completely removing the $\sqrt{\log d}$ factor. The statement (Corollary 28) and a more detailed discussion on this can be found in Appendix F.

[3] While Theorem 3 shows that ρ is part of the input parameter $\widetilde{\delta} = C \cdot (\rho + \delta)$, some stability-based algorithms do not require $\widetilde{\delta}$ as an input parameter (Diakonikolas et al., 2020, Appendix A).

In light of the above, Theorem 3 and Corollary 28 simultaneously answer Questions 1 to 3 for robust mean estimation. In particular, (i) both weak and strong local contamination yield the same computationally-efficient rates, answering Question 2 and (ii) local contamination (whether weak or strong) does not induce new information-computation gap, refuting the hypothesis in Nietert et al. (2024) and answering Question 3.

Another important class for which algorithms have been developed is distributions with certifiably bounded moments, which have low-degree *sum of squares proofs* for the bounded moment conditions (see Appendix I for related definitions and background). Importantly, their covariance may differ from identity and be unknown to the learning algorithm. For this class, we obtain the following:

Theorem 4 (Optimal asymptotic error for certifiably bounded distributions; informal) *Let $\epsilon \in (0, c)$ for a sufficiently small absolute constant c. Let P be a distribution over $\mathbb{R}^d$ with mean μ and t-th moment certifiably bounded by M_t. Then there is an algorithm that draws $n = \mathrm{poly}(d^t, 1/\epsilon)$ samples, runs in time $\mathrm{poly}(n^t, d^{t^2})$, and outputs an estimate $\widehat{\mu} \in \mathbb{R}^d$ such that with high constant probability $\|\widehat{\mu} - \mu\|_2 = O(M_t \epsilon^{1-\frac{1}{t}})$.*

The asymptotic error $\epsilon^{1-1/t}$ is again optimal for this class of distributions, and the sample complexity is qualitatively optimal for a broad family of algorithms such as statistical query algorithms and low-degree polynomials (Diakonikolas et al., 2022a). We prove Theorem 4 in Appendix I.

Distribution learning We now move beyond mean estimation to the problem of distribution learning with respect to the (sliced)-Wasserstein metric, defined below.

Definition 5 (Sliced Wasserstein distance) *Let P, Q be two distributions and denote by $\mathcal{V}_k$ the set of all rank-k projection matrices. The k-sliced p-Wasserstein Distance is defined as $W_{p,k}(P,Q) := \max_{\mathbf{V} \in \mathcal{V}_k} \inf_{\pi \in \Pi(P,Q)} \mathbf{E}_{(x,x') \sim \pi} \left[\|\mathbf{V}(x - x')\|_2^p\right]^{1/p}$, where $\Pi(P,Q)$ is the set of all couplings of P and Q. By slightly overloading our notation, when S and $\hat{S}$ are sets of points, we denote by $W_{p,k}(S, \hat{S})$ the k-sliced p-Wasserstein distance between the uniform distributions over S and $\hat{S}$.*

Distribution learning in this metric has applications in distributionally robust optimization (Nietert et al., 2024). To present our result in full generality, we consider the following contamination model, that interpolates between StrongLC Model and WeakLC Model for $k = 1$ and $k = d$, respectively.

StrongWC (Strong Wasserstein Contamination) Model Let $\rho > 0$, and denote by $\mathcal{V}_k$ the set of all rank-k projections matrices. Let $S_0 = \{x_1, \dots, x_n\}$ be a multiset of n points in $\mathbb{R}^d$. Define the following set of local perturbations which are small in all rank-k subspaces: $\mathcal{W}_{1,k}^{\mathrm{strong}}(S_0, \rho) := \left\{ S = \{\widetilde{x}_1, \dots, \widetilde{x}_n\} \subset \mathbb{R}^d : \max_{\mathbf{V} \in \mathcal{V}_k} \frac{1}{n} \sum_{i \in [n]} \|\mathbf{V}(\widetilde{x}_i - x_i)\|_2 \leq \rho. \right\}$ The adversary can return any set $S \in \mathcal{W}_{1,k}^{\mathrm{strong}}(S_0, \rho)$ after possibly reordering the points. We call the set S ρ-contaminated version of the set S_0 under the k-sliced Wasserstein adversary.

The distribution problem we consider is as follows: Let S_0 be a stable set of inliers, corrupted by a combination of local (StrongWC Model) and global corruptions (Global Contamination Model). The goal is to output a distribution $\hat{P}$ close to the uniform distribution over S_0 with respect to the $W_{1,k}$ metric (a natural choice since the corruptions in StrongWC Model are also measured in $W_{1,k}$).

Theorem 6 (Distribution learning) *Let parameters $\epsilon \in (0, c)$ where c is a sufficiently small absolute constant, $\rho > 0$ and $\delta > \epsilon$. Let $S_0 \subset \mathbb{R}^d$ be a set that is (ϵ, δ)-stable with respect to an* *(unknown) $\mu \in \mathbb{R}^d$. For a slicing parameter $k \in [d]$, let T be the corrupted dataset after a combination of local and global corruptions from StrongWC Model and Global Contamination Model with parameters ρ and ϵ, respectively (i.e., $T \in \mathcal{O}(S, \epsilon)$ for some $S \in \mathcal{W}_{1,k}^{\text{strong}}(S_0, \rho)$). Then, there exists a polynomial-time algorithm that on input $T, \epsilon, \rho, k, \delta$, the algorithm outputs an estimate $\widehat{S} \subset T$ such that, with high constant probability, for all $k' \in [k]$ it holds that $W_{1,k'}(\widehat{S}, S_0) = O(\delta\sqrt{k'} + \rho)$.*

Nietert et al. (2024) also provides rates for distribution estimation in the $W_{1,k}$ metric. However, their adversary for the local contamination is much weaker (it measures contamination using $W_{1,d}$ instead of $W_{1,k}$).In contrast, our rates for both the local perturbation and accuracy are measured in $W_{1,k}$.

Our mean estimation result, Theorem 3, is a special case for $k' = 1$.[4] For general $k \geq 1$, our algorithm follows a filtering-based approach that uses a certificate lemma from Nietert et al. (2024), which we directly optimize for the optimal error $\delta\sqrt{k'} + \rho$. In contrast, Nietert et al. (2024) optimizes an approximation of their certificate, leading to the larger error $\max(\delta, \sqrt{\epsilon})\sqrt{k'} + \rho$.[5]

We now discuss the error guarantee of Theorem 6 in more detail. First, the error of $\Omega(\rho)$ is trivially needed because each point could be shifted by distance ρ along the same direction. Next, the error term $\delta\sqrt{k'}$ is also optimal; see Nietert et al. (2024, Corollary 5).

We now show how to instantiate the error guarantee of Theorem 6 for learning a distribution P over $\mathbb{R}^d$ as opposed to the uniform distribution on S_0. First, by the triangle inequality $W_{1,k'}(\hat{S}, P) \leq W_{1,k'}(\hat{S}, S_0) + W_{1,k'}(S_0, P) = O(\delta\sqrt{k'} + \rho) + W_{1,k'}(S_0, P)$. While the first two terms are optimal (as shown above), the third is $\widetilde{O}(\sqrt{d}\, k'\, n^{-\frac{1}{\max(k',2)}})$ by Boedihardjo (2024). On the other hand, it has been shown in Niles-Weed and Rigollet (2022) that even for clean i.i.d. data, any estimator $\hat{P}$ has error $W_{1,k'}(\hat{P}, P) = \Omega(c_d n^{-1/\max(k,2)} + \sqrt{d/n})$ for a dimension-dependent term c_d. Thus, the error guarantee is tight up to the suboptimality of $W_{1,k'}(S_0, P)$, which we leave for future work.

1.3. Overview of Techniques

We start by highlighting the key challenge toward showing Theorem 3.

Local corruptions can destroy higher moment structure Consider a stable set $S_0 = \{x_i\}_{i \in [n]}$ with identity covariance and its perturbed version $S = \{x_1 + 0.5\rho n v, x_2 - 0.5\rho n v, x_3, \ldots, x_n\}$ where $\|v\|_2 = 1$. The covariance of S is inflated by $\Omega(\rho^2 n)$ which violates the stability condition for large n. However, the issue here is caused by only a few points, which can be ignored (as global outliers). We aim to show this always holds, i.e., there exists a *large subset* $S' \subset S$ of the perturbed data that remains stable, implying that stability-based algorithms can still work, proving Theorem 3. For our distribution learning result, we need an analogous claim for a generalized notion of stability, defined below (we denote $\mu_S = \frac{1}{|S|}\sum_{x \in S} x$, and $\overline{\mathbf{\Sigma}}_S = \frac{1}{|S|}\sum_{x \in S}(x - \mu)(x - \mu)^\top$, $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{A}\mathbf{B}^\top)$):

Definition 7 (Generalized stability) *Let $\epsilon \in (0, 1/2)$ and $\delta \in [\epsilon, \infty)$. We say that a set S of points in $\mathbb{R}^d$ satisfies the (ϵ, δ, k)-generalized-stability with respect to $\mu \in \mathbb{R}^d$ if for all $S' \subseteq S$ with $|S'| \geq (1 - \epsilon)S$, the following hold (where $\mathcal{V}_k$ denotes the set of all rank-k projection matrices): (i) $\|\mu_{S'} - \mu\|_2 \leq \delta$, and (ii) for every $\mathbf{V} \in \mathcal{V}_k$, $\left|\left\langle \mathbf{V}, \overline{\mathbf{\Sigma}}_{S'} - \mathbf{I} \right\rangle\right| \leq \delta^2/\epsilon$.*

As highlighted above, the difficulty comes from showing the second property in Definition 7. Let us consider $S' = S$ for simplicity for now. The variance-like quantity $\langle \mathbf{V}, \overline{\mathbf{\Sigma}}_S - \mathbf{I} \rangle$ is mainly

[4] By a standard property of sliced-Wasserstein distance, $\|\mu_{\hat{S}} - \mu_{S_0}\|_2 \lesssim W_{1,1}(\hat{S}, S_0)$.

[5] Recall that for nice distribution families, δ is a function of ϵ from Theorem 20; importantly, $\delta = o(\sqrt{\epsilon})$ for Gaussians and distributions with $t > 2$ bounded moments.

composed of two terms (ignoring the cross terms): (i) the covariance of the unperturbed data S_0, $\langle \mathbf{V}, \overline{\mathbf{\Sigma}}_{S_0} - \mathbf{I} \rangle$ and (ii) the second moment of the local perturbations: $\frac{1}{n}\sum_{i=1}^n \|\mathbf{V}\Delta_i\|_2^2$. If S_0 is (ϵ, δ, k)-stable the first term is at most δ^2/ϵ. If we could additionally show that $\frac{1}{n}\sum_{i=1}^n \|\mathbf{V}\Delta_i\|_2^2 \lesssim \rho^2/\epsilon$, then overall, $\langle \mathbf{V}, \overline{\mathbf{\Sigma}}_S - \mathbf{I} \rangle$ would be $O((\delta + \rho)^2/\epsilon)$, meaning that S is $(\epsilon, \widetilde{\delta}, k)$-generalized stable with $\widetilde{\delta} = O(\delta + \rho)$. The formal version of the argument of this paragraph is Proposition 14.

In order to show Theorem 3 thus, it suffices to show existence of a $(1-\epsilon)n$-sized subset $\mathcal{I} \in [n]$ with $\frac{1}{\mathcal{I}}\sum_{i \in \mathcal{I}} \|\mathbf{V}\Delta_i\|_2^2 \lesssim \rho^2/\epsilon$. As we highlight below, achieving this differs significantly for the weak and strong local contamination models.

Differences between weak and strong local contamination Let $\{\Delta_i\}_{i\in[n]}$ and $\{\Delta_i'\}_{i\in[n]}$ be the perturbation vectors in $\mathbb{R}^d$ defined as $\widetilde{x}_i - x_i$ in StrongWC Model and WeakLC Model respectively, i.e.,

$$\sup_{\mathbf{V}\in\mathcal{V}_k} \tfrac{1}{n}\sum_{i\in[n]} \|\mathbf{V}\Delta_i\|_2 \le \rho, \quad \tfrac{1}{n}\sum_{i\in[n]} \|\Delta_i'\|_2 \le \rho, \tag{2}$$

where $\mathcal{V}_k$ is the set of all rank-k projection matrices. For weak local contamination, we can take $\mathcal{I} \subset [n]$ as the indices of the $(1-\epsilon)n$ vectors in $\{\Delta_i'\}_{i\in[n]}$ with the smallest Eucledian norms and verify that for any $\mathbf{V} \in \mathcal{V}_k$, $\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \|\Delta_i'\|_2^2 \le \max_{i\in\mathcal{I}} \|\Delta_i'\|_2 \cdot \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \|\Delta_i'\|_2 \lesssim \frac{\rho}{\epsilon}\cdot\rho \lesssim \frac{\rho^2}{\epsilon}$. Here, we applied Markov's inequality to bound $\|\Delta_i'\|_2 \le \rho/\epsilon$ for all $i \in \mathcal{I}$ and used the weak local perturbation condition $\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \|\Delta_i'\|_2 \le \rho$. This strategy is used implicitly in Nietert et al. (2024).[6]

Let $\{\Delta_i\}_{i\in[n]}$ now be perturbations according to the strong local contamination (i.e., satisfying the first inequality in (2)). A natural idea is to adapt this "truncation & Markov" proof strategy in a *direction-dependent* manner. For a "direction" $\mathbf{V} \in \mathcal{V}_k$, we can define $\mathcal{I}_{\mathbf{V}} \subset [n]$ to be the set of $(1-\epsilon)n$ many indices with the smallest $\|\mathbf{V}\Delta_i\|_2$. Following the arguments of the previous paragraph, we find that $\max_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{|\mathcal{I}_{\mathbf{V}}|}\sum_{i\in\mathcal{I}_{\mathbf{V}}} \|\mathbf{V}\Delta_i\|_2^2 \lesssim \frac{\rho^2}{\epsilon}$. However, the order of quantifiers of $\mathbf{V}$ and the $\mathcal{I}_{\mathbf{V}}$ is reversed compared to what we want: we would like to find a single subset that works for every $\mathbf{V}$.

Toward tackling strong local contamination In what follows, we show that the order of quantifiers can actually be fixed by establishing the following statement in this section:

Proposition 8 *Let points $\Delta_i \in \mathbb{R}^d$ with $\max_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{n}\sum_{i\in[n]} \|\mathbf{V}\Delta_i\|_2 \le \rho$. Then for every $\epsilon \in (0,1)$ there exists $\mathcal{I} \subseteq [n]$ such that (i) $|\mathcal{I}| \ge (1-\epsilon)n$ and (ii) for all $\max_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \|\mathbf{V}\Delta_i\|_2^2 \lesssim \rho^2/\epsilon$.*

Our proof strategy builds on Steinhardt et al. (2018) and Diakonikolas et al. (2020), with key differences. While the former includes a similar result for the $k=1$ case, their approach does not seem to capture the rank-k sliced distance, preventing a generalization to Proposition 8, which is crucial for our distribution learning result. Meanwhile, Diakonikolas et al. (2020) focuses on sample complexity for stability (Theorem 20) in the $k=1$ case, rather than the deterministic statement above.

We now sketch the proof of Proposition 8. Instead of directly optimizing over all large subsets $\mathcal{I}$, we first perform a convex relaxation by defining $\Delta_{n,\epsilon} := \{w \in \mathbb{R}_+^n : \sum_{i=1}^n w_i = 1; 0 \le w_i \le 1/(n(1-\epsilon))\}$. A rounding argument (Lemma 41 from Diakonikolas et al. (2020)) shows that for proving Proposition 8 it suffices to show that $\min_{w\in\Delta_{n,\epsilon}} \max_{\mathbf{V}\in\mathcal{V}_k} \sum_i w_i \|\mathbf{V}\Delta_i\|_2^2 \lesssim \rho^2/\epsilon$. As noted earlier, our "truncation & Markov" argument from the earlier paragraphs could show this bound if the order of quantifiers for w and $\mathbf{V}$ were reversed. To achieve this reversal, we first convexify the maximization variable by defining $\mathcal{M}_k := \{\mathbf{M} \in \mathbb{R}^{d\times d} : 0 \preceq \mathbf{M} \preceq \mathbf{I}; \operatorname{tr}(\mathbf{M}) = k\}$ (which is the convex hull of $\mathbf{V} \in \mathcal{V}_k$), and then apply min-max duality for bilinear programs over convex compact sets:[7]

[6] However Nietert et al. (2024) does not achieve the optimal dependence on the stability parameter after truncation.

[7] Throught the section, we will often also use that $\|\mathbf{V}\Delta_i\|_2^2 = \Delta_i^\top \mathbf{V}^\top \mathbf{V}\Delta_i = \Delta_i^\top \mathbf{V}\Delta_i$.

$$\min_{w\in\Delta_{n,\epsilon}}\max_{\mathbf{V}\in\mathcal{V}_k}\sum_{i=1}^{n}w_i\|\mathbf{V}\Delta_i\|_2^2=\min_{w\in\Delta_{n,\epsilon}}\max_{\mathbf{M}\in\mathcal{M}_k}\sum_{i=1}^{n}w_i\Delta_i^\top\mathbf{M}\Delta_i=\max_{\mathbf{M}\in\mathcal{M}_k}\min_{w\in\Delta_{n,\epsilon}}\sum_{i=1}^{n}w_i\Delta_i^\top\mathbf{M}\Delta_i. \quad (3)$$

While the order of quantifiers is reversed we are now faced with the new challenge: Even though the left hand side above is small $O(\rho^2/\epsilon)$, the right hand side (over $\mathcal{M}_k$) might be large. [8] At a high-level, this is because we do not have guarantees on the behavior of $\{\Delta_i^\top\mathbf{M}\Delta_i\}_{i=1}^n$ for a general $\mathbf{M}\in\mathcal{M}_k$.

Towards that, we note that if $\sup_{\mathbf{M}\in\mathcal{M}_k}\frac{1}{n}\sum_{i=1}^n\sqrt{\Delta_i^\top\mathbf{M}\Delta_i}\lesssim\rho$, then the "truncation & Markov" argument would become applicable, showing that the right-hand side in Eq. (3) is indeed $O(\rho^2/\epsilon)$. In order to show this upper bound, we develop a Gaussian rounding scheme, extending the approach of Depersin and Lecué (2022). This is done Proposition 13 where we show that $\sup_{\mathbf{M}\in\mathcal{M}_k}\sum_{i=1}^n\frac{1}{n}\sqrt{\Delta_i^\top\mathbf{M}\Delta_i}\lesssim\sup_{\mathbf{V}\in\mathcal{V}_k}\frac{1}{n}\sum_{i\in[n]}\|\mathbf{V}\Delta_i\|_2$. Noting that the RHS is at most ρ by definition of our contamination model completes the proof.

Distribution learning We again use Propositions 8, 14 to argue that there is large stable subset of our dataset. However, there are no pre-existing stability-based algorithms for distribution learning. We develop one inspired by standard filtering techniques from Diakonikolas and Kane (2023) and a certificate lemma from Nietert et al. (2024). This is outlined in Section 5, with the full proof in Appendix G.

2. Preliminaries

We only provide the necessary preliminaries here; the full version can be found in Appendix B.

Basic notation We use $[n]:=\{1,\dots,n\}$, $\|x\|_2$ for the Eucledian norm, and $\mathbf{I}$ for the identity matrix. A symmetric matrix $\mathbf{A}$ is PSD (positive semidefinite), written as $\mathbf{A}\succeq 0$, if $x^\top\mathbf{A}x\geq 0$ for all $x\in\mathbb{R}^d$. We use $\|\mathbf{A}\|_{\text{op}}$ for the operator norm, $\text{tr}(\mathbf{A})$ for the trace, and $\langle\mathbf{A},\mathbf{B}\rangle=\text{tr}(\mathbf{A}\mathbf{B}^\top)$ for the Frobenius inner product. The *Mahalanobis* norm w.r.t. a PSD matrix $\mathbf{M}$ is $\|x\|_{\mathbf{M}}:=\sqrt{x^\top\mathbf{M}x}$. The indicator function of an event $\mathcal{E}$ is $\mathbb{1}_{\mathcal{E}}$. We write $a\lesssim b$ for $a\leq Cb$ for some absolute constant $C>0$. For $S\subset\mathbb{R}^d$, the sample mean, covariance, and centered second moment matrix (relative to μ, clear from context) are $\mu_S=\frac{1}{|S|}\sum_{x\in S}x$, $\mathbf{\Sigma}_S=\frac{1}{|S|}\sum_{x\in S}(x-\mu_S)(x-\mu_S)^\top$, $\overline{\mathbf{\Sigma}}_S=\frac{1}{|S|}\sum_{x\in S}(x-\mu)(x-\mu)^\top$.

Projection matrices and convex relaxations We use $\mathcal{V}_k$ to denote the set of all rank-k projection matrices in $\mathbb{R}^{d\times d}$ Recall that for any $\mathbf{V}\in\mathcal{V}_k$, $\mathbf{V}$ is symmetric, PSD, and idempotent. We use $\mathcal{M}_k$ to denote the set of convex relaxation of $\mathcal{V}_k$, i.e., $\mathcal{M}_k:=\{\mathbf{M}\in\mathbb{R}^{d\times d}:\mathbf{M}\succeq 0,\ \mathbf{M}\preceq\mathbf{I},\ \text{tr}(\mathbf{M})=k\}$.

2.1. Generalized Rank-k Stability

Recall Definition 7, generalizing Definition 1 for $k>1$. By convexity, it can be seen that we can replace "for every $\mathbf{V}\in\mathcal{V}_k$" with "for every $\mathbf{M}\in\mathcal{M}_k$" in Definition 7. We also note the alternative definitions:

Definition 9 (Generalized stability; alternative definitions) *Let $\epsilon\in(0,1/2)$ and $\delta\in[\epsilon,\infty)$. Let $S\subset\mathbb{R}^d$ and μ be a vector. Each of the following two bullets is equivalent*[9] *to Definition 7:*

- *S satisfies: (i) $\|\mu_S-\mu\|_2\leq\delta$, (ii) for all $\mathbf{M}\in\mathcal{M}_k$, $\left\langle\mathbf{M},\overline{\mathbf{\Sigma}}_S-\mathbf{I}\right\rangle\leq\delta^2/\epsilon$, and (iii) for all $S'\subset S$ with $|S'|\geq(1-\epsilon)|S|$ and for all $\mathbf{M}\in\mathcal{M}_k$, it holds $\left\langle\mathbf{M},\overline{\mathbf{\Sigma}}_{S'}-\mathbf{I}\right\rangle\geq-\delta^2/\epsilon$.*
- *S satisfies (i),(ii) as above, and for all $T\subset S$ with $|T|\leq\epsilon|S|$ and all $\mathbf{M}\in\mathcal{M}_k$, $\frac{|T|}{|S|}\langle\mathbf{M},\overline{\mathbf{\Sigma}}_T\rangle\leq\frac{\delta^2}{\epsilon}$.*

[8] This is because of the non-linearity induced by the $\min_w$ operator (if it was linear in $\mathbf{M}$, then the maximum over $\mathcal{V}_k$ and the analogous maximum over $\mathbf{M}\in\mathcal{M}_k$ would have been equal by convexity).

[9] Up to a constant factor in the resulting stability parameter δ.

2.2. Consequences of (Generalized) Stability

We state key stability facts used throughout this paper, with proofs in Appendix B. The first connects standard stability (Definition 1) to its generalized variant (Definitions 7,9). The second shows that all large subsets of a stable set remain stable. The last shows that all large subsets of a stable set are close in the sliced-Wasserstein metrics (Definition 5).

Lemma 10 *Let S be an (ϵ, δ)-stable set with respect to μ. Then S is also (ϵ, δ', k)-generalized stable with respect to μ with $\delta' \lesssim \delta\sqrt{k}$.*

Lemma 11 *Let S be (ϵ, δ, k)-generalized stable with respect to μ. Let $r \geq 1$ be such that $r\epsilon \leq 1/2$. Then: (i) S is also $(r\epsilon, \delta', k)$-generalized stable with respect to μ for $\delta' \lesssim \delta\sqrt{r}$, (ii) any subset $S' \subset S$ such that $|S'| \geq (1 - r\epsilon)|S|$, is also (ϵ, δ', k)-generalized-stable with respect to μ with $\delta' \lesssim \delta\sqrt{r}$.*

Lemma 12 *Let S_0 be a set satisfying (ϵ, δ, k)-generalized stability, and S_0' be a subset of S_0 with $|S_0'| \geq (1 - \epsilon)|S_0|$ for $\epsilon \leq 1/2$. Then, $W_{1,k}(S_0, S_0') \lesssim \epsilon\sqrt{k} + \delta$ and $W_{2,k}(S_0, S_0') \lesssim \sqrt{\epsilon k} + \delta/\sqrt{\epsilon}$.*

3. Averages of Low-rank Projections and Their Convex Relaxations

In this section, we establish a key structural property of local perturbations and their (generalized) projections. Let $\{\Delta_i\}_{i=1}^n$ be perturbations satisfying StrongWC Model (sliced Wasserstein distance). A first step toward our main result of ensuring stability of the perturbed data is to show that $\{\|\Delta_i\|_{\mathbf{M}}\}_{i\in[n]}$ are nicely bounded for any $\mathbf{M} \in \mathcal{M}_k$. While this holds for projections $\mathbf{V} \in \mathcal{V}_k$ by definition of the StrongWC Model, our proof strategy in Proposition 8 required similar bounds for $\mathbf{M} \in \mathcal{M}_k$, because it relied on a convex relaxation to apply min-max duality. The following proposition thus will complete the missing step in the proof of Proposition 8.[10]

Proposition 13 (Bound on average projections) *Let $y_1, y_2, \ldots, y_n$ be vectors in $\mathbb{R}^d$. Then the following holds: $\sup_{\mathbf{M}\in\mathcal{M}_k} \frac{1}{n}\sum_{i=1}^n \|y_i\|_{\mathbf{M}} \lesssim \sup_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{n}\sum_{i=1}^n \|y_i\|_{\mathbf{V}}$.*

We prove this using a Gaussian rounding scheme inspired by Depersin and Lecué (2022). Although the proof is technical, we present an informal proof sketch is below; the full proof is in Appendix D.

The idea comes from viewing elements $\mathbf{M}$ of $\mathcal{M}_k$ as $\mathbf{M} = \mathbf{E}_{g_1,\ldots,g_k\sim\mathcal{N}(0,\mathbf{M})}[\frac{1}{k}\sum_{i=1}^k g_i g_i^\top]$ and trying to rewrite the LHS of the inequality that we want to show (the conclusion of Proposition 13) as an expectation of projections $\|y_i\|_{\mathbf{V}}$. More concretely, let the random variable $Z = \frac{1}{n}\sum_{i=1}^n \sqrt{y_i^\top \mathbf{B} y_i}$, where $\mathbf{B} = \frac{1}{k}\sum_{i=1}^k g_i g_i^\top$ and $g_1, \ldots, g_k \sim \mathcal{N}(0, \mathbf{M})$. The proof strategy is to establish the following bounds for every $\mathbf{M} \in \mathcal{M}_k$:

$$\frac{1}{n}\sum_{i=1}^n \sqrt{y_i^\top \mathbf{M} y_i} \lesssim \mathbf{E}_{g_1,\ldots,g_k}[Z] \lesssim \sup_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{n}\sum_{i=1}^n \sqrt{y_i^\top \mathbf{V} y_i}\ . \tag{4}$$

The first bound above, when rewritten as $\frac{1}{n}\sum_{i=1}^n \sqrt{\mathbf{E}[y_i^\top \mathbf{B} y_i]} \lesssim \frac{1}{n}\sum_{i=1}^n \mathbf{E}[\sqrt{y_i^\top \mathbf{B} y_i}]$, it resembles the opposite of Jensen's inequality. However, because $\mathbf{B}$ has the very special form of the average of rank-one Gaussian vectors, we can prove the inequality using Gaussian properties (as shown in the formal proof in Appendix D, the bound essentially reduces to the fact that for a Gaussian $x \sim \mathcal{N}(0, 1)$, $\sqrt{\mathbf{E}[x^4]} \leq \sqrt{3}\,\mathbf{E}[x^2]$). Regarding the second bound in Equation (4), applying the

[10] Recall our notation $\|y\|_{\mathbf{M}} = \sqrt{y^\top \mathbf{M} y}$.

Cauchy–Schwarz inequality gives $Z \leq \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{B}^{1/2}\|_{\mathrm{op}}\|y_i\|_{\mathbf{B}}$. Then, we note that: (i) with high probability, $\|\mathbf{B}^{1/2}\|_{\mathrm{op}} = O(1)$, and (ii) $\|y_i\|_{\mathbf{B}} \lesssim \sup_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{n}\sum_{i=1}^{n}\sqrt{y_i^\top \mathbf{V} y_i}$ because $\mathbf{B}$ is rank-k with high probability. These imply $Z \leq \sup_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{n}\sum_{i=1}^{n}\sqrt{y_i^\top \mathbf{V} y_i}$ with high probability. Although for Equation (4), we need this inequality to hold in expectation (rather than with high probability), this can be achieved by modifying the proof slightly (see the formal argument in Appendix D).

4. Stability Is Preserved Under Local Perturbations

We now present the main technical result (Proposition 14) behind Theorems 3 and 6. First, recall Proposition 8. We will use it to get the conclusion that $\max_{\mathbf{M}\in\mathcal{M}_k} \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \|\Delta_i\|_{\mathbf{M}}^2 = \max_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \|\mathbf{V}\Delta_i\|_2^2 \lesssim \rho^2/\epsilon$, where the first step is because $\mathcal{M}_k$ is the convex hull of $\mathcal{V}_k$ and $\mathbf{V}\mathbf{V}^\top = \mathbf{V}$ for all $\mathbf{V} \in \mathcal{V}_k$. Then, the plan is to use the above with Proposition 14 below, stating that as long as local perturbations have bounded second moment, the stability parameter degrades in a dimension-independent manner and the the set remains close to the original in Wasserstein metric. [11]

Proposition 14 *Let $S_0' = \{x_1, \ldots, x_n\}$ be an (ϵ, δ, k)-generalized stable set with respect to $\mu \in \mathbb{R}^d$ and $\Delta_i \in \mathbb{R}^d$ such that*

$$\max_{\mathbf{M}\in\mathcal{M}_k} \frac{1}{n}\sum_{i\in[n]} \|\Delta_i\|_{\mathbf{M}} \lesssim \rho \qquad \textit{and} \qquad \max_{\mathbf{M}\in\mathcal{M}_k} \frac{1}{n}\sum_{i\in[n]} \|\Delta_i\|_{\mathbf{M}}^2 \lesssim \rho^2/\epsilon\ . \tag{5}$$

Define $\widetilde{x}_i := x_i + \Delta_i$ for all $i \in [n]$ and define the set S' to be $\{\widetilde{x}_i\}_{i\in[n]}$. Then the following hold:

- *S' satisfies $(\epsilon, \widetilde{\delta}, k)$-generalized stability with respect to μ (Definition 7) for $\widetilde{\delta} \lesssim \delta + \rho$.*
- *For all subsets $S'' \subset S'$ with $|S''| \geq (1-\epsilon)|S'|$ we have $W_{1,k}(S_0', S'') \lesssim \rho + \epsilon\sqrt{k} + \delta$ and $W_{2,k}(S_0', S'') \lesssim \sqrt{\epsilon k} + \delta/\sqrt{\epsilon} + \rho/\sqrt{\epsilon}$.*

We prove each bullet point of Proposition 14 separately in Section 4.1 and Section 4.2.

4.1. Proof of First Part of Proposition 14

We use $\mu = 0$ throughout this proof without loss of generality. By Lemma 18 our goal is equivalent (up to a constant factor in $\widetilde{\delta}$) to establishing the following conditions (simultaneously). We will establish these three conditions separately.

1. (Mean) $\|\mu_{S'}\|_2 \leq \widetilde{\delta}$.
2. (Upper bound on covariance) For all $\mathbf{M} \in \mathcal{M}_k$ we have $\left\langle \mathbf{M}, \overline{\mathbf{\Sigma}}_{S'} - \mathbf{I}\right\rangle \leq \widetilde{\delta}^2/\epsilon$.
3. (Lower bound on large subsets) For all $S'' \subset S'$ with $|S''| \geq (1-\epsilon)|S'|$ and all $\mathbf{M} \in \mathcal{M}_k$ we have $\left\langle \mathbf{M}, \overline{\mathbf{\Sigma}}_{S'} - \mathbf{I}\right\rangle \geq -\widetilde{\delta}^2/\epsilon$.

Proof of the mean condition We start with the mean condition, which follows directly by the triangle inequality, the stability of the original points $\{x_i\}_{i\in[n]}$, and the assumption (5): $\|\frac{1}{n}\sum_{i\in[n]} \widetilde{x}_i\|_2 \leq \|\frac{1}{n}\sum_{i\in[n]} x_i\|_2 + \|\frac{1}{n}\sum_{i\in[n]} \Delta_i\|_2 \lesssim \delta + \sup_{v\in\mathbb{R}^d:\|v\|_2=1} \sum_{i=1}^{n} |v_i^\top \Delta_i| \lesssim \delta + \rho$.

[11] The last bullet of Proposition 14 is only relevant to Theorem 6.

Proof of upper bound on covariance Using the decomposition $\widetilde{x}_i = x_i + \Delta_i$ yields the following:

$$\left\langle \mathbf{M}, \overline{\mathbf{\Sigma}}_{S'} - \mathbf{I} \right\rangle = \tfrac{1}{n}\textstyle\sum_{i\in[n]} \|\widetilde{x}_i\|_{\mathbf{M}}^2 - \operatorname{tr}(\mathbf{M}) = \tfrac{1}{n}\textstyle\sum_{i\in[n]} \left(\|x_i\|_{\mathbf{M}}^2 - \operatorname{tr}(\mathbf{M}) + \|\Delta_i\|_{\mathbf{M}}^2 + 2x_i^\top \mathbf{M}\Delta_i \right) . \quad (6)$$

We now bound each of the terms above separately.

- The first part of Equation (6) can be handled by the generalized stability of the original points in S_0'. Since $\mathbf{M} \in \mathcal{M}_k$, and S_0 is (ϵ, δ, k)-stable, we have $\frac{1}{n}\sum_{i\in[n]} \|x_i\|_{\mathbf{M}}^2 - \operatorname{tr}(\mathbf{M}) \le \delta^2/\epsilon$.
- The next term in (6) is at most ρ^2/ϵ by Equation (5).
- We finally bound the cross term in (6). For any fixed $\mathbf{M} \in \mathcal{M}$, define $\mathcal{I}_{\mathbf{M}}$ to be the set of indices in $[n]$ such that $\|x_i\|_{\mathbf{M}}$ is bigger than $C\delta/\epsilon$. For a large enough constant C, the condition in Definition 9 implies that $|\mathcal{I}_{\mathbf{M}}| \le \epsilon n$. Combining this upper bound on the cardinality of $\mathcal{I}_{\mathbf{M}}$ with the last condition in Definition 9 we have that $\frac{1}{n}\sum_{i\in\mathcal{I}_{\mathbf{M}}} \|x_i\|_{\mathbf{M}}^2 \lesssim \frac{\delta^2}{\epsilon}$. Using the Cauchy-Schwarz inequality, we obtain that for any $\mathbf{M} \in \mathcal{M}_k$:

$$\frac{1}{n}\sum_{i\in[n]} |x_i^\top \mathbf{M}\Delta_i| \le \frac{1}{n}\sum_{i\in[n]} \|\Delta_i\|_{\mathbf{M}}\|x_i\|_{\mathbf{M}} \le \frac{1}{n}\sum_{i\in[n]} \|\Delta_i\|_{\mathbf{M}}\|x_i\|_{\mathbf{M}} \mathbb{1}_{i\notin\mathcal{I}_{\mathbf{M}}} + \frac{1}{n}\sum_{i\in[n]} \|\Delta_i\|_{\mathbf{M}}\|x_i\|_{\mathbf{M}} \mathbb{1}_{i\in\mathcal{I}_{\mathbf{M}}}$$

$$\lesssim \frac{1}{n}\sum_{i\in[n]} \frac{\delta}{\epsilon}\|\Delta_i\|_{\mathbf{M}} + \sqrt{\frac{1}{n}\sum_{i\in[n]} \|\Delta_i\|_{\mathbf{M}}^2}\sqrt{\frac{1}{n}\sum_{i\in\mathcal{I}_{\mathbf{M}}} \|x_i\|_{\mathbf{M}}^2} \lesssim \frac{\delta\rho}{\epsilon} + \sqrt{\frac{\rho^2}{\epsilon}}\sqrt{\frac{\delta^2}{\epsilon}} \lesssim \frac{\delta\rho}{\epsilon} , \quad (7)$$

where the last step uses Equation (5) and $\frac{1}{n}\sum_{i\in\mathcal{I}_{\mathbf{M}}} \|x_i\|_{\mathbf{M}}^2 \lesssim \frac{\delta^2}{\epsilon}$ which we have shown earlier.

Combining everything, we have shown that for all matrices $\mathbf{M} \in \mathcal{M}_k$, the following bound holds: $\langle \mathbf{M}, \overline{\mathbf{\Sigma}}_{S'} - \mathbf{I} \rangle \lesssim \frac{\delta^2}{\epsilon} + \frac{\rho^2}{\epsilon} + \frac{\delta\rho}{\epsilon}$. Therefore $\langle \mathbf{M}, \overline{\mathbf{\Sigma}}_{S'} - \mathbf{I} \rangle \lesssim \widetilde{\delta}^2/\epsilon$ for some $\widetilde{\delta} \lesssim \delta + \rho$.

Proof of lower bound on covariance For any $\mathbf{M}$ and any $S'' \subset S'$, similarly to Equation (6) we have

$$\tfrac{1}{|S''|}\textstyle\sum_{i:\widetilde{x}_i\in S''} \|\widetilde{x}_i\|_{\mathbf{M}}^2 - \operatorname{tr}(\mathbf{M}) = \tfrac{1}{|S''|}\textstyle\sum_{i:\widetilde{x}_i\in S''} \left(\left(\|x_i\|_{\mathbf{M}}^2 - \operatorname{tr}(\mathbf{M})\right) + \|\Delta_i\|_{\mathbf{M}}^2 + 2x_i^\top \mathbf{M}\Delta_i \right) .$$

Since $|S''| \ge (1-\epsilon)|S'| \ge (1-2\epsilon)|S_0'|$, the first term $(1/n)\sum_i \|x_i\|_{\mathbf{M}}^2 - \operatorname{tr}(\mathbf{M})$ is at least $-O(\delta^2/\epsilon)$ by the generalized-stability of S_0' (the first bullet in Definition 9) and Lemma 11. The second term is non-negative. The last (cross term) has absolute value at most $\delta\rho/\epsilon$ because of Equation (7). Thus, we have shown the following lower bound on the covariance: for all subsets $S'' \subset S$ with $|S''| \ge (1-\epsilon)|S'|$, it holds that $\langle \mathbf{M}, \overline{\mathbf{\Sigma}}_{S''} - \mathbf{I} \rangle \ge -\widetilde{\delta}^2/\epsilon$ for some $\widetilde{\delta} \lesssim \delta + \rho$.

4.2. Proof of Second Part of Proposition 14

Due to space constraints, the full version with additional details on each step is provided in Appendix E. Using the triangle inequality we have $W_{1,k}(S_0', S'') \le W_{1,k}(S_0', S') + W_{1,k}(S', S'')$. The first term is bounded by definition of Definition 5: if x_i denote the points in S_0' and $x_i + \Delta_i$ are the ones in S', then $W_{1,k}(S_0', S') \le \max_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{|S_0'|}\sum_{i:i\in S_0'} \|\Delta_i\|_2 \lesssim \rho$, where the last inequality follows by Equation (5). For the second term we have $W_{1,k}(S', S'') \lesssim \epsilon\sqrt{k} + \rho + \delta$ by Lemma 12 (S' is $(\epsilon, O(\delta + \rho), k)$-generalized-stable). The bound for $W_{2,k}(S_0', S'')$ can be shown following similar steps.

5. Mean Estimation and Distribution Learning

In this section, we provide informal proof sketches for Theorems 3 and 6. The full versions are in Appendices F and G, respecticely.

Mean estimation If $S_0 = \{x_i\}_{i \in [n]}$ denotes the original (ϵ, δ)-stable set (Definition 1), and Δ_i are the perturbations introduced to each point by the StrongLC Model, then by Proposition 8 (with k=1), there exists a $(1-\epsilon)n$-sized subset $\mathcal{I} \subset [n]$ such that $\max_{\mathbf{M} \in \mathcal{M}_1} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|\Delta_i\|_{\mathbf{M}}^2 = \max_{\mathbf{V} \in \mathcal{V}_1} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|\Delta_i\|_{\mathbf{V}}^2 \lesssim \rho^2/\epsilon$ (the first step follows by convexity). The, using Proposition 13, $\max_{\mathbf{M} \in \mathcal{M}_1} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|\Delta_i\|_{\mathbf{M}} \lesssim \max_{\mathbf{V} \in \mathcal{V}_1} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|\Delta_i\|_{\mathbf{V}} \leq \rho$. Finally, Proposition 14 implies that $S' = \{x_i + \Delta_i : i \in \mathcal{I}\}$ is $(\epsilon, O(\delta + \rho))$-stable and Theorem 3 follows.

Algorithm 1 Distribution learning under global and strong local corruptions

```
Input: (Multi)-Set of samples $T \subset \mathbb{R}^d$, and parameters $\epsilon \in (0, c), \delta \geq \epsilon, \rho \geq 0$.
Output: $\widehat{S} \subset \mathbb{R}^d$ and $\widehat{\mu} \in \mathbb{R}^d$ such that $\|\widehat{\mu} - \mu\|_2 \lesssim \delta + \rho$ and $W_{1,k'}(\widehat{S}, S) \lesssim \delta\sqrt{k'} + \rho, \forall k' \in [k]$.
Let $C$ be a sufficiently large absolute constant
for $k' = 1, 2, \ldots, k$ do
    Define $\widetilde{\delta} = \delta\sqrt{k'} + \rho$
    while true do
        Compute $\mathbf{M} \in \mathbb{R}^{d \times d}$ maximizing $\langle \mathbf{M}, \mathbf{\Sigma}_T \rangle$ under the constraints $0 \preceq \mathbf{M} \preceq \mathbf{I}$, $\mathrm{tr}(\mathbf{M}) = k'$
        if $\langle \mathbf{M}, \mathbf{\Sigma}_T \rangle \leq k' + C\widetilde{\delta}^2/\epsilon$ then        // Stopping condition
            Continue (exit while loop)
        else
            Let $L \subset T$ consist of the $\epsilon|T|$ points with the largest score $g(x) = (x - \mu_T)^\top \mathbf{M}(x - \mu_T)$
            Define the thresholded scores $\tau(x) := g(x)\mathbf{1}_{x \in L}$ for $x \in T$
            for each $x \in T$: Delete $x$ from $T$ with probability $\tau(x)/\max_{x \in T} \tau(x)$
Let $\widehat{S} \leftarrow T$
return $\widehat{S}$
```

Distribution learning The pseudocode is provided in Algorithm 1. For simplicity, let us ignore the outer for loop in the algorithm, and only discuss the case $k' = k$.

Let $S_0 = \{x_i\}_{i \in [n]}$ be the (ϵ, δ)-stable input set (also $(\epsilon, \delta\sqrt{k}, k)$-generalized-stable by Lemma 10). First, using Proposition 8, Proposition 13 and Proposition 14 we have that if Δ_i denote the local perturbations, there exists a core $\mathcal{I} \subset [n]$ with $|\mathcal{I}| \geq (1-\epsilon)n$ such that the set $S' := \{x_i + \Delta_i : i \in \mathcal{I}\}$ is $(\epsilon, O(\delta\sqrt{k} + \rho), k)$-generalized-stable.[12] The algorithm runs in iterations (*while* loop) with each iteration removing points from the current dataset T (*else* block) until the stopping condition $\langle \mathbf{M}, \mathbf{\Sigma}_T \rangle \leq k + C\widetilde{\delta}^2/\epsilon$ becomes true. Its correctness follows from two arguments, discussed individually next.

- If the stopping condition is true and T contains at least $(1-20\epsilon)n$ inliers, $W_{1,k}(T, S_0) \lesssim \delta\sqrt{k} + \rho$.
- If the stopping condition is false, filtering removes more outliers than inliers in expectation.

Certification of solutions The stopping condition is informed by Nietert et al. (2024) (Lemma 29), which bounds the $W_{1,k}$ distance between S' and the dataset T by a variance-like quantity of the outliers $T \setminus S'$. Since S' is unknown, the algorithm cannot compute this directly. Instead, we further bound it (Lemma 30) by a variance-like quantity of the entire T. Below, we state a simplified version for the purpose of this proof sketch; obtaining this formally requires careful technical work combining Lemmata 29 and 30, the bounds on $W_{1,k}, W_{2,k}$ from Proposition 14 and Lemmata 11 and 12.

[12] S' also satisfies the $W_{1,k}$ and $W_{2,k}$ bounds in the second part of Proposition 14, which will also be needed for the technical work of this proof but these calculations will not be highlighted in the proof sketch of this section.

Lemma 15 (Certificate lemma (informal)) *Consider the context of Algorithm 1, where T denotes the current dataset. Denote $\lambda := \langle \mathbf{M}, \mathbf{\Sigma}_T \rangle - k'$, let S_0 and S' be the sets defined earlier and assume $|T \cap S'| \geq (1 - 20\epsilon)$. We have $W_{1,k'}(T, S_0) \lesssim \delta\sqrt{k'} + \rho + \sqrt{\lambda\epsilon}$.*

When stopping condition becomes true, $\lambda \lesssim (\delta\sqrt{k'} + \rho)^2/\epsilon$, thus $W_{1,k'}(T, S_0) \lesssim \delta\sqrt{k'} + \rho$, as desired.

Filtering We can show that the scores used by the algorithm for rejecting points (inside the *else* block) assign significantly bigger values to global outliers than inliers, whenever the stopping condition is not true. The full version can be found in Lemma 31.

Lemma 16 (Filtering lemma (informal; see Lemma 31)) *Define $\widetilde{\delta} = \delta + \rho$. Let S' as above, and T with $|T \cap S'| \geq (1 - 20\epsilon)$. Let $\tau(x)$ be the scores defined in Algorithm 1. Then if $\langle \mathbf{M}, \mathbf{\Sigma}_T \rangle > k' + C\widetilde{\delta}^2/\epsilon$ we have that $\sum_{x \in S' \cap T} \tau(x) \leq 0.1 \sum_{x \in T} \tau(x)$.*

Thus, until the while loop exits, removing each $x \in T$ with probability $\tau(x)/\max_{x \in T} \tau(x)$ in expectation rejects a constant factor more outliers than inliers. A standard martingale analysis (e.g., Diakonikolas and Kane (2023)) shows that with probability at least $9/10$, $|T \triangle S'| \leq 20\epsilon$ holds throughout the algorithm's execution.

So far, we have ignored the outer for-loop over $k' \in [k]$. To bound $W_{1,k'}(T, S_0) \lesssim \delta\sqrt{k'} + \rho$ uniformly for all $k' \in [k]$, note that S' from the first paragraph is independent of k' and satisfies $(\epsilon, O(\delta\sqrt{k'} + \rho), k')$-generalized stability for all k'. Repeating the algorithm for each $k' \in [k]$ ensures the filtered set meets the $W_{1,k'}$ bound for all k'.

Acknowledgments

We thank Po-Ling Loh for helpful comments and discussion on prior work.

Acknowledgments

We thank Po-Ling Loh for helpful comments and discussion on prior work. We also thank Sloan Nietert for his insightful question that led to a stronger conclusion in Theorem 6.

References

P. Abdalla and N. Zhivotovskiy. Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *Journal of the European Mathematical Society*, 2024.

Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

B. Barak and D. Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares. 1, 2016. URL `http://www.sumofsquares.org/public/index.html`.

M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar. The security of machine learning. *Machine Learning*, 2010.

B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *Proc. 29th International Conference on Machine Learning (ICML)*, 2012.

M. T. Boedihardjo. Sharp bounds for the max-sliced wasserstein distance. *arXiv preprint arXiv:2403.00666*, 2024.

A. Bora, E. Price, and A. G. Dimakis. Ambientgan: Generative models from lossy measurements. In *Proc. 6th International Conference on Learning Representations (ICLR)*, 2018.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5.

O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

P. Chao and E. Dobriban. Statistical estimation under distribution shift: Wasserstein perturbations and minimax theory. *arXiv preprint arXiv:2308.01853*, 2023.

Y. Cheng, I. Diakonikolas, and R. Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proc. 30th Annual Symposium on Discrete Algorithms (SODA)*, pages 2755–2771. SIAM, 2019. doi: 10.1137/1.9781611975482.171.

Y. Cheng, I. Diakonikolas, R. Ge, and M. Soltanolkotabi. High-Dimensional Robust Mean Estimation via Gradient Descent. In *Proc. 37th International Conference on Machine Learning (ICML)*, 2020.

S. Chewi, J. Niles-Weed, and P. Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 2024.

J. Depersin and G. Lecué. Robust sub-Gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 2022.

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

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust Estimators in High Dimensions without the Computational Intractability. In *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2016. doi: 10.1109/FOCS.2016.85.

I. Diakonikolas, D. M. Kane, and A. Pensia. Outlier Robust Mean Estimation with Subgaussian Rates via Stability. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. arXiv preprint: arXiv:2007.15618.

I. Diakonikolas, D. M. Kane, S. Karmalkar, A. Pensia, and T. Pittas. Robust sparse mean estimation via sum of squares. In *Proc. 35th Annual Conference on Learning Theory (COLT)*, 2022a.

I. Diakonikolas, D. M. Kane, A. Pensia, and T. Pittas. Streaming Algorithms for High-Dimensional Robust Statistics. In *Proc. 39th International Conference on Machine Learning (ICML)*, 2022b.

I. Diakonikolas, D. M. Kane, A. Pensia, and T. Pittas. Nearly-linear time and streaming algorithms for outlier-robust PCA. In *Proc. 40th International Conference on Machine Learning (ICML)*, 2023.

I. Diakonikolas, S. B. Hopkins, A. Pensia, and S. Tiegel. Sos certifiability of subgaussian distributions and its algorithmic applications. 2024. Available on arXiv: https://arxiv.org/abs/2410.21194.

Y. Dong, S. B. Hopkins, and J. Li. Quantum Entropy Scoring for Fast Robust Mean Estimation and Improved Outlier Detection. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

N. Fleming, P. Kothari, and T. Pitassi. Semialgebraic proofs and efficient algorithm design. *Found. Trends Theor. Comput. Sci.*, 2019.

E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.

J. Hayase, W. Kong, R. Somani, and S. Oh. Spectre: defending against backdoor attacks using robust statistics. In *Proc. 38th International Conference on Machine Learning (ICML)*, 2021.

S. B. Hopkins and J. Li. Mixture models, robustness, and sum of squares proofs. In *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*, 2018.

S. B. Hopkins and J. Li. How Hard Is Robust Mean Estimation? In *Proc. 32nd Annual Conference on Learning Theory (COLT)*, 2019.

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, March 1964. ISSN 0003-4851. doi: 10.1214/aoms/1177703732.

A. Jambulapati, J. Li, and K. Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

A. Jambulapati, S. Kumar, J. Li, S. Pandey, A. Pensia, and K. Tian. Black-box k-to-1-pca reductions: Theory and applications. In *Proc. 37th Annual Conference on Learning Theory (COLT)*, 2024.

M. Jirak, S. Minsker, Y. Shen, and M. Wahl. Concentration and moment inequalities for heavy-tailed random matrices. *arXiv preprint arXiv:2407.12948*, 2024.

V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 2017.

W. Kong, R. Somani, S. Kakade, and S. Oh. Robust meta-learning for mixed linear regression with small batches. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

P. K Kothari and D. Steurer. Outlier-robust moment-estimation via sum-of-squares. *arXiv preprint arXiv:1711.11581*, 2017.

P. K. Kothari, J. Steinhardt, and D. Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*, 2018. doi: 10.1145/3188745.3188970.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic Estimation of Mean and Covariance. In *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2016. doi: 10.1109/FOCS.2016.76.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991. ISBN 978-3-642-20211-7 978-3-642-20212-4. doi: 10.1007/978-3-642-20212-4.

J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 2008.

Z. Liu and P. Loh Loh. Robust W-GAN-based estimation under Wasserstein contamination. *Information and Inference: A Journal of the IMA*, 2022.

G. Lugosi and S. Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021. doi: 10.1214/20-AOS1961.

T. Manole, S. Balakrishnan, and L. Wasserman. Minimax confidence intervals for the Sliced Wasserstein distance. *Electronic Journal of Statistics*, 2022.

K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Simsekli. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. 2004. ISBN 978-1-4613-4691-3 978-1-4419-8853-9.

S. Nietert, Z. Goldfeld, and R. Cummings. Outlier-robust optimal transport: Duality, structure, and statistical analysis. 2022.

S. Nietert, Z. Goldfeld, and S. Shafiee. Robust distribution learning with local and global adversarial corruptions (extended abstract). In *Proc. 37th Annual Conference on Learning Theory (COLT)*, 2024.

J. Niles-Weed and P. Rigollet. Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 2022.

Roberto I Oliveira and Zoraida F Rico. Improved covariance estimation: optimal robustness and sub-gaussian guarantees under heavy tails. *The Annals of Statistics*, 52(5):1953–1977, 2024.

P. Paschou, J. Lewis, A. Javed, and P. Drineas. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of Medical Genetics*, 2010.

A. Pensia, V. Jog, and P. Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *CoRR*, abs/2009.12976, September 2020.

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, 2012.

N. Rosenberg, J. Pritchard, J. Weber, H. Cann, K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. Genetic structure of human populations. *Science*, 2002.

J. Steinhardt, P. W. Koh, and P. Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

J. Steinhardt, M. Charikar, and G. Valiant. Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers. In *Proc. 9th Innovations in Theoretical Computer Science Conference (ITCS)*, 2018.

K. Tikhomirov. Sample covariance matrices of heavy-tailed distributions. *International Mathematics Research Notices*, 2018(20):6254–6289, 2018.

B. Tran, J. Li, and A. Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

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

R. Van Handel. Structured random matrices. In *Convexity and concentration*, 2017.

H. Yang and L. Carlone. Certifiably optimal outlier-robust geometric perception: Semidefinite relaxations and scalable global optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 2024.

B. Zhu, J. Jiao, and J. Steinhardt. Generalized Resilience and Robust Statistics. *The Annals of Statistics*, 2019. doi: 10.1214/22-AOS2186.

B. Zhu, J. Jiao, and J. Steinhardt. Robust estimation via generalized quasi-gradients. *Information and Inference: A Journal of the IMA*, 2022a. doi: 10.1093/imaiai/iaab018.

B. Zhu, J. Jiao, and D. Tse. Deconstructing generative adversarial networks. *IEEE Transactions on Information Theory*, 2022b.

Appendix

Appendix A. Related Work

Robust statistics Our work lies broadly in the field of algorithmic robust statistics. We refer the reader to Diakonikolas and Kane (2023) for a recent book on the topic. Our work is most closely related to Steinhardt et al. (2018); Depersin and Lecué (2022); Diakonikolas et al. (2020), which we discuss in Section 1.3. As mentioned earlier, robust statistics has primarily focused on Global Contamination Model. Some notable exceptions include Zhu et al. (2022b, 2019); Liu and Loh (2022); Chao and Dobriban (2023); Nietert et al. (2024), discussed later on in detail.

Stability condition Regarding the stability condition and stability-based algorithms (Definitions 1, 2), there exist such algorithms based on convex programming (Diakonikolas et al., 2016; Steinhardt et al., 2018; Cheng et al., 2019), iterative filtering (Diakonikolas et al., 2016), gradient descent (Cheng et al., 2020; Zhu et al., 2022a). Over the years, these stability-based algorithms have been optimized to be near-optimal in other important aspects: runtime (Cheng et al., 2019; Dong et al., 2019; Depersin and Lecué, 2022), sample complexity (Diakonikolas et al., 2020), and memory (Diakonikolas et al., 2022b). Stability-based algorithms have found black-box consequences on other problems such as principal component analysis (Appendix H) and linear regression (Pensia et al., 2020).

Wasserstein perturbations Zhu et al. (2019) and Liu and Loh (2022) studied the problems of covariance estimation and linear regression under the Wasserstein-1 perturbations. Similarly, Chao and Dobriban (2023) investigated the Wasserstein-2 perturbations and developed minmax-optimal estimators under those models. To the best of our knowledge, the combined contamination model (Global+WeakLC Model) was first proposed and studied in Zhu et al. (2022b), inspired by chained perturbations in the computer vision literature (Bora et al., 2018). Liu and Loh (2022) also studied Global+WeakLC Model by focusing on the problems of covariance estimation and linear regression. However, all of these works focused on the statistical aspects (and *weak* local contamination), and did not provide computationally-efficient algorithms. Our work is most closely related to Nietert et al. (2024) who developed computationally-efficient algorithms for the combined contamination model (with weak local perturbations) in Global+WeakLC Model. In contrast, we study the stronger Global+StrongLC Model. We also obtain the improved dependence on ϵ for certain distribution families, which was phrased as an open question in their work.

Optimal transport Finally, we mention related works from the theory of optimal transport. In fact, the combined contamination model (Global+WeakLC Model) is closely related to the notion of *outlier-robust* optimal transport cost (Balaji et al., 2020; Nietert et al., 2022), but their focus is rather different. In fact, our results can be seen as learning when the samples are perturbed in outlier-robust *sliced* optimal transport cost. The sliced-Wasserstein distance has been studied in several recent works because it avoids the curse of dimensionality fundamental to the usual Wasserstein distance (Rabin et al., 2012; Nadjahi et al., 2020; Manole et al., 2022; Boedihardjo, 2024; Chewi et al., 2024).

Appendix B. Preliminaries

In this section, we provide the full version of the prelimiaries.

Basic notation We use $\mathbb{Z}_+$ for the set of positive integers and $[n]$ to denote $\{1, \dots, n\}$. For a vector x we denote by $\|x\|_2$ its Euclidean norm. Let $\mathbf{I}_d$ denote the $d \times d$ identity matrix (omitting the subscript when it is clear from the context). We use $\mathcal{S}^{d-1}$ to denote the set of points $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$. We use $\top$ for the transpose of matrices and vectors. For a subspace $\mathcal{V}$ of $\mathbb{R}^d$ of dimension m, we denote by $\mathbb{P}_{\mathcal{V}} \in \mathbb{R}^{d\times d}$ the orthogonal projection matrix of $\mathcal{V}$. That is, if the subspace $\mathcal{H}$ is spanned by the columns of the matrix $\mathbf{A}$, then $\mathbb{P}_{\mathcal{H}} := \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$. By slightly overloading notation, if $\mathbf{A}$ is a matrix, we will also use $\mathbb{P}_{\mathbf{A}}$ to denote the orthogonal projection matrix for the subspace spanned by the columns of $\mathbf{A}$. We say that a symmetric $d \times d$ matrix $\mathbf{A}$ is PSD (positive semidefinite) and write $\mathbf{A} \succcurlyeq 0$ if for all $x \in \mathbb{R}^d$ it holds $x^\top\mathbf{A}x \geq 0$. We use $\|\mathbf{A}\|_{\mathrm{op}}$ for the operator (or spectral) norm of the matrix $\mathbf{A}$. We use $\mathrm{tr}(\mathbf{A})$ to denote the *trace* of the matrix $\mathbf{A}$ and $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{tr}(\mathbf{A}\mathbf{B}^\top)$ to denote the *Frobenius inner product* between matrices $\mathbf{A}$ and $\mathbf{B}$. For a PSD matrix $\mathbf{M}$ and a vector x, $\|x\|_{\mathbf{M}} := \sqrt{x^\top\mathbf{M}x}$ denotes the *Mahalanobis norm* of x with respect to $\mathbf{M}$.

We write $x \sim D$ for a random variable x following the distribution D and use $\mathbf{E}[x]$ for its expectation. We use $\mathcal{N}(\mu, \mathbf{\Sigma})$ to denote the Gaussian distribution with mean μ and covariance matrix $\mathbf{\Sigma}$. We write $\mathbf{Pr}(\mathcal{E})$ for the probability of an event $\mathcal{E}$. We write $\mathbb{1}_{\mathcal{E}}$ for the indicator function of the event $\mathcal{E}$.

We use $a \lesssim b$ to denote that there exists an absolute universal constant $C > 0$ (independent of the variables or parameters on which a and b depend) such that $a \leq Cb$. Sometime, we shall abuse the notation and use $a = O(b)$ to denote the same to save space.

Projection matrices and convex relaxations We use $\mathcal{V}_k$ to denote the set of all rank-k projection matrices in $\mathbb{R}^{d\times d}$. Recall that for any $\mathbf{V} \in \mathcal{V}_k$, $\mathbf{V}$ is symmetric, PSD, and idempotent. We use $\mathcal{M}_k$ to denote the set of convex relaxation of $\mathcal{V}_k$, i.e.,

$$\mathcal{M}_k := \{\mathbf{M} \in \mathbb{R}^{d\times d} : \mathbf{M} \succeq 0,\ \mathbf{M} \preceq \mathbf{I},\ \mathrm{tr}(\mathbf{M}) = k\}. \tag{8}$$

Empirical mean and second moment matrices For a $S \subset \mathbb{R}^d$, we use the following notation for the sample mean, sample covariance, and the centered second moment matrix with respect to μ (which shall be clear from context), respectively:

$$\mu_S := \frac{1}{|S|}\sum_{x\in S} x, \quad \mathbf{\Sigma}_S := \frac{1}{|S|}\sum_{x\in S}(x-\mu_S)(x-\mu_S)^\top, \quad \overline{\mathbf{\Sigma}}_S := \frac{1}{|S|}\sum_{x\in S}(x-\mu)(x-\mu)^\top . \tag{9}$$

B.1. Generalized Rank-k Stability

As outlined in Section 1, our algorithm for mean estimation relies on the exact same *stability condition* developed in prior work. However, for our distribution learning result, our algorithm is a multi-dimensional generalization of the standard filtering, which requires us to consider an appropriate generalization of the stability condition, presented in Definition 7. For $k = 1$ this definition reduces to the standard stability condition.

Definition 7 (Generalized stability) *Let $\epsilon \in (0, 1/2)$ and $\delta \in [\epsilon, \infty)$. We say that a set S of points in $\mathbb{R}^d$ satisfies the (ϵ, δ, k)-generalized-stability with respect to $\mu \in \mathbb{R}^d$ if for all $S' \subseteq S$ with $|S'| \geq (1-\epsilon)S$, the following hold (where $\mathcal{V}_k$ denotes the set of all rank-k projection matrices): (i) $\|\mu_{S'} - \mu\|_2 \leq \delta$, and (ii) for every $\mathbf{V} \in \mathcal{V}_k$, $\left|\left\langle \mathbf{V}, \overline{\mathbf{\Sigma}}_{S'} - \mathbf{I}\right\rangle\right| \leq \delta^2/\epsilon$.*

Remark 17 *Using convexity arguments, it can be seen that we can replace the condition "for every $\mathbf{V} \in \mathcal{V}_k$" with "for every $\mathbf{M} \in \mathcal{M}_k$" (cf. (8)) in the second condition of Definition 7.*

B.1.1. EQUIVALENT DEFINITIONS OF GENERALIZED STABILITY

We will often need to use basic properties of the stability condition that follow directly from its definition. These properties are presented in Lemma 18 as equivalent ways to define the stability condition. These equivalences have been shown in the literature for the special case of $k = 1$ (see, e.g., Claim 4.1 in Diakonikolas et al. (2020) and Lemma 3.1 in Diakonikolas and Kane (2023)), but the proof readily extends to general k.

Lemma 18 *Definition 7, and the two definitions given in Definition 9 are all equivalent to each other, up to an absolute constant factor in front of the parameter δ.*

Definition 9 (Generalized stability; alternative definitions) *Let $\epsilon \in (0, 1/2)$ and $\delta \in [\epsilon, \infty)$. Let $S \subset \mathbb{R}^d$ and μ be a vector. Each of the following two bullets is equivalent*[13] *to Definition 7:*

- *S satisfies: (i) $\|\mu_S - \mu\|_2 \leq \delta$, (ii) for all $\mathbf{M} \in \mathcal{M}_k$, $\langle \mathbf{M}, \overline{\mathbf{\Sigma}}_S - \mathbf{I} \rangle \leq \delta^2/\epsilon$, and (iii) for all $S' \subset S$ with $|S'| \geq (1-\epsilon)|S|$ and for all $\mathbf{M} \in \mathcal{M}_k$, it holds $\langle \mathbf{M}, \overline{\mathbf{\Sigma}}_{S'} - \mathbf{I} \rangle \geq -\delta^2/\epsilon$.*
- *S satisfies (i),(ii) as above, and for all $T \subset S$ with $|T| \leq \epsilon|S|$ and all $\mathbf{M} \in \mathcal{M}_k$, $\frac{|T|}{|S|}\langle \mathbf{M}, \overline{\mathbf{\Sigma}}_T \rangle \leq \frac{\delta^2}{\epsilon}$.*

B.2. Consequences of (Generalized) Stability

The next result gives a bound on the average of $\|x - \mu\|_{\mathbf{M}}$ over a small subset of a stable set.

Lemma 19 *Let S be a finite multiset of n points satisfying the (ϵ, δ, k)-generalized-stability condition with respect to $\mu \in \mathbb{R}^d$. Then $\max_{\mathbf{M} \in \mathcal{M}_k} \max_{T \subset S : |T| \leq \epsilon |S|} \frac{1}{|S|} \sum_{x \in T} \|x - \mu\|_{\mathbf{M}} \lesssim \delta$.*

Proof Using Cauchy-Schwarz inequality and the last condition in Definition 9, we obtain

$$\begin{aligned}\max_{\mathbf{M} \in \mathcal{M}_k} \max_{\substack{T \subset S \\ |T| \leq \epsilon n}} \frac{1}{|S|} \sum_{x \in T} \|x\|_{\mathbf{M}} &\leq \max_{\mathbf{M} \in \mathcal{M}_k} \max_{\substack{T \subset S \\ |T| \leq \epsilon n}} \frac{|T|}{|S|} \sqrt{\frac{1}{|T|} \sum_{x \in T} \|x\|_{\mathbf{M}}^2} \\ &= \max_{\mathbf{M} \in \mathcal{M}_k} \max_{\substack{T \subset S \\ |T| \leq \epsilon n}} \sqrt{\frac{|T|}{|S|}} \sqrt{\frac{1}{|S|} \sum_{x \in T} \|x\|_{\mathbf{M}}^2} \lesssim \sqrt{\epsilon} \sqrt{\frac{\delta^2}{\epsilon}} \leq \delta.\end{aligned}$$

■

Lemma 10 *Let S be an (ϵ, δ)-stable set with respect to μ. Then S is also (ϵ, δ', k)-generalized stable with respect to μ with $\delta' \lesssim \delta\sqrt{k}$.*

Proof Since the mean condition in the definition of generalized stability is the same as the one in the plain stability, this condition holds trivially. For the covariance condition, let $\mathbf{M} = \sum_{i=1}^d \lambda_i v_i v_i^\top$ be the spectral decomposition of $\mathbf{M}$. Then, for a subset $S' \subseteq S$ with $|S'| \geq (1 - \epsilon)|S|$ we have that

$$\left| \left\langle \mathbf{M}, \frac{1}{|S'|} \sum_{x \in S'} x x^\top - \mathbf{I} \right\rangle \right| \leq \sum_{i=1}^d \lambda_i \left| \left\langle v_i v_i^\top, \frac{1}{|S'|} \sum_{x \in S'} x x^\top - \mathbf{I} \right\rangle \right|$$

[13] Up to a constant factor in the resulting stability parameter δ.

$$\leq \sum_{i=1}^{d} \lambda_i \delta^2/\epsilon = \mathrm{tr}(\mathbf{M})\delta^2/\epsilon = k\delta^2/\epsilon \ ,$$

where the second inequality uses the (ϵ, δ)-stability of S. ■

The next result shows that all large subsets of a stable set are stable, and the contamination parameter ϵ is "robust" to constant prefactors.

Lemma 11 *Let S be (ϵ, δ, k)-generalized stable with respect to μ. Let $r \geq 1$ be such that $r\epsilon \leq 1/2$. Then: (i) S is also $(r\epsilon, \delta', k)$-generalized stable with respect to μ for $\delta' \lesssim \delta\sqrt{r}$, (ii) any subset $S' \subset S$ such that $|S'| \geq (1 - r\epsilon)|S|$, is also (ϵ, δ', k)-generalized-stable with respect to μ with $\delta' \lesssim \delta\sqrt{r}$.*

Proof We start with the first claim, which we show using Definition 9 for the definition of generalized stability (and Lemma 18, stating that all definitions are equivalent to each other up to an absolute constant in front of the δ). The first two conditions in Definition 9 (about the mean and second moment over all the points in S) hold trivially by the (ϵ, δ, k)-generalized stability of S. It remains to show the last condition, that for every set $T \subseteq S$ with $|T| \leq r\epsilon|S|$ it holds $\frac{1}{|S|}\sum_{x\in T} \|x - \mu\|_{\mathbf{M}}^2 \leq \delta'^2/\epsilon$. This can be seen by splitting T into at most r disjoint sets of size at most $\epsilon|S|$ each, and apply the corresponding condition from the (ϵ, δ, k)-generalized stability of S. That is, write $T = T_1 \cup \cdots \cup T_{r'}$ where T_j are disjoint, $|T_j| \leq \epsilon|S|$ and $r' \leq r$. Then

$$\frac{1}{|S|}\sum_{x\in T} \|x - \mu\|_{\mathbf{M}}^2 = \sum_{j=1}^{r'} \frac{1}{|S|} \sum_{x\in T_j} \|x - \mu\|_{\mathbf{M}}^2 \leq r'\delta^2/\epsilon \ .$$

We move to the second claim. Using the first claim, we have that S is $((r+1)\epsilon, \delta', k)$-generalized stable with $\delta' \lesssim \sqrt{r}\delta$ (since $r \geq 1$). It remains to check that the two conditions from Definition 7 hold for every subset S'' of size $|S''| \geq (1 - \epsilon)|S'|$. Since $(1 - \epsilon)|S'| \geq (1 - \epsilon)(1 - r\epsilon)|S| \geq (1 - (r + 1)\epsilon))|S|$, the desired conditions follow by the $((r + 1)\epsilon, \delta', k)$-generalized stability of S. ■

Finally, the next result shows that all large subsets of a stable set are close in the sliced-Wasserstein metrics (Definition 5).

Lemma 12 *Let S_0 be a set satisfying (ϵ, δ, k)-generalized stability, and S_0' be a subset of S_0 with $|S_0'| \geq (1 - \epsilon)|S_0|$ for $\epsilon \leq 1/2$. Then, $W_{1,k}(S_0, S_0') \lesssim \epsilon\sqrt{k} + \delta$ and $W_{2,k}(S_0, S_0') \lesssim \sqrt{\epsilon k} + \delta/\sqrt{\epsilon}$.*

Proof Let us use the notation $S_0 = \{x_1, \ldots, x_n\}$, for our set satisfying (ϵ, δ, k)-generalized stability with respect to μ. Let us use $\mu = 0$ thought the proof without loss of generality. Define $\mathcal{J} \subset [n]$ for the set of indices corresponding to the points in S_0' (and $\mathcal{J}^{\complement} = [n] \setminus \mathcal{J}$ for the rest of the points), and denote $m := |\mathcal{J}| = |S_0'|$. Recall the definition of sliced-Wasserstein distance from Definition 5 for $p \in \{1, 2\}$:

$$W_{p,k}(S_0, S_0') = \sup_{\mathbf{V}\in\mathcal{V}_k} \inf_{\pi\in\Pi(S_0,S_0')} \mathop{\mathbf{E}}_{(x,x')\sim\pi} \left[\|\mathbf{V}(x - x')\|_2^p\right]^{1/p} \ .$$

We will use the following coupling π on (x, x'):

- First, $x' = x_i$ for an index i chosen uniformly at random from $\mathcal{J}$.

- Then, conditioned on $x' = x_i$, with probability m/n, x is set to be x_i and with probability $1 - m/n$, x is chosen to be x_i for an index chosen uniformly at random from $\mathcal{J}^{\complement}$.

It can be checked that this is a valid coupling: The marginal of x' is the uniform distribution on S_0' (by definition), and the marginal of x is uniform on S_0 since for every $i \in \mathcal{J}$ we have $\mathbb{P}[x = x_i] = \frac{1}{m}\frac{m}{n} = 1/n$ and for every $i \in \mathcal{J}^{\complement}$ we have $\mathbb{P}[x = x_i] = \sum_{i\in[m]} \frac{1}{m}(1 - m/n)\frac{1}{n-m} = 1/n$. We can thus bound $W_{p,k}(S_0, S_0')$ as follows:

$$\begin{aligned} W_{p,k'}(S_0, S_0')^p &\le \sup_{\mathbf{V}\in\mathcal{V}_k} \mathop{\mathbf{E}}_{(x,x')\sim\pi}\left[\|\mathbf{V}(x - x')\|_2^p\right] \\ &\lesssim \sup_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{m}\sum_{i\in\mathcal{J}} \mathop{\mathbf{E}}_{(x,x')\sim\pi}\left[\|\mathbf{V}(x - x')\|_2^p \,|\, x' = x_i\right] \\ &= \sup_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{m}\sum_{i\in\mathcal{J}}\left(\frac{m}{n}\|\mathbf{V}(x_i - x_i)\|_2^p + \frac{n-m}{n}\frac{1}{n-m}\sum_{j\in\mathcal{J}^{\complement}}\|\mathbf{V}(x_j - x_i)\|_2^p\right) \\ &= \sup_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{mn}\sum_{i\in\mathcal{J}, j\in\mathcal{J}^{\complement}}\|\mathbf{V}(x_j - x_i)\|_2^p \,. \end{aligned} \tag{10}$$

Controlling $W_{1,k}$ We first consider the easy case of $p = 1$, for which we can use the triangle inequality to obtain the following:

$$\begin{aligned} \frac{1}{mn}\sum_{i\in\mathcal{J}, j\in\mathcal{J}^{\complement}}\|\mathbf{V}(x_j - x_i)\|_2 &\le \frac{1}{mn}\sum_{i\in\mathcal{J}, j\in\mathcal{J}^{\complement}}\left(\|\mathbf{V}x_j\|_2 + \|\mathbf{V}x_i\|_2\right) \\ &\lesssim \frac{1}{n}\sum_{j\in\mathcal{J}^{\complement}}\|\mathbf{V}x_j\|_2 + \epsilon\cdot\frac{1}{m}\sum_{i\in\mathcal{J}}\|\mathbf{V}x_i\|_2 \\ &\lesssim \delta + \epsilon\cdot\frac{1}{n}\sum_{i\in[n]}\|\mathbf{V}x_i\|_2 \,, \end{aligned} \tag{11}$$

where the bound on the first term follows by Lemma 19, and the bound on the second term uses that $n \lesssim m$. We now use the (ϵ, δ, k)-generalized-stability of S_0 and Cauchy-Schwarz inequality to obtain the following:

$$\frac{1}{n}\sum_{i\in[n]}\|\mathbf{V}x_i\|_2 \le \sqrt{\frac{1}{n}\sum_{i\in[n]}\|\mathbf{V}x_i\|_2^2} = \sqrt{\langle\mathbf{V}, \mathbf{\Sigma}_{S_0}\rangle} \lesssim \sqrt{k + \delta^2/\epsilon} \lesssim \sqrt{k} + \delta/\sqrt{\epsilon}\,.$$

This concludes an upper bound on $W_{1,k}(S_0, S_0')$ of the order $\delta + \epsilon\sqrt{k} + \delta\sqrt{\epsilon}$.

Controlling $W_{2,k}$ We now turn our attention to $W_{2,k}(S_0, S_0')$ using $p = 2$ in Equation (10). Using the inequality $(a + b)^2 \le 2a^2 + 2b^2$ to analyze the cross term, we obtain:

$$\frac{1}{nm}\sum_{i\in\mathcal{J}}\sum_{j\in\mathcal{J}^{\complement}}\|\mathbf{V}(x_j - x_i)\|_2^2 \lesssim \frac{\epsilon}{m}\sum_{i\in\mathcal{J}}\|\mathbf{V}x_i\|_2^2 + \frac{1}{n}\sum_{j\in\mathcal{J}^{\complement}}\|\mathbf{V}x_j\|_2^2$$

$$\lesssim \epsilon\langle \mathbf{V}, \mathbf{\Sigma}_{S_0}\rangle + \frac{\delta^2}{\epsilon} \lesssim \epsilon\left(k + \frac{\delta^2}{\epsilon}\right) + \frac{\delta^2}{\epsilon} \lesssim \epsilon k + \frac{\delta^2}{\epsilon}.$$

where the bounds in the last line follow by the generalized stability of the original points S_0. Thus, we conclude that $W_{2,k}(S_0, S')^2 \lesssim \epsilon k + \delta^2/\epsilon$.

■

Appendix C. Stability Rates for Well-Behaved Distributions

The statistical rates for stability (Definition 1) were well understood due to Diakonikolas et al. (2020), which established bounds that are optimal up to a $\sqrt{\log d}$ factor for distributions with bounded covariance or higher moments. In this paper, we further tighten these bounds by eliminating the logarithmic term entirely. This section presents the improved bounds, while the proof (which strengthens the approach of Diakonikolas et al. (2020) using recent concentration results for heavy-tailed random matrices) is deferred to Appendix C.1.

Theorem 20 (Stability rates for well-behaved distribution families (Diakonikolas et al., 2020)) *Let $\mathcal{D}$ be a family of distributions. Let c be a small enough absolute constant. Fix a $D \in \mathcal{D}$ and let S_0 be a set of n samples drawn i.i.d. from D. For each of the cases below and $\epsilon + \frac{\log(1/\tau)}{n} \in (0, c)$, there exists an $S_1 \subset S_0$ with $|S_1| \geq (1-\epsilon)|S_0|$ which is (ϵ, δ)-stable with probability at least $1-\tau$, for the following parameter δ:*

- *If $\mathcal{D}$ is the family of isotropic subgaussian distributions, then $\delta \lesssim \epsilon\sqrt{\log(1/\epsilon)} + \sqrt{d/n} + \sqrt{\log(1/\tau)/n}$.*
- *If $\mathcal{D}$ is the family of distributions with isotropic covariance and bounded k-th moments,[14] then $\delta \lesssim \epsilon^{1-\frac{1}{k}} + \rho + \sqrt{d/n} + \sqrt{\log(1/\tau)/n}$.*
- *If $\mathcal{D}$ is the family of distributions with covariance $\mathbf{\Sigma} \preceq \mathbf{I}$, then $\delta \lesssim \sqrt{\epsilon} + \sqrt{\mathrm{tr}(\mathbf{\Sigma})/n} + \sqrt{\log(1/\tau)/n}$.*

In comparison, the stability rates given in Diakonikolas et al. (2020) are the same as Theorem 20 for isotropic subgaussian distributions, $\delta \lesssim \epsilon^{1-\frac{1}{k}} + \rho + \sqrt{(d\log d)/n} + \sqrt{\log(1/\tau)/n}$ for the family of distributions with isotropic covariance and bounded k-th moments, and $\delta \lesssim \sqrt{\epsilon} + \sqrt{(\mathrm{tr}(\mathbf{\Sigma})\log(r(\mathbf{\Sigma}))/n} + \sqrt{\log(1/\tau)/n}$ for the family of distributions with covariance $\mathbf{\Sigma} \preceq \mathbf{I}$ (where $r(\mathbf{\Sigma})$ denotes the rank of the matrix $\mathbf{\Sigma}$).

Recall the family of *stability-based algorithms* from Definition 2. An immediate corollary concerns robust mean estimation of the mean of well-behaved distributions with high probability, with optimal bounds on the error rate.

Corollary 21 *Let $\mathcal{D}$ be a family of distributions. Fix a $P \in \mathcal{P}$ and let S_0 be a set of n i.i.d. samples from P. Let T be a corrupted version of S_0 with ϵ-fraction of global outliers (Global Contamination Model). Let μ be the (unknown) mean of P. There exist computationally-efficient algorithms that take as input T, ϵ, and ρ and output $\widehat{\mu}$ with the following guarantees:*

[14] That is, $\mathbf{E}_{x\sim P}[|v^\top(X-\mu_P)|^k]^{1/k} \leq \sigma_k$ for constant σ_k.

- *If $\mathcal{D}$ is the family of isotropic subgaussian distributions, then $\|\widehat{\mu}-\mu\|_2 \lesssim \epsilon\sqrt{\log(1/\epsilon)} + \sqrt{d/n} + \sqrt{\log(1/\tau)/n}$.*
- *If $\mathcal{D}$ is the family of distributions with isotropic covariance and bounded k-th moments, then $\|\widehat{\mu}-\mu\|_2 \lesssim \epsilon^{1-\frac{1}{k}} + \sqrt{d/n} + \sqrt{\log(1/\tau)/n}$.*
- *If $\mathcal{D}$ is the family of distributions with covariance $\mathbf{\Sigma} \preceq \mathbf{I}$ then $\|\widehat{\mu}-\mu\|_2 \lesssim \sqrt{\epsilon} + \sqrt{d/n} + \sqrt{\log(1/\tau)/n}$.*

The above bounds for robust mean estimation are optimal with respect to the parameters ϵ, d, n, τ. Specifically for ϵ, the matching lower bound can be derived using a simple identifiability argument (see Section 1.3 in Diakonikolas and Kane (2023) for formal details). The (near-)optimality with respect to the remaining parameters are also folklore facts: $\Omega(\sqrt{d/n})$ error is necessary even for estimating the mean of $\mathcal{N}(\mu, \mathbf{I})$ without outliers (as a standard application of Fano's method), and $\Omega(\sqrt{\log(1/\tau)/n})$ is necessary even for one-dimensional $\mathcal{N}(\mu, 1)$ without outliers (see e.g., Proposition 6.1 in Catoni (2012)).

C.1. Proof of Improved Stability Rates

The two subsections that follow discuss the proof of the last two bullets in Theorem 20.

C.1.1. Bounded Covariance

We show how to remove the $\log d$ factor from Theorem 1.4 in Diakonikolas et al. (2020), by leveraging recent concentration inequalities for heavy-tailed matrices (Jirak et al., 2024).

Theorem 22 *Fix $\tau \in (0,1)$. Let $x_1, \dots, x_n$ be i.i.d. points from a distribution in $\mathbb{R}^d$ with mean μ and covariance $\mathbf{\Sigma} \preceq \mathbf{I}$. Let $\epsilon' = O(\log(1/\tau)/n + \epsilon) \leq c$ for a sufficiently small positive constant c. Then, with probability at least $1-\tau$, there exists a subset $S' \subseteq S$ s.t. $|S'| \geq (1-\epsilon')n$ and S' $(2\epsilon', \delta)$-stable (Definition 1) with respect to μ with $\delta = O(\sqrt{\mathrm{tr}(\mathbf{\Sigma})/n} + \sqrt{\epsilon'} + \sqrt{\log(1/\tau)/n})$.*

We follow the proof structure of Diakonikolas et al. (2020), with our improvement coming from an alternate proof of Lemmata 2.3 and 2.4 in that paper. Since the reminder of the proof is lengthy and does not need to be changed we will present only the revised versions of these lemmas.

First, we need the following helper lemma.

Lemma 23 *Let C be a sufficiently large absolute constant. Then for any $x \in \mathbb{R}^d$, $R \in \mathbb{R}$, and PSD $d \times d$ matrix $\mathbf{M}$ with $\mathrm{tr}(\mathbf{M}) \leq 1$, it holds*

$$\min(x^\top \mathbf{M} x, R) \lesssim \mathop{\mathbf{E}}_{v \sim \mathcal{N}(0,\mathbf{M})}\left[\min\left((v^\top x)^2, R\right) \mathbb{1}_{\|v\|_2 \leq C}\right] . \tag{12}$$

Proof The first goal is to show

$$\min(x^\top \mathbf{M} x, R) \lesssim \mathop{\mathbf{E}}_{v \sim \mathcal{N}(0,\mathbf{M})}\left[\min\left((v^\top x)^2, R\right)\right] . \tag{13}$$

To do that, we can write $\min(x^\top \mathbf{M} x, R) = \min\left(\mathbf{E}_{v \sim \mathcal{N}(0,\mathbf{M})}[(v^\top x)^2], R\right)$. Define $A := v^\top x$, which since $v \sim \mathcal{N}(0, \mathbf{M})$, we have that $A \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = x^\top \mathbf{M} x$. Therefore the goal expressed in a slightly simpler form is is to show that $\min(\sigma^2, R) \lesssim \mathbf{E}[\min(A^2, R)]$ for $A \sim \mathcal{N}(0, \sigma^2)$. We examine two cases.

If $R \leq \sigma^2$, then the LHS is equal to R and it thus remains to show that the RHS is at least R. This can be seen as follows

$$\mathbf{E}[\min(A^2, R)] \geq R \cdot \mathbf{Pr}[R \leq A^2] \geq R \cdot \mathbf{Pr}[\sigma^2 \leq A^2] = \Omega(R) \ , \tag{14}$$

where the first step follows because $R, A^2 \geq 0$, the second step uses that $R \leq \sigma^2$ and the last step uses that $\mathbf{Pr}_{x \sim \mathcal{N}(0,1)}[x^2 \geq 1] = \Omega(1)$.

In the opposite case that $R > \sigma^2$, the target inequality $\min(\sigma^2, R) \lesssim \mathbf{E}[\min(A^2, R)]$ has LHS equal to σ^2, and thus, we need to show that $\mathbf{E}[\min(A^2, R)] \gtrsim \sigma^2$. This is indeed true because

$$\mathbf{E}[\min(A^2, R)] \geq \mathbf{E}[\min(A^2, \sigma^2)] = \sigma^2\, \mathbf{E}[\min(A^2/\sigma^2, 1)] = \Omega(\sigma^2) \ ,$$

where the first step uses that $R > \sigma^2$ and the last step uses $\mathbf{E}_{x \sim \mathcal{N}(0,1)}[\min(x^2, 1)] = \Omega(1)$.

So far we have shown Equation (13). This is indeed sufficient to conclude the proof of Lemma 23 because of the following:

$$\begin{aligned}
\mathop{\mathbf{E}}_{v \sim \mathcal{N}(0,\mathbf{M})}\left[\min\left(A^2, R\right) \mathbb{1}_{\|v\|_2 > C}\right] &\leq \sqrt{\mathop{\mathbf{E}}_{v \sim \mathcal{N}(0,\mathbf{M})}[\min\left(A^2, R\right)]}\sqrt{\mathbf{Pr}[\|v\|_2 > C]} \\
&\leq \sqrt{\mathop{\mathbf{E}}_{v \sim \mathcal{N}(0,\mathbf{M})}[\min\left(A^4, R^2\right)]}/C' \qquad \text{(since } \mathrm{tr}(\mathbf{M}) \leq 1) \\
&\lesssim \min(x^\top \mathbf{M} x, R)/C' \ ,
\end{aligned}$$

where the last step is the claim that remains to be shown. This is equivalent to $\mathbf{E}\left[\min\left(A^4, R^2\right)\right] \leq \min(\sigma^4, R^2)$. By scaling, let us consider $\sigma^2 = 1$ and $A \sim \mathcal{N}(0,1)$. The claim then can be checked by taking cases: If $1 < R$ we have $\mathbf{E}\left[\min\left(A^4, R^2\right)\right] \leq \mathbf{E}[A^4] = O(1) = O(\min(1, R^2))$. If $1 \geq R$, then $\mathbf{E}[\min(A^4, R^2)] \leq R^2 = \min(1, R^2)$. ■

We now show the revised version of Lemma 2.3 and 2.4 in Diakonikolas et al. (2020) (we do this in the single lemma statement below that combines the two). Regarding notation, $\Delta_{n,\epsilon} := \{w \in \mathbb{R}^n_+ : \sum_{i=1}^n w_i = 1\,; 0 \leq w_i \leq 1/(n(1-\epsilon))\ \}$ and $\overline{\mathbf{\Sigma}}_w = w_i(x_i - \mu)(x_i - \mu)^\top$.

Lemma 24 *Let $x_1, \ldots, x_n$ be i.i.d. points from a distribution in $\mathbb{R}^d$ with mean μ and covariance $\mathbf{\Sigma} \preceq \mathbf{I}$. Further assume that for each i, $\|x_i - \mu\|_2 = O(\sqrt{\mathrm{tr}(\mathbf{\Sigma})/\epsilon})$. There exists $c, c' > 0$ such that for $\epsilon \in (0, c')$, with probability $1 - 2\exp(-cn\epsilon)$, we have that $\min_{w \in \Delta_{n,\epsilon}} \|\overline{\mathbf{\Sigma}}_w - \mathbf{I}\|_{\mathrm{op}} \leq \delta^2/\epsilon$ for some $\delta = O(\sqrt{\mathrm{tr}(\mathbf{\Sigma})/n} + \sqrt{\epsilon})$.*

Proof The first part is the same as Diakonikolas et al. (2020). We consider $\mu = 0$ without loss of generality. Let $Q = \Theta(1/\sqrt{\epsilon} + (1/\epsilon)\sqrt{\mathrm{tr}(\mathbf{\Sigma})/n})$.

$$\begin{aligned}
\min_{w \in \Delta_{w,\epsilon}} \|\overline{\mathbf{\Sigma}}_w - \mathbf{I}\|_{\mathrm{op}} &\leq 1 + \min_{w \in \Delta_{w,\epsilon}} \|\overline{\mathbf{\Sigma}}_w\|_{\mathrm{op}} \\
&= 1 + \min_{w \in \Delta_{w,\epsilon}} \max_{\mathbf{M} \in \mathcal{M}} \sum_{i=1}^n w_i x_i^\top \mathbf{M} x_i \\
&= 1 + \max_{\mathbf{M} \in \mathcal{M}} \min_{w \in \Delta_{w,\epsilon}} \sum_{i=1}^n w_i x_i^\top \mathbf{M} x_i
\end{aligned}$$

$$\leq 1 + \frac{1}{(1-\epsilon)n} \max_{\mathbf{M} \in \mathcal{M}} \sum_{i=1}^{n} \min(x_i^\top \mathbf{M} x_i, Q^2) \ ,$$

where the last step uses that with probability $1 - \exp(-cn\epsilon)$, for every $\mathbf{M} \in \mathcal{M}$ the set $S_{\mathbf{M}} = \{i \in [n] : x_i^\top \mathbf{M} x_i \leq Q^2\}$ has cardinatliy larger than $(1-\epsilon)n$ (established in another lemma in Diakonikolas et al. (2020)), meaning that the uniform distribution over $S_{\mathbf{M}}$ belongs to $\Delta_{n,\epsilon}$. The proof now will deviate from Diakonikolas et al. (2020). Using Lemma 23 we can further bound the RHS as follows:

$$\max_{\mathbf{M} \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \min(x_i^\top \mathbf{M} x_i, Q^2) \lesssim \max_{\mathbf{M} \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbf{E}}_{v \sim \mathcal{N}(0,\mathbf{M})} [\min((v^\top x_i)^2, Q^2) \mathbb{1}_{\|v\|_2 \leq C}]$$
$$\lesssim \sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \frac{1}{n} \sum_{i=1}^{n} \min((v^\top x_i)^2, Q^2) \ .$$

The goal is to show that the RHS is at most $O(\delta^2/\epsilon)$. Let $g_v(x) = \min((v^\top x)^2, Q^2)$ to save space. We can write the following (where the expectations are with respect to x_i):

$$\sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \frac{1}{n} \sum_{i=1}^{n} g_v(x_i) \leq \sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \frac{1}{n} \sum_{i=1}^{n} (g_v(x_i) - \mathbf{E}[g_v(x_i)]) + \sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[g_v(x_i)] \ .$$

We will bound each term separately. For the second term we have

$$\sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[g_v(x_i)] \leq \sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[(v^\top x_i)^2] = O(1) \ ,$$

which follows by the assumption that x_i have covariance matrix $\mathbf{\Sigma} \preceq \mathbf{I}$ and $\|v\|_2 = O(1)$.

We now move to the deviation term, which we call Y to save space. We first bound the expectation of this term using symmetrization and contraction property for Rademacher averages (Giné and Zinn, 1984; Ledoux and Talagrand, 1991; Boucheron et al., 2013):

$$\mathbf{E}[Y] = \mathbf{E}\left[\sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \frac{1}{n} \sum_{i=1}^{n} (g_v(x_i) - \mathbf{E}[g_v(x_i)])\right] \leq 2\,\mathbf{E}\left[\sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \frac{1}{n} \left|\sum_{i=1}^{n} \xi_i g_v(x_i)\right|\right]$$
$$\lesssim Q\,\mathbf{E}\left[\sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \frac{1}{n} \left|\sum_{i=1}^{n} \xi_i v^\top x_i\right|\right] = Q\,\mathbf{E}\left[\left\|\frac{1}{n} \sum_{i=1}^{n} x_i\right\|_{\mathrm{op}}\right] \leq Q\sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})}{n}} \ .$$

Then, we apply Talagrand's concentration inequality to obtain that, with probability at least $1 - \exp(-n\epsilon)$, the following holds,

$$Y \lesssim \mathbf{E}[Y] + \frac{1}{n}\sqrt{\sigma^2 t} + \frac{1}{n} L t \ ,$$

where $t := \epsilon n$, $\sigma^2 := \sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \sum_{i=1}^{n} (g_v(x_i) - \mathbf{E}[g_v(x_i)])^2 \leq \sup_{v \in \mathbb{R}^d : \|v\|_2 \leq C} \sum_{i=1}^{n} g_v(x_i)^2 \lesssim Q^2 \sup_{v \in \mathbb{R}^d : \|v\|_2 \lesssim C} \sum_{i=1}^{n} \mathbf{E}[g_v(x_i)] \lesssim Q^2 n$ and $L := Q^2$. Plugging these value above we obtain:

$$Y \lesssim \mathbf{E}[Y] + \frac{1}{n} Q \sqrt{n} \sqrt{\epsilon n} + \frac{1}{n} Q^2 \epsilon n$$

$$
\begin{aligned}
&\lesssim Q\sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})}{n}} + Q\sqrt{\epsilon} + Q^2\epsilon \\
&= \frac{1}{\epsilon}\left(Q\epsilon\sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})}{n}} + Q\epsilon\sqrt{\epsilon} + (Q\epsilon)^2\right) \\
&\lesssim \frac{1}{\epsilon}\left(\frac{\mathrm{tr}(\mathbf{\Sigma})}{n} + \sqrt{\epsilon\frac{\mathrm{tr}(\mathbf{\Sigma})}{n}} + Q\epsilon\sqrt{\epsilon} + (Q\epsilon)^2\right) \qquad (Q = \Theta(1/\sqrt{\epsilon} + (1/\epsilon)\sqrt{\mathrm{tr}(\mathbf{\Sigma})/n})) \\
&\lesssim \frac{1}{\epsilon}\left(\sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma})}{n}} + \sqrt{\epsilon} + \epsilon Q\right)^2 .
\end{aligned}
$$

Therefore, setting $\delta = \Theta(\sqrt{\mathrm{tr}(\mathbf{\Sigma})/n} + \sqrt{\epsilon} + \epsilon Q) = \Theta(\sqrt{\mathrm{tr}(\mathbf{\Sigma})/n} + \sqrt{\epsilon})$ the right hand side above is at most δ^2/ϵ. ■

C.1.2. Bounded Higher Moments

Unlike the previous section, we now consider a variant of the stability definition (Definition 1) that uses weights for each point (i.e., soft sets). Given a weight function $w : \mathbb{R}^d \to [0,1]$, and a distribution D we define the re-weighted distribution D_w to be $D_w(x) := D(x)w(x)/\int_{\mathbb{R}^d} w(x)D(x)\mathrm{d}x$. We use $\mu_{w,D} = \mathbf{E}_{X\sim D_w}[X]$ for its mean and $\overline{\mathbf{\Sigma}}_{w,D} = \mathbf{E}_{X\sim D_w}[(X-\mu)(X-\mu)^T]$ for the second moment that is centered with respect to μ (it will be clear from the context what μ is). The weighted-variant of the stability definition is the following:

Definition 25 (Stability condition; weighted version) *Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. A distribution G on $\mathbb{R}^d$ is (ϵ, δ)-stable with respect to $\mu \in \mathbb{R}^d$ if for any weight function $w : \mathbb{R}^d \to [0,1]$ with $\mathbf{E}_{X\sim G}[w(X)] \geq 1 - \epsilon$ we have that $\|\mu_{w,G} - \mu\|_2 \leq \delta$ and $\|\overline{\mathbf{\Sigma}}_{w,G} - \mathbf{I}_d\|_{\mathrm{op}} \leq \delta^2/\epsilon$.*

We call a set of points (ϵ, δ)-stable when the uniform distribution on S is stable. This means that every large *soft* subset of S has bounded (weighted) mean and covariance as shown above.

We note that using weights instead of subsets is not a substantial difference since many of the stability-based algorithms mentioned in Appendix A work under this definition. The result of this section improves the stability rates from Diakonikolas et al. (2020) by leveraging recent concentration results for heavy-tailed random matrices from Jirak et al. (2024):

Theorem 26 *Fix $\tau \in (0,1)$. Let S be a multiset of n i.i.d. samples from a distribution P on $\mathbb{R}^d$ with mean μ, covariance $\mathbf{\Sigma} = \mathbf{I}$, and bounded central moment $\sigma_k := \sup_{v\in\S^{d-1}} \mathbf{E}_{x\sim P}[|v^\top(x - \mu_P)|^k]^{1/k} < \infty$ for some $k > 4$. Let $\epsilon' = \Theta(\log(1/\tau)/n + \epsilon)$ and assume that $\epsilon < c$ and $n > C_1 \log(1/\tau) + C_2(\sigma_k, k)d$, where c is a sufficiently small constant, C_1 is a sufficiently large constant, and $C_2(\sigma_k, k)$ is some quantity that depends only on σ_k and k. Then, with probability at least $1 - \tau$, there exist positive weights $\{w_i\}_{i\in[n]}$ with $\sum_{i\in[n]} w_i \geq 1 - \epsilon$ such that the uniform distribution on S, re-weighted by the w_i's is $(2\epsilon', \delta)$-stable with respect to μ, where $\delta = O(\sqrt{d/n} + \sigma_k\epsilon^{1-1/k} + \sigma_4\sqrt{\log(1/\tau)/n})$.*

We highlight, that the above result holds for $k > 4$. Handling $k \in (2,4]$ was an open problem in the matrix concentration literature (Tikhomirov, 2018; Abdalla and Zhivotovskiy, 2024), and was only recently resolved in Oliveira and Rico (2024). The statement also assumes $n > C_1 \log(1/\tau) + C_2(\sigma_k, k)d$, and the dependence on $\log(1/\tau)$ in δ scales with σ_4. If k, σ_4, σ_k are constants, then so are C_1, C_2. For general k, the best covariance estimation results (Abdalla and Zhivotovskiy, 2024) include similar dependencies on $C_2(\sigma_k, k)$ and σ_4, thus it is unclear whether removing them is easy.

Again, the proof follows by refining an argument from Diakonikolas et al. (2020). Since the proof is lengthy and the improvement stems from modifying one lemma, we do not repeat the full proof but instead highlight where the improvement occurs.

The key modification is in Lemma E.3 of Diakonikolas et al. (2020), specifically the bound on the expected value of the empirical process $\sup_{\mathbf{M}\in\mathcal{M}} f(x_i^\top \mathbf{M} x_i) - \mathbf{E}[f(x_i^\top \mathbf{M} x_i)]$ considered there. $\mathcal{M}$ and f have the same meaning as in Diakonikolas et al. (2020): $\mathcal{M}$ corresponds to $\mathcal{M}_1$ in our notation (the set $\{\mathbf{M} \in \mathbb{R}^{d\times d} : \mathbf{M} \succeq 0,\ \mathbf{M} \preceq \mathbf{I},\ \operatorname{tr}(\mathbf{M}) = 1\}$), and f is defined as $f(x) = \min(x, Q_k^2)$ for some threshold Q_k (irrelevant to our proof). The initial steps, like in Diakonikolas et al. (2020), rely on symmetrization and contraction principles (Giné and Zinn, 1984; Ledoux and Talagrand, 1991; Boucheron et al., 2013), after which the proof deviates to make use of the improved concentration for covariance matrices. The improved bounds are as follows

$$
\begin{aligned}
\mathbf{E}\left[\sup_{\mathbf{M}\in\mathcal{M}} f(x_i^\top \mathbf{M} x_i) - \mathbf{E}[f(x_i^\top \mathbf{M} x_i)]\right] &\le 2\,\mathbf{E}\left[\sup_{\mathbf{M}\in\mathcal{M}} \sum_{i=1}^n \epsilon_i f(x_i^\top \mathbf{M} x_i)\right] && \text{(symmetrization)}\\
&\le 2\,\mathbf{E}\left[\sup_{\mathbf{M}\in\mathcal{M}} \sum_{i=1}^n \epsilon_i x_i^\top \mathbf{M} x_i\right] && \text{(contraction)}\\
&= 2\,\mathbf{E}\left[\sup_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n \epsilon_i (v^\top x_i)^2\right] && \text{(convexity; $\mathcal{M}$ is the convex hull of $\{vv^\top : v \in \mathcal{S}^{d-1}\}$)}\\
&= 2\,\mathbf{E}\left[\sup_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n \epsilon_i \left((v^\top x_i)^2 - 1 + 1\right)\right]\\
&\le 2\,\mathbf{E}\left[\sup_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n \epsilon_i \left((v^\top x_i)^2 - 1\right)\right] + 2\,\mathbf{E}\left[\sum_{i=1}^n \epsilon_i\right]\\
&\le 2\,\mathbf{E}\left[\sup_{v\in\mathcal{S}^{d-1}} \sum_{i=1}^n \epsilon_i \left((v^\top x_i)^2 - 1\right)\right] && \text{(since $\mathbf{E}[\epsilon_i] = 0$)}\\
&\le 4\,\mathbf{E}\left[\sup_{v\in\mathcal{S}^{d-1}} \left|\sum_{i=1}^n \left((v^\top x_i)^2 - 1\right)\right|\right] && \text{(symmetrization (Boucheron et al., 2013))}\\
&= 4\,\mathbf{E}\left[\left\|\sum_{i=1}^n x_i x_i^\top - nI\right\|_{\mathrm{op}}\right]\\
&\le 4\,\mathbf{E}\left[\left\|\sum_{i=1}^n x_i x_i^\top - nI\right\|_{\mathrm{op}}^2\right]^{1/2} && \text{(Jensen's inequality)}\\
&\lesssim \left(\sqrt{nd} + \left(\mathbf{E}[\max_i \|x_i\|_2^k]\right)^{2/k}\right) && \text{(Theorem 3.3 in Jirak et al. (2024))}
\end{aligned}
$$

$$\lesssim C(\sigma_k, k)\left(\sqrt{nd} + \sigma_k^2 d/\epsilon^{-2/k}\right) \ . \qquad \text{(assumption that } \max_{i\in[n]}\|x_i\|_2 = O(\sigma_k\sqrt{d}\epsilon^{-1/k})\text{)}$$

The application of Theorem 3.1 in Jirak et al. (2024) requires us to make the assumption $n > C_2(\sigma_k, k)d$, where $C_2(\sigma_k, k)$ is a constant that depends only on k and σ_k. This assumption is not included in the setting of Diakonikolas et al. (2020), therefore, the improved result Theorem 26 will need to include it. Regarding the last line above, the assumption that $\max_{i\in[n]}\|x_i\|_2 = O(\sigma_k\sqrt{d}\epsilon^{-1/k})$ comes from the setting of Lemma E.3 in Diakonikolas et al. (2020), i.e., it is not an additional assumption. Finally, the RHS above, can be further bounded by $O\left(\sqrt{\frac{\sigma_k^2 nd}{\epsilon^{2/k}}} + \frac{\sigma_k^2 d}{\epsilon^{2/k}}\right)$, using the fact that $\sigma_k^2 \geq 1$ and $\epsilon^{-2/k} \geq 1$, and also using that $n \gg C_2(\sigma_k, k)d$. Comparing this with the bound in Diakonikolas et al. (2020), we see that it is identical but without the $\log d$ factor. Thus, the rest of the proof strategy from Diakonikolas et al. (2020) can be carried out as in that paper but, with two differences: we add the assumption $n > C_2(\sigma_k, k)d$, and we avoid the randomized rounding lemma from Diakonikolas et al. (2020), which converts weights w into a hard set because it introduces a $\log d$ factor (this is the reason why we use Definition 25 in this section). With these adjustments, Theorem 26 follows.

Appendix D. Omitted Details from Section 3

We restate and proof the following technical result, which follows from a Gaussian rounding scheme, inspired by Depersin and Lecué (2022).

Proposition 13 (Bound on average projections) *Let $y_1, y_2, \ldots, y_n$ be vectors in $\mathbb{R}^d$. Then the following holds:* $\sup_{\mathbf{M}\in\mathcal{M}_k}\frac{1}{n}\sum_{i=1}^n\|y_i\|_{\mathbf{M}} \lesssim \sup_{\mathbf{V}\in\mathcal{V}_k}\frac{1}{n}\sum_{i=1}^n\|y_i\|_{\mathbf{V}}$.

Proof Let $\rho := \sup_{\mathbf{V}\in\mathcal{V}_k}\frac{1}{n}\sum_{i=1}^n\|\mathbf{V}y_i\|_2$. Suppose that there exists a $\mathbf{M}\in\mathcal{M}_k$ with the property $\frac{1}{n}\sum_{i=1}^n\sqrt{y_i^\top\mathbf{M}y_i} \geq C'\cdot\rho$ for a sufficiently large constant C'. We will show that this leads to a contradiction.

For a $r > 0$, let $\mathcal{B}_r := \{\mathbf{B}\in\mathbb{R}^{d\times d}\succeq 0 : \|\mathbf{B}\|_{\mathrm{op}} \leq r,\ \mathbf{B}\text{ is rank-}k\}$ be the set of rank-k PSD matrices with bounded operator norm. Let $g_1, \ldots, g_k$ be i.i.d. samples from $\mathcal{N}(0, \mathbf{M})$ and define $\mathbf{B} := \frac{1}{k}\sum_{i=1}^k g_i g_i^\top$ to be the empirical second moment matrix, which is an unbiased estimate of $\mathbf{M}$. We define the random variable

$$Z = \frac{1}{n}\sum_{i=1}^n\|y_i\|_{\mathbf{B}}\mathbb{1}_{\mathbf{B}\in\mathcal{B}_r} \ .$$

On the one hand, we have that

$$Z \leq \frac{1}{n}\sum_{i=1}^n\|\mathbf{B}^{1/2}\|_{\mathrm{op}}\|y_i\|_{\mathbb{P}_{\mathbf{B}}}\mathbb{1}_{\mathbf{B}\in\mathcal{B}_r} \leq \sqrt{r}\sup_{\mathbf{V}\in\mathcal{V}_k}\frac{1}{n}\sum_{i=1}^n\|y_i\|_{\mathbf{V}} \leq \sqrt{r}\rho, \tag{15}$$

where we use that $\mathbf{B}$ is a rank-k matrix if $\mathbf{B}\in\mathcal{B}_r$. The above implies that $\mathbf{E}[Z] \leq \sqrt{r}\rho$. We shall now show contradiction by deriving a lower bound on $\mathbf{E}[Z]$.

Towards that goal, we use the following decomposition to handle the indicator event $\mathbf{B}\in\mathcal{B}_r$:

$$\mathbf{E}[Z] = \frac{1}{n}\sum_{i=1}^n\mathbf{E}\left[\|y_i\|_{\mathbf{B}}\right] - \frac{1}{n}\sum_{i=1}^n\mathbf{E}\left[\|y_i\|_{\mathbf{B}}\mathbb{1}_{\mathbf{B}\notin\mathcal{B}_r}\right] \ , \tag{16}$$

where the expectation is taken with respect to the random matrix $\mathbf{B}$.

We first obtain a lower bound on the first term above. Consider the random variables $R_i := \|y_i\|_{\mathbf{B}}^2 = \frac{1}{k}\sum_{j=1}^k (g_j^\top y_i)^2$, which is a degree-two polynomial of the Gaussian samples. Then $\mathbf{E}[R_i] = \|y_i\|_{\mathbf{M}}^2$. To obtain a lower bound on $\mathbf{E}[\sqrt{R_i}]$, we shall prove an upper bound on $\mathbf{E}[R_i^2]$. Using $\mathbf{E}[G^4] \leq 3(\mathbf{E}[G^2])^2$ for a Gaussian variable G, we obtain $\mathbf{E}[R_i^2] = \frac{1}{k^2}\sum_{j=1}^k \mathbf{E}[(g_j^\top y_i)^4] + \frac{k(k-1)}{k^2}(\mathbf{E}[R_i])^2 < \frac{k+3}{k}(\mathbf{E}[R_i])^2 \leq 4(\mathbf{E}[R_i])^2$. We shall use this upper bound in the following inequality: $\mathbf{E}[|X|]^{2/3}\,\mathbf{E}[|X|^4]^{1/3} \geq \mathbf{E}[|X|^2]$ which holds for any real-valued random variable X with finite fourth moments. Applying it to $\sqrt{R_i}$, we obtain $\mathbf{E}[\sqrt{R_i}] \geq \frac{\mathbf{E}[R_i]^{3/2}}{\mathbf{E}[R_i^2]^{1/2}} \geq \frac{\mathbf{E}[R_i]^{3/2}}{2\,\mathbf{E}[R_i]} = \frac{1}{2}\sqrt{\mathbf{E}[R_i]}$, where the middle step uses the aforementioned upper bound on $\mathbf{E}[R_i^2]$. Combining everything, we have shown the following lower bound on the first term:

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|y_i\|_{\mathbf{B}}\right] \geq \frac{1}{2}\frac{1}{n}\sum_{i=1}^n \sqrt{\mathbf{E}\left[\|y_i\|_{\mathbf{B}}^2\right]} = \frac{1}{2}\frac{1}{n}\sum_{i=1}^n \sqrt{\|y_i\|_{\mathbf{M}}^2}\,,$$

where the second step uses that $\mathbf{E}[\mathbf{B}] = \mathbf{M}$. We now show that the second term in (16) can be ignored as follows:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|y_i\|_{\mathbf{B}}\mathbb{1}_{\mathbf{B}\notin\mathcal{B}_r}\right] &\leq \frac{1}{n}\sum_{i=1}^n \sqrt{\mathbf{E}\left[\|y_i\|_{\mathbf{B}}^2\right]}\sqrt{\mathbf{Pr}[\mathbf{B}\notin\mathcal{B}_r]} \qquad \text{(Cauchy-Schwarz)}\\ &\leq \frac{1}{n}\sum_{i=1}^n \|y_i\|_{\mathbf{M}}\sqrt{\frac{4C}{r}},\end{aligned}$$

where the last inequality follows by concentration of covariance of Gaussians, which we establish next. First, we observe that $\mathbf{M}$ must have rank at least $\mathrm{tr}(\mathbf{M})/\|\mathbf{M}\|_{\mathrm{op}} \geq k$. Since $g_1, \ldots, g_k$ are sampled i.i.d. from $\mathcal{N}(0, \mathbf{M})$ with $\mathrm{rank}(\mathbf{M}) \geq k$, the matrix $\mathbf{B} := \sum_i g_i g_i^\top$ has rank exactly k has rank exactly k with probability 1; This is because the Lebesgue measure of a rank-deficient subspace is zero. Thus, it remains to show that $\mathbf{B}$ has operator norm at most r with probability at least $1 - 1/32$. Observe that $\mathbf{B}$ is the second moment matrix of k independent Gaussians whose covariance matrix has trace equal to k. Applying the Gaussian covariance concentration results to this setting (see, e.g., (Koltchinskii and Lounici, 2017, Theorem 4) or (Van Handel, 2017, Theorem 5.1)), we obtain that for an absolute constant C:

$$\mathbf{E}[\|\mathbf{B}-\mathbf{M}\|_{\mathrm{op}}] \leq C\|\mathbf{M}\|_{\mathrm{op}}\left(\sqrt{\frac{1}{\|\mathbf{M}\|_{\mathrm{op}}}} + \frac{1}{\|\mathbf{M}\|_{\mathrm{op}}}\right) \leq C(1+\sqrt{\|\mathbf{M}\|_{\mathrm{op}}}) \leq 2C.$$

Since for any $r \geq 2$, $\mathbf{B} \notin \mathcal{B}_r$ implies that $\|\mathbf{B}-\mathbf{M}\|_{\mathrm{op}} \geq r - 1 \geq r/2$, applying the Markov inequality, we obtain the desired inequality $\mathbf{Pr}(\mathbf{B}\notin\mathcal{B}_{r/2}) \leq \frac{4C}{r}$.

Putting everything together and taking $r = 64C$, we obtain the following:

$$\mathbf{E}[Z] \geq \left(\frac{1}{2} - \sqrt{\frac{4C}{r}}\right)\frac{1}{n}\sum_{i=1}^n \|y_i\|_{\mathbf{M}} = \frac{1}{4}\frac{1}{n}\sum_{i=1}^n \|y_i\|_{\mathbf{M}} \geq C'\rho/4, \tag{17}$$

where we used the assumption $\sum_{i=1}^n \|y_i\|_{\mathbf{M}} > C'\rho$. If $C'/4 > \sqrt{r} = \sqrt{128C}$, this contradicts the upper bound $\mathbf{E}[Z] \leq \sqrt{r}\rho$ established earlier. ■

Appendix E. Omitted Details from Section 4.2

Lemma 27 *Consider the setting in Proposition 14. Then for all $|S''| \geq (1-\epsilon)|S'|$, we have $W_{1,k}(S_0', S'') \lesssim \rho + \epsilon\sqrt{k} + \delta$ and $W_{2,k}(S_0', S'') \lesssim \sqrt{\epsilon k} + \delta/\sqrt{\epsilon} + \rho/\sqrt{\epsilon}$.*

Proof Using the triangle inequality:

$$W_{1,k}(S_0', S'') \leq W_{1,k}(S_0', S') + W_{1,k}(S', S'') \; .$$

For the first term, note that S_0' and S' have the same cardinality, and for every point $x_i \in S_0'$ we have exactly one point $\widetilde{x}_i \in S'$ with $x_i - \widetilde{x}_i =: \Delta_i$. Thus we can use this simple coupling in Definition 5 to get this upper bound:

$$W_{1,k}(S_0', S') \leq \max_{\mathbf{V} \in \mathcal{V}_k} \frac{1}{|S_0'|} \sum_{i : i \in S_0'} \|\Delta_i\|_2 \lesssim \rho \; ,$$

where the last inequality follows by (5). For the second term we have $W_{1,k}(S', S'') \lesssim \epsilon\sqrt{k} + \widetilde{\delta} \leq \epsilon\sqrt{k} + \rho + \delta$ by Lemma 12, which is applicable because S'' is an $(1-\epsilon)$-subset of S' and S' is $(\epsilon, \widetilde{\delta}, k)$-generalized stable with $\widetilde{\delta} \lesssim \delta + \rho$ (we have shown this in the proof of the first part of Proposition 14).

The bound for $W_{2,k}(S_0', S'')$ is similar. First, $W_{2,k}(S_0', S'') \leq W_{2,k}(S_0', S') + W_{2,k}(S', S'')$. The first term is $O(\rho/\sqrt{\epsilon})$ by (5) and the second term is at most $O(\sqrt{\epsilon k} + \widetilde{\delta}/\sqrt{\epsilon}) = O(\sqrt{\epsilon k} + \delta/\sqrt{\epsilon} + \rho/\sqrt{\epsilon})$ by Lemma 12. ■

Appendix F. Omitted Details from Section 5: Mean Estimation

In this section, we restate and prove the first algorithmic result of our paper (Theorem 3).

Theorem 3 (Mean estimation) *Let C be a sufficiently large constant, $c = 1/C$. Let $\epsilon \in (0, c)$ and $\rho > 0$ be parameters. Let $S_0 \subset \mathbb{R}^d$ be an (ϵ, δ)-stable set with respect to an unknown $\mu \in \mathbb{R}^d$, where $\delta > \epsilon$. Let T be a corrupted dataset with an ϵ-fraction of global outliers and ρ-strong local corruptions (Global+StrongLC Model). Then, any stability-based algorithm $\mathcal{A}$ on input T, ϵ and $\widetilde{\delta} := C \cdot (\delta + \rho)$, outputs $\widehat{\mu} \in \mathbb{R}^d$ such that, with high probability, $\|\widehat{\mu} - \mu\|_2 = O(\delta + \rho)$.*

Proof Let $S_0 = \{x_i\}_{i \in [n]}$ be an (ϵ, δ)-stable set with respect to $\mu \in \mathbb{R}^d$. The final set T is constructed by first picking $S \in \mathcal{W}^{\text{strong}}(S_0, \rho)$ (cf. StrongLC Model) and then $T \in \mathcal{O}(S, \epsilon)$ (cf. Global Contamination Model) by the corresponding adversaries. Let Δ_i denote the perturbabtions of $\mathcal{W}^{\text{strong}}$ adversary, i.e., $S = \{\widetilde{x}_i\}_{i \in [n]}$ where $\widetilde{x}_i = x_i + \Delta_i$. By Proposition 8 (applied with $k = 1$), there exists a subset $S_0' \subset S_0$ of size at least $(1-\epsilon)n$ for which $\max_{\mathbf{M} \in \mathcal{M}_1} \frac{1}{|S_0'|} \sum_{i : x_i \in S_0'} \|\Delta_i\|_{\mathbf{M}}^2 \lesssim \rho^2/\epsilon$. Applying Proposition 13 with $k = 1$ and $y_i = \Delta_i$ for the set S_0', we have that $\sup_{\mathbf{M} \in \mathcal{M}_1} \frac{1}{|S_0'|} \sum_{i : x_i \in S_0'} \|\Delta_i\|_{\mathbf{M}} \lesssim \sup_{\mathbf{V} \in \mathcal{V}_1} \frac{1}{|S_0'|} \sum_{i : x_i \in S_0'} \|\Delta_i\|_{\mathbf{V}} \lesssim \rho$ (where the last inequality follows by the definition of the local perturbations model).

So far, we have established the necessary conditions in Equation (5) for applying Proposition 14 with $k = 1$. The condition in Proposition 14 that S_0' is $(\epsilon, O(\delta), 1)$-generalized stable is satisfied because (i) S_0 is (ϵ, δ) and (ii) S_0' is a large subset of S_0 (Lemma 11).

Proposition 14 guarantees that if S' denotes the set $\{x_i + \Delta_i : x_i \in S_0'\}$ (i.e., the points in S_0' after the local perturbations), then S' is $(\epsilon, \widetilde{\delta})$-stable with respect to μ (cf. Definition 1), for $\widetilde{\delta} \lesssim \rho + \delta$. After T is chosen by the second adversary (the one associated with the global outliers), we have that $|T \cap S| \geq (1-\epsilon)|T|$ which implies that $|T \cap S'| \geq |T \cap S| - |S \setminus S'| \geq (1-2\epsilon)|T|$. This means that $T \in \mathcal{O}(S', 2\epsilon)$ for an $(2\epsilon, O(\widetilde{\delta}))$-stable set S' (which follows from $(\epsilon, \widetilde{\delta})$-stability of S' and Lemma 11), and thus any stability-based algorithm outputs a $\hat{\mu}$ such that $\|\hat{\mu} - \mu\|_2 \lesssim \widetilde{\delta} \lesssim \rho + \delta$. ■

We conclude this section with a corollary of Theorem 3. In Appendix C we recorded known bounds for the stability parameter of a set of points sampled from well-behaved distributions, and presented an additional tightening of these bounds that we developed in this paper by using more recent concentration results. Combining the results of that section (Theorem 20) with Theorem 3 we immediately obtain robust mean estimators for these families that work in the combined contamination model.

Corollary 28 *Let $\mathcal{P}$ be a family of distributions. Fix a $P \in \mathcal{P}$ with the (unknown) mean μ, Let S_0 be a set of n i.i.d. samples from P. Let T be a corrupted version of S_0 with local contamination parameter ρ and global contamination rate ϵ (Global+StrongLC Model). Let τ be the failure probability such that $\log(1/\tau)/n$ is less than an absolute constant. There exist computationally-efficient algorithms that (i) take as input T, ϵ, $\mathcal{P}$, τ, and ρ and (ii) output $\widehat{\mu}$ that satisfies the following guarantees with probability $1-\tau$:*

- *If $\mathcal{P}$ is the family of isotropic subgaussian distributions, then $\|\widehat{\mu} - \mu\|_2 \lesssim \epsilon\sqrt{\log(1/\epsilon)} + \rho + \sqrt{d/n} + \sqrt{\log(1/\tau)/n}$.*
- *If $\mathcal{P}$ is the family of distributions with isotropic covariance and bounded k-th moments, then $\|\widehat{\mu} - \mu\|_2 \lesssim \epsilon^{1-\frac{1}{k}} + \rho + \sqrt{d/n} + \sqrt{\log(1/\tau)/n}$.*
- *If $\mathcal{P}$ is the family of distributions with covariance $\mathbf{\Sigma} \preceq \mathbf{I}$ then $\|\widehat{\mu} - \mu\|_2 \lesssim \sqrt{\epsilon} + \rho + \sqrt{\operatorname{tr}(\mathbf{\Sigma})/n} + \sqrt{\log(1/\tau)/n}$.*

We highlight that the error rates above are information-theoretically optimal in all the parameters $\epsilon, \rho, d, n, \tau$; See Appendix C for further discussion on optimality.

Appendix G. Omitted Details from Section 5: Distribution Learning

In this section, we prove Theorem 6 (restated below), which yields guarantees for distribution learning in the presence of the combined contamination model.

Theorem 6 (Distribution learning) *Let parameters $\epsilon \in (0, c)$ where c is a sufficiently small absolute constant, $\rho > 0$ and $\delta > \epsilon$. Let $S_0 \subset \mathbb{R}^d$ be a set that is (ϵ, δ)-stable with respect to an (unknown) $\mu \in \mathbb{R}^d$. For a slicing parameter $k \in [d]$, let T be the corrupted dataset after a combination of local and global corruptions from StrongWC Model and Global Contamination Model with parameters ρ and ϵ, respectively (i.e., $T \in \mathcal{O}(S, \epsilon)$ for some $S \in \mathcal{W}^{\text{strong}}_{1,k}(S_0, \rho)$). Then, there exists a polynomial-time algorithm that on input $T, \epsilon, \rho, k, \delta$, the algorithm outputs an estimate $\widehat{S} \subset T$ such that, with high constant probability, for all $k' \in [k]$ it holds that $W_{1,k'}(\widehat{S}, S_0) = O(\delta\sqrt{k'} + \rho)$.*

This result also relies on the structural result of Proposition 8 and Proposition 14, provided in Section 4. For the distribution learning result, we provide an algorithm which uses a multi-dimensional variant of the standard iterative filtering procedure, given in Algorithm 1. Then, leveraging Proposition 14, a certification lemma from Nietert et al. (2024) in Appendix G.1, and a now-standard analysis of the iterative filtering procedure, we prove Theorem 6.

The pseudocode is provided in Appendix G. However, we provide a copy here for convenience:

Algorithm 2 Distribution learning under global and strong local corruptions (Algorithm 1 reproduced)

```
Input: (Multi)-Set of samples $T \subset \mathbb{R}^d$, and parameters $\epsilon \in (0, c), \delta \geq \epsilon, \rho \geq 0$.
Output: $\widehat{S} \subset \mathbb{R}^d$ and $\widehat{\mu} \in \mathbb{R}^d$ such that $\|\widehat{\mu} - \mu\|_2 \lesssim \delta + \rho$ and $W_{1,k'}(\widehat{S}, S) \lesssim \delta\sqrt{k'} + \rho, \forall k' \in [k]$.
Let $C$ be a sufficiently large absolute constant
for $k' = 1, 2, \dots, k$ do
    Define $\widetilde{\delta} = \delta\sqrt{k'} + \rho$
    while true do
        Compute $\mathbf{M} \in \mathbb{R}^{d \times d}$ maximizing $\langle \mathbf{M}, \mathbf{\Sigma}_T \rangle$ under the constraints $0 \preceq \mathbf{M} \preceq \mathbf{I}$, $\mathrm{tr}(\mathbf{M}) = k'$
        if $\langle \mathbf{M}, \mathbf{\Sigma}_T \rangle \leq k' + C\widetilde{\delta}^2/\epsilon$ then          // Stopping condition
            Continue (exit while loop)
        else
            Let $L \subset T$ consist of the $\epsilon|T|$ points with the largest score $g(x) = (x - \mu_T)^\top \mathbf{M}(x - \mu_T)$
            Define the thresholded scores $\tau(x) := g(x)\mathbf{1}_{x \in L}$ for $x \in T$
            for each $x \in T$: Delete $x$ from $T$ with probability $\tau(x)/\max_{x \in T} \tau(x)$
Let $\widehat{S} \leftarrow T$
return $\widehat{S}$
```

G.1. Certification of Solutions

In this section, we state the certificate lemma (from Nietert et al. (2024)) that provides a way to bound the $W_{1,k'}$ distance between the current filtered version of the dataset T and the uniform distribution over the original inliers S_0. This bound is expressed as a function of a variance-like quantity of the dataset. This insight informs the design of our algorithm's stopping condition. Consequently, we can guarantee that upon termination, if the variance is sufficiently small, the solution output by the algorithm will be close to the uniform distribution over the inliers S_0 in the $W_{1,k'}$ metric.

Lemma 29 (Lemma 20 in Nietert et al. (2024)) *Let S_0 be an (ϵ, δ)-stable set. Let S' be any set satisfying $W_{2,k'}(S', S_0) \leq r$ and T be a set with $|T \cap S'| \geq (1 - \epsilon)|T|$. Then, the following holds:*[15]

$$W_{1,k'}(T, S') \lesssim \delta\sqrt{k'} + r\sqrt{\epsilon} + \epsilon\sqrt{r'} \ ,$$

where $r' = \sup_{\mathbf{M} \in \mathcal{M}_{k'}} \langle \mathbf{M}, \mathbf{\Sigma}_{T \setminus S'} \rangle + (\mu_{T \setminus S'} - \mu_{S'})^\top \mathbf{M}(\mu_{T \setminus S'} - \mu_{S'})$.

The next result provides an upper bound for the quantity r' appearing in Lemma 29 in terms of the simpler quantity $\langle \mathbf{M}, \mathbf{\Sigma}_T \rangle - k'$. This simpler quantity acts as the stopping condition of our algorithm. In addition, the lemma below also bounds $\|\mu_T - \mu\|_{\mathbf{M}}$, which is the error of the empirical mean over the current dataset T.

[15] The following result is implied by (Nietert et al., 2024, Lemma 20) after using $P' = P$ both being equal to the uniform distribution over S', and Q being the uniform distribution over T.

Lemma 30 *Let k' be a positive integer, $\mathbf{M} \in \mathcal{M}_{k'}$, and $\widetilde{\delta} \geq \epsilon$. Let S' be a set satisfying the $(\epsilon, \widetilde{\delta}, k')$-generalized-stability with respect to the vector $\mu \in \mathbb{R}^d$ (cf. Definition 7). Let T be a set such that $|T \cap S'| \geq (1-\epsilon)|T|$ and denote $\lambda := \langle \mathbf{M}, \mathbf{\Sigma}_T \rangle - k'$. Then, the following hold:*[16]

1. $\|\mu_T - \mu\|_{\mathbf{M}} \lesssim \widetilde{\delta} + \epsilon\sqrt{k'} + \sqrt{\lambda\epsilon}$.
2. $\max\left(\langle \mathbf{M}, \mathbf{\Sigma}_{T\setminus S'} \rangle, \|\mu_{S'\cap T} - \mu_{T\setminus S'}\|_{\mathbf{M}}^2\right) \lesssim \lambda/\epsilon + \widetilde{\delta}^2/\epsilon^2 + k'$.

Proof The covariance matrix can be decomposed as follows:

$$\mathbf{\Sigma}_T = (1-\epsilon)\mathbf{\Sigma}_{S'\cap T} + \epsilon\mathbf{\Sigma}_{T\setminus S'} + \epsilon(1-\epsilon)(\mu_{S'\cap T} - \mu_{T\setminus S'})(\mu_{S'\cap T} - \mu_{T\setminus S'})^\top .$$

Using the decomposition above with our assumptions, we obtain

$$\begin{aligned} k' + \lambda &\geq \langle \mathbf{M}, \mathbf{\Sigma}_T \rangle \\ &= (1-\epsilon)\langle \mathbf{M}, \mathbf{\Sigma}_{S'\cap T} \rangle + \epsilon\langle \mathbf{M}, \mathbf{\Sigma}_{T\setminus S'} \rangle + \epsilon(1-\epsilon)(\mu_{S'\cap T} - \mu_{T\setminus S'})^\top \mathbf{M}(\mu_{S'\cap T} - \mu_{T\setminus S'}) \\ &\geq (1-\epsilon)\left(k' - O\left(\frac{\widetilde{\delta}^2}{\epsilon}\right)\right) + \epsilon\langle \mathbf{M}, \mathbf{\Sigma}_{T\setminus S'} \rangle + \epsilon(1-\epsilon)(\mu_{S'\cap T} - \mu_{T\setminus S'})^\top \mathbf{M}(\mu_{S'\cap T} - \mu_{T\setminus S'}) , \end{aligned} \tag{18}$$

where the second line is derived by the assumption that S' is a set satisfying the $(\epsilon, \widetilde{\delta}, k')$-generalized-stability as follows:

$$\begin{aligned} \langle \mathbf{M}, \mathbf{\Sigma}_{S'\cap T} \rangle &= \left\langle \mathbf{M}, \frac{1}{|S'\cap T|}\sum_{x\in S'\cap T}(x - \mu_{S'\cap T})(x - \mu_{S'\cap T})^\top \right\rangle \\ &= \left\langle \mathbf{M}, \frac{1}{|S'\cap T|}\sum_{x\in S'\cap T}(x-\mu)(x-\mu)^\top \right\rangle + \left\langle \mathbf{M}, (\mu - \mu_{S'\cap T})(\mu - \mu_{S'\cap T})^\top \right\rangle \\ &+ \left\langle \mathbf{M}, \frac{1}{|S'\cap T|}\sum_{x\in S'\cap T}(x-\mu)(\mu-\mu_{S'\cap T})^\top \right\rangle + \left\langle \mathbf{M}, \frac{1}{|S'\cap T|}\sum_{x\in S'\cap T}(\mu-\mu_{S'\cap T})(x-\mu)^\top \right\rangle \\ &\geq \left\langle \mathbf{M}, \frac{1}{|S'\cap T|}\sum_{x\in S'\cap T}(x-\mu)(x-\mu)^\top \right\rangle - (\mu - \mu_{S'\cap T})^\top \mathbf{M}(\mu - \mu_{S'\cap T}) \\ &\geq k' - \frac{\widetilde{\delta}^2}{\epsilon} - 2\|\mathbf{M}\|_{\text{op}}\|\mu - \mu_{S'\cap T}\|_2 \geq k' - \frac{\widetilde{\delta}^2}{\epsilon} - 2\widetilde{\delta} \geq k' - \frac{3\widetilde{\delta}^2}{\epsilon} , \end{aligned}$$

where we used that $\mathbf{M} \preceq \mathbf{I}$, $\|\mu - \mu_{S'\cap T}\|_2 \lesssim \widetilde{\delta}$ by the stability assumption for S' and $\widetilde{\delta} \geq \epsilon$. Rearranging (18) yields

$$\langle \mathbf{M}, \mathbf{\Sigma}_{T\setminus S'} \rangle + (1-\epsilon)(\mu_{S'\cap T} - \mu_{T\setminus S'})^\top \mathbf{M}(\mu_{S'\cap T} - \mu_{T\setminus S'}) \lesssim \lambda/\epsilon + \widetilde{\delta}^2/\epsilon^2 + k' . \tag{19}$$

This implies that both $\langle \mathbf{M}, \mathbf{\Sigma}_{T\setminus S'} \rangle$ and $(\mu_{S'\cap T} - \mu_{T\setminus S'})^\top \mathbf{M}(\mu_{S'\cap T} - \mu_{T\setminus S'})$ are at most $O(\lambda/\epsilon + \widetilde{\delta}^2/\epsilon^2 + k')$, which shows the second part of our lemma. The first part of Lemma 30 follows simply by the decomposition below:

$$\|\mu_T - \mu\|_{\mathbf{M}} = \|(1-\epsilon)\mu_{S'\cap T} + \epsilon\mu_{T\setminus S'} - \mu\|_{\mathbf{M}}$$

[16] Recall that $\|x\|_{\mathbf{M}} = \sqrt{x^\top \mathbf{M} x}$ denotes the Mahalanobis norm of x with respect to $\mathbf{M}$.

$$
\begin{aligned}
&\le \|\mu_{S'\cap T} - \mu\|_{\mathbf{M}} + \epsilon \|\mu_{S'\cap T} - \mu_{T\setminus S'}\|_{\mathbf{M}} \\
&\lesssim \widetilde{\delta} + \epsilon\sqrt{k'} + \sqrt{\lambda\epsilon}\ .
\end{aligned}
$$

where we used the generalized-stability assumption to bound the first term, and $(\mu_{S'\cap T} - \mu_{T\setminus S'})^\top \mathbf{M}(\mu_{S'\cap T} - \mu_{T\setminus S'}) = O(\lambda/\epsilon + \widetilde{\delta}^2/\epsilon^2 + k')$ from earlier to bound the second term.

■

G.2. Filtering Scores

In this subsection, we show that the scores $\tau(x)$ used in Algorithm 1 remove more outliers than inliers in expectation in each round.

Lemma 31 (Analysis of one round of filtering: scores of outliers > scores of inliers) *Let S' be a multi-set of $\mathbb{R}^d$ satisfying $(\epsilon, \widetilde{\delta}, k')$-generalized-stability with respect to $\mu \in \mathbb{R}^d$. Assume the following: $\epsilon \in (0, c)$ for a sufficiently small absolute constant c, and $\widetilde{\delta} \ge \epsilon\sqrt{k'}$. Let T be a multiset such that $|T \cap S'| \ge (1 - 20\epsilon)$. Let C be a sufficiently large absolute constant. Let $\tau(x)$ be the scores as defined in Line 2 of Algorithm 1, i.e., $g(x) := \|x - \mu_T\|_{\mathbf{M}}^2$, L is the set of points in T with the $\epsilon \cdot |T|$ largest scores $g(x)$, and $\tau(x) := g(x)\mathbb{1}_{x\in L}$. If $\mathbf{M} \in \mathcal{M}_{k'}$ is a matrix with $\langle \mathbf{M}, \mathbf{\Sigma}_T\rangle > k' + C\widetilde{\delta}^2/\epsilon$, then $\sum_{x\in S'\cap T} \tau(x) \le 0.1 \sum_{x\in T} \tau(x)$.*

Proof

Denote $\lambda := \langle \mathbf{M}, \mathbf{\Sigma}_T\rangle - k'$. For the inlier points, we have the following (explanations are provided after the inequalities):

$$
\begin{aligned}
\sum_{x\in S'\cap T} \tau(x) &= \sum_{x\in S'\cap L} g(x) = \sum_{x\in S'\cap L} \|x - \mu_T\|_{\mathbf{M}}^2 && (20)\\
&\le 2 \sum_{x\in S'\cap L} (x-\mu)^\top \mathbf{M}(x-\mu) + 2|S'\cap L| \cdot \|\mu - \mu_T\|_{\mathbf{M}}^2 && (21)\\
&\lesssim |S'|\widetilde{\delta}^2/\epsilon + \epsilon|S'|(\widetilde{\delta}^2 + \epsilon^2 k' + \epsilon\lambda) \le 0.01\lambda|S'|\ , && (22)
\end{aligned}
$$

where the steps used were the following: (21) used the triangle inequality combined with the inequality $(a+b)^2 \le 2a^2 + 2b^2$, the first term in (22) was bounded using the third condition in Definition 9 (and our assumption that S' satisfies the generalized stability condition), the second term in (22) used that $\|\mu - \mu_T\|_{\mathbf{M}}^2 \lesssim \widetilde{\delta}^2 + \epsilon^2 k' + \epsilon\lambda$ by the certificate lemma (Lemma 30), and we also used that $|S' \cap L| \le |L| = \epsilon|T| \le \frac{\epsilon}{1-\Omega(\epsilon)}|S'|$. The last inequality in (22) uses our assumptions $\lambda \ge C\widetilde{\delta}^2/\epsilon$, $\widetilde{\delta} \ge \epsilon$, $\widetilde{\delta} \ge \epsilon\sqrt{k'}$, $C \gg 1$, $\epsilon \ll 1$.

We now show the lower bound for the sum of the scores over all points:

$$
\sum_{x\in T} \tau(x) = \sum_{x\in T\cap L} g(x) \ge \sum_{x\in T\setminus S'} g(x) \ge \sum_{x\in T} g(x) - \sum_{x\in S'\cap T} g(x)\ . \tag{23}
$$

where the second step above is based on the fact that $|T \setminus S'| \le \epsilon|T|$ and that L is defined to be the points with the largest $\epsilon|T|$ scores. For the first term, we have that, by definition:

$$
\sum_{x\in T} g(x) = \langle \mathbf{M}, \mathbf{\Sigma}_T\rangle = (k' + \lambda)|T| \ge (k' + \lambda)(1-\epsilon)|S'|\ . \tag{24}
$$

For the second term in the RHS of (23), we have the following:

$$
\begin{aligned}
\sum_{x\in S'\cap T} g(x) \le \sum_{x\in S'} g(x) &= \sum_{x\in S'}(x-\mu_T)^\top \mathbf{M}(x-\mu_T) \\
&= \sum_{x\in S'}(x-\mu)^\top \mathbf{M}(x-\mu) + |S'|(\mu-\mu_T)^\top \mathbf{M}(\mu-\mu_T) + 2\sum_{x\in S'}(x-\mu)^\top \mathbf{M}(\mu-\mu_T)\,.
\end{aligned}
\tag{25}
$$

The first term is at most $(k'+\widetilde{\delta}^2/\epsilon)|S'|$ by our generalized-stability assumption. The second term is at most $|S'|(\widetilde{\delta}^2+\epsilon^2 k'+\epsilon\lambda)$ by Lemma 30. For the third term, we have that

$$
\begin{aligned}
\frac{1}{|S'|}\sum_{x\in S'}(x-\mu_T)^\top \mathbf{M}(\mu-\mu_T) = (\mu_{S'}-\mu_T)^\top \mathbf{M}(\mu-\mu_T) &\le \|\mu_{S'}-\mu_T\|_2\|\mathbf{M}\|_{\mathrm{op}}\|\mu-\mu_T\|_2 \\
&\lesssim (\|\mu-\mu_{S'}\|_2+\|\mu-\mu_T\|_2)\|\mu-\mu_T\|_2 \lesssim \widetilde{\delta}^2+\epsilon^2 k'+\lambda\epsilon\,,
\end{aligned}
\tag{26}
$$

where we used that $\|\mathbf{M}\|_{\mathrm{op}}\le 1$, and then we applied the triangle inequality and Lemma 30. Putting (23)-(26) together, we have that

$$
\begin{aligned}
\sum_{x\in T}\tau(x) &\ge (k'+\lambda)(1-\epsilon)|S'| - (k'+\widetilde{\delta}^2+\widetilde{\delta}^2/\epsilon+\epsilon^2 k'+\lambda\epsilon)|S'| \\
&\ge (k'+\lambda)(1-\epsilon)|S'| - (k'+0.001\lambda)|S'| \\
&\qquad\qquad \text{(using } \lambda\ge C\widetilde{\delta}^2/\epsilon,\ \widetilde{\delta}\ge\epsilon,\ \widetilde{\delta}\ge\epsilon\sqrt{k'},\ C\gg 1,\ \epsilon\ll 1\text{)} \\
&\ge (\lambda-\epsilon k'-\lambda\epsilon-0.001\lambda)|S'| \\
&\ge 0.9\lambda|S'|\,. \qquad\qquad \text{(using } \epsilon\ll 1,\ \lambda\ge C\widetilde{\delta}^2/\epsilon\ge C\epsilon k'\text{)}
\end{aligned}
$$

Combining with (22) concludes the proof of this lemma.

■

G.3. Proof of Theorem 6

We are now ready to combine the previous components to conclude the analysis of our algorithm and complete the proof of Theorem 6.

Proof (Proof of Theorem 6) We briefly recall the notation. As in the theorem statement, $S_0=\{x_1,\dots,x_n\}$ is the original set of inliers (before any kind of corruptions), which is assumed to satisfy the stability conditions, $S=\{x_i+\Delta_i\}_{i\in[n]}$ is the set after the strong Wasserstein corruptions of StrongWC Model, (Δ_i's denote the shift that each point undergoes), and T is the final dataset after globally corrupting S (Global Contamination Model).

The set $S_0=\{x_1,\dots,x_n\}$ is (ϵ,δ)-stable by assumption, which means that, by Lemma 10, for any $k'\in[k]$, it also satisfies (ϵ,δ',k')-generalized stability, with $\delta'\lesssim\sqrt{k'}\delta$. By Proposition 8 we have the existence of a set $S_0'\subset S_0$ with $|S_0'|\ge(1-\epsilon)|S_0|$ such that for every $k'\in[k]$ it holds $\max_{\mathbf{M}\in\mathcal{M}_{k'}}\frac{1}{|S_0'|}\sum_{i:x_i\in S_0'}\|\Delta_i\|_{\mathbf{M}}^2\le\max_{\mathbf{M}\in\mathcal{M}_k}\frac{1}{|S_0'|}\sum_{i:x_i\in S_0'}\|\Delta_i\|_{\mathbf{M}}^2\lesssim\rho^2/\epsilon$ (the last inequality is the conclusion of Proposition 8; the first one follows trivially by the definition of $\mathcal{M}_k$). By Proposition 13 we have that the same set S_0' also satisfies the following for every $k'\in[k]$: $\max_{\mathbf{M}\in\mathcal{M}_{k'}}\frac{1}{|S_0'|}\sum_{i:x_i\in S_0'}\|\Delta_i\|_{\mathbf{M}}\le\max_{\mathbf{M}\in\mathcal{M}_k}\frac{1}{|S_0'|}\sum_{i:x_i\in S_0'}\|\Delta_i\|_{\mathbf{M}}\lesssim$

$\max_{\mathbf{V}\in\mathcal{V}_k} \frac{1}{|S_0'|}\sum_{i:x_i\in S_0'} \|\Delta_i\|_{\mathbf{V}} \lesssim \rho$ (again, the first inequality is trivial, the second inequality is the actual application of the proposition, and the last inequality is by definition of our local perturbation model).

Now define $S' \overset{\text{def}}{=} \{x_i + \Delta_i : x_i \in S_0'\}$. By applying Proposition 14 for each $k' \in [k]$, we obtain that for all $k' \in [k]$ S' satisfies $(\epsilon, \widetilde{\delta}_{k'}, k')$-generalized-stability with respect to μ, for $\widetilde{\delta}_{k'} = C \cdot (\delta\sqrt{k'} + \rho)$ for some sufficiently large constant. Moreover, for any $(1-\epsilon)|S'|$ sized subset S'' of S', and any $k' \in [k]$, it holds that

$$\begin{aligned} W_{1,k'}(S_0', S'') &\lesssim \rho + \epsilon\sqrt{k'} + \widetilde{\delta}_{k'} \lesssim \rho + \delta\sqrt{k'} \\ \text{and } W_{2,k'}(S_0', S'') &\lesssim \sqrt{\epsilon k'} + \widetilde{\delta}_{k'}/\sqrt{\epsilon} + \rho/\sqrt{\epsilon} \lesssim (\sqrt{k'}\delta + \rho)/\sqrt{\epsilon}, \end{aligned} \tag{27}$$

where we used that $\delta \geq \epsilon$.

We now argue that filtering does not remove too many "stable inliers" ($T \cap S'$) throughout its execution. Lemma 31 states that, as long as the main while loop of the algorithm has not been terminated, the scores $\tau(x)$ that the algorithm assigns to inlier points in T is substantially bigger (at least by a constant factor) than the ones for outlier points. Following the standard analysis of filtering algorithms in Diakonikolas and Kane (2023), we obtain that with probability at least $9/10$, we have that $|T \triangle S'| \leq 20\epsilon$ throughout the execution.

Let $\widehat{S}$ denote the set T at the end of the nested loop and condition on the high probability event that satisfies (i) $|\widehat{S} \cap S'| \leq 20\epsilon$ and (ii) for all $k' \in [k]$ and for all $\mathbf{M} \in \mathcal{M}_{k'}$ it holds $\langle \mathbf{M}, \mathbf{\Sigma}_{\widehat{S}} \rangle \leq k' + C\widetilde{\delta}_{k'}^2/\epsilon$. Fix any $k' \in [k]$. In what follows we show that $W_{1,k'}(S_0, S') \lesssim \delta\sqrt{k'} + \rho$. To apply Lemma 29, we need a bound on $W_{2,k'}(S_0, S')$, which we obtain below:

$$W_{2,k'}(S_0, S') \leq W_{2,k'}(S_0, S_0') + W_{2,k'}(S_0', S') \leq \sqrt{k'\epsilon} + \delta\sqrt{k'/\epsilon} + \sqrt{\rho^2/\epsilon},$$

where use Lemma 12 for the first term and (27) for the second term (with $S'' = S'$). With this bound on $W_{2,k'}(S', S_0')$, applying Lemma 29 yields

$$W_{1,k'}(\widehat{S}, S') \lesssim \delta\sqrt{k'} + \rho + \epsilon\sqrt{r'}\ ,$$

where r' is defined in Lemma 29. To upper bound r', we apply Lemma 30 with $T = \widehat{S}$ and $\epsilon' = 20\epsilon$ in place of the parameter ϵ appearing in the statement of that lemma (regarding the applicability of the lemma, $\hat{S}$ is $(\epsilon, O(\widetilde{\delta}_{k'}), k')$-generalized-stable due to the $(\epsilon, \widetilde{\delta}_{k'}, k')$-generalized-stability of S' and Lemma 11). This gives that

$$\begin{aligned} r' &= \sup_{\mathbf{M}\in\mathcal{M}_{k'}} \langle \mathbf{M}, \mathbf{\Sigma}_{\widehat{S}\setminus S'} \rangle + \|\mu_{\widehat{S}\setminus S'} - \mu_{S'}\|_{\mathbf{M}}^2 \\ &\lesssim \sup_{\mathbf{M}\in\mathcal{M}_{k'}} \langle \mathbf{M}, \mathbf{\Sigma}_{\widehat{S}\setminus S'} \rangle + \|\mu_{\widehat{S}\setminus S'} - \mu_{S'\cap\widehat{S}}\|_{\mathbf{M}}^2 + \|\mu_{S'\cap\widehat{S}} - \mu_{S'}\|_{\mathbf{M}}^2 \\ &\lesssim \widetilde{\delta}_{k'}^2/\epsilon^2 + k' + \|\mu_{S'\cap\widehat{S}} - \mu_{S'}\|_{\mathbf{M}}^2 \\ &\lesssim \widetilde{\delta}_{k'}^2/\epsilon^2\ . \end{aligned}$$

where the last line used $\|\mu_{S'\cap\widehat{S}} - \mu_{S'}\|_{\mathbf{M}}^2 \lesssim \|\mu_{S'\cap\widehat{S}} - \mu\|_{\mathbf{M}}^2 + \|\mu - \mu_{S'}\|_{\mathbf{M}}^2 \lesssim \widetilde{\delta}_{k'}^2/\epsilon^2$ by the generalized stability of S' (we also used that $\widetilde{\delta}_{k'} \geq \epsilon$ and $\widetilde{\delta}_{k'} \geq \epsilon\sqrt{k'}$). Plugging this back, we obtain a bound of $W_{1,k'}(\widehat{S}, S') \lesssim \delta\sqrt{k'} + \rho$. We can translate this into a bound for $W_{1,k'}(S_0, \widehat{S})$ using the triangle inequality as follows:

$$W_{1,k'}(S_0, \widehat{S}) \leq W_{1,k'}(S_0, S_0') + W_{1,k'}(S_0', S') + W_{1,k'}(S', \widehat{S}).$$

The first term above is upper bounded by $\epsilon\sqrt{k'} + \delta$ by Lemma 12, the second term is upper bounded by $\delta\sqrt{k'} + \rho$ by Equation (27), and the last term was shown to be at most $\delta\sqrt{k'} + \rho$. Combining these three terms yields the desired result.

Runtime Note that the algorithm removes at least one point per iteration and the set of inliers $S' \cap T$ satisfies the stopping condition of Algorithm 1. This is because of stability of S' and therefore stability of any large subset of S' (cf. Lemma 11). This means that the algorithm will terminate after $O(n)$ iterations. In each iteration, the algorithm requires solving an SDP (Algorithm 2), which can be done in polynomial time by ellipsoid method or interior point method Nesterov (2004). ■

Appendix H. Principal Component Analysis

In this section, we present our result for robust principal component analasysis (PCA) in Theorem 35 below. The goal for PCA is to output a high-variance direction v of the unknown covariance $\mathbf{\Sigma}$ in the following sense: $v^\top \mathbf{\Sigma} v \geq (1-\gamma)\|\mathbf{\Sigma}\|_{\text{op}}$ for γ as small as possible. Under the global contamination model, robust PCA algorithms have been developed in Kong et al. (2020); Jambulapati et al. (2020); Diakonikolas et al. (2023); Jambulapati et al. (2024).

We will work with zero-mean distributions (for inliers) in this section, which is without loss of generality because one can always reduce to this setting by taking differences of pairs of samples.

We state an appropriate version of the stability condition which is more relevant to PCA.

Definition 32 (PCA stability) *Let $0 < \epsilon \leq \gamma$. A finite multiset $S \subset \mathbb{R}^d$ is called (ϵ, γ)-PCA-stable with respect to a PSD matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ if for every $S' \subseteq S$ with $|S'| \geq (1-\epsilon)|S|$, the following holds: $(1-\gamma)\mathbf{\Sigma} \preceq \frac{1}{|S'|}\sum_{x\in S'} xx^\top \preceq (1+\gamma)\mathbf{\Sigma}$.*

The definition above is very closely related to Definition 7 as shown by the following observation.

Fact 33 *Let $\mathbf{\Sigma}$ be a positive definite matrix. If a set of samples $\{\mathbf{\Sigma}^{-1/2}x_i\}_{i\in[n]}$ is (ϵ, δ)-stable (Definition 1) with respect to $\mu = 0$ (Definition 7), then $\{x_i\}_{i\in[n]}$ is (ϵ, γ)-PCA-stable with respect to $\mathbf{\Sigma}$ (Definition 32) for $\gamma \lesssim \delta^2/\epsilon$.*

Using the connection above, it can be seen that the stability definition above is satisfied by many distribution families of interest. Similarly, a set of i.i.d. samples from such distributions continue to satisfy this definition with high probability (Jambulapati et al., 2020, 2024). Consequently, the stability-based algorithms obtain the state of the art results for robust PCA for many distribution families (Jambulapati et al., 2020; Diakonikolas et al., 2023; Jambulapati et al., 2024).

Definition 34 (Stability-based algorithms for PCA) *Let S be an (ϵ, γ)-PCA-stable set with respect to an (unknown) PSD matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ (Definition 32). Let T be any set in $\mathcal{O}(S, \epsilon)$ (cf. Global Contamination Model). We call an algorithm stability-based PCA-algorithm if it takes as an input T, ϵ, and γ, and outputs a unit vector $v \in \mathbb{R}^d$ in polynomial time such that $v^\top \mathbf{\Sigma} v \geq (1 - O(\gamma))\|\mathbf{\Sigma}\|_{\text{op}}$.*

For this section, we consider a version of the contamination model where local corruptions are introduced to the data after whitening.

Contamination Model 1 (Strong local contamination after whitening) *Let $\overline{\rho} \geq 0$. Let $S_0 = \{x_1, \dots, x_n\}$ be an n-sized set in $\mathbb{R}^d$ and $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ be a PSD matrix. Consider an adversary that perturbs each point x_i to $\widetilde{x}_i$ with the only restriction that in each direction, the average perturbation is at most $\overline{\rho}$. Formally, we define*

$$\mathcal{W}^{\text{strong}}(S_0, \overline{\rho}, \mathbf{\Sigma}) := \left\{ S = \{\widetilde{x}_1, \dots, \widetilde{x}_n\} \subset \mathbb{R}^d : \sup_{v \in \mathbb{R}^d : \|v\|_2 = 1} \frac{1}{n} \sum_{i \in [n]} \left| v^\top \mathbf{\Sigma}^{-1/2} (\widetilde{x}_i - x_i) \right| \leq \overline{\rho} \right\} .$$

The adversary returns an arbitrary set $S \in \mathcal{W}^{\text{strong}}(S_0, \overline{\rho}, \mathbf{\Sigma})$ after possibly reordering the points.

As a remark, we note that Contamination Model 1 and StrongLC Model are equivalent to each other, as long as the matrix $\mathbf{\Sigma}$ is well-conditioned. In particular, $\overline{\rho} \leq \rho / \sqrt{\lambda_{\min}}$ and $\rho \leq \overline{\rho} \sqrt{\lambda_{\max}}$ where $\lambda_{\max}, \lambda_{\min}$ denote the largest and smallest eigenvalues of $\mathbf{\Sigma}$ respectively. The appealing property of the whitened local perturbations is that (i) it allows the amount of local perturbations to increase in high-variance directions, and (ii) it decouples the local contamination parameter $\overline{\rho}$ from the scale of the covariance matrix $\mathbf{\Sigma}$.[17]

Theorem 35 *Let c be a sufficiently small positive constant and C a sufficiently large constant. Let outlier rate $\epsilon \in (0, c)$ and contamination radius $\overline{\rho} \in (0, \sqrt{\epsilon})$. Let S_0 be a set of samples satisfying (ϵ, γ)-PCA-stability with respect to a PSD matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ (Definition 32). Let T be a corrupted dataset after ϵ-fraction of outliers and $\overline{\rho}$-strong local corruptions after whitening (as per Contamination Model 1). Then, any stability-based PCA algorithm (Definition 34) on input $T, \epsilon, \widetilde{\gamma} = C \cdot (\gamma + \frac{\overline{\rho}^2}{\epsilon})$, outputs a unit vector v such that with high probability (over the internal randomness of the algorithm): $v^\top \mathbf{\Sigma} v \geq \left(1 - O\left(\gamma + \frac{\overline{\rho}^2}{\epsilon}\right)\right) \|\mathbf{\Sigma}\|_{\text{op}}$.*

Proof For a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ and set $S \subset \mathbb{R}^d$, we use $\mathbf{A}[S]$ to denote the set $\{\mathbf{A}x : x \in S\}$. Let S be the set after the local perturbations of S_0 (as per Contamination Model 1). It suffices to show that S contains a subset $S' \subset S$ such that $|S'| \geq (1 - \epsilon)$ and S' is $(\epsilon, \widetilde{\gamma})$-PCA stable with respect to $\mathbf{\Sigma}$ for $\widetilde{\gamma} \lesssim \gamma + \overline{\rho}^2 / \epsilon$. By Fact 33, it suffices to show that the whitened data $\mathbf{\Sigma}^{-1/2}[S]$ contains a large subset S' that is $(\epsilon, \widetilde{\gamma})$-PCA-stable with respect to $\mathbf{I}$. Leveraging the connections between PCA-stability and the usual stability (Definition 7), it suffices to show that S' satisfies the conditions pertaining to the second moment of $(\epsilon, \sqrt{\epsilon \gamma} + \overline{\rho})$-stability (with respect to $\mu = 0$). The existence of a large S' with the desired stability can be shown by following the proof in Proposition 14 mutatis mutandis for $k = 1$ and $\delta = \sqrt{\epsilon \gamma}$.[18]

■

[17] Indeed, it can be seen that the range of parameter ρ where robust PCA is non-trivial depends on the scale of $\mathbf{\Sigma}$. For example, consider the case when the inlier distribution is $\mathcal{N}(0, \sigma^2(\mathbf{I} + vv^\top))$ for a unit vector v and a scalar σ^2. Then the 2-Wasserstein distance between $\mathcal{N}(0, \sigma^2(\mathbf{I} + vv^\top))$ and $\mathcal{N}(0, \sigma^2 \mathbf{I})$ is $\Theta(\sigma)$. Hence, for $\rho \gtrsim \sigma$, the local adversary can simulate samples from an isotropic distribution, removing any signal from the direction of interest v. On the other hand, as shown in Theorem 35, the range of $\overline{\rho}$ does not depend on $\mathbf{\Sigma}$.

[18] In fact, if we make the stronger assumption in the theorem that $\mathbf{\Sigma}^{-1/2}[S_0]$ is $(\epsilon, \delta, 1)$-generalized stable (as opposed to PCA stable), then the desired conclusion follows as a direct corollary from Proposition 14 and Fact 33.

Finally, we briefly mention how to generalize Theorem 35 to k-robust PCA for $k > 1$. Observe that in the proof above, we have shown that S contains a large subset that is $(\epsilon, \widetilde{\gamma})$-PCA-stable for $\widetilde{\gamma} \lesssim \gamma + \overline{\rho}^2/\epsilon$. Generalization to $k > 1$ then follows directly from (Jambulapati et al., 2024, Corollary 3).

Appendix I. Sum-of-Squares Based Algorithm: Proof of Theorem 4

In this section, we prove the result on mean estimation under the combined contamination model for distributions with certifiably bounded moments in the sum-of-squares (SoS) proof system. We show that the approach of Kothari et al. (2018); Hopkins and Li (2018) extends to the contamination model of this paper.

We refer the reader to Barak and Steurer (2016); Fleming et al. (2019) for the necessary definitions of the terms such as degree-d SoS proofs and pseudoexpectations. We list only a few basic facts that we use and refer the reader to the aforementioned references for the full background.

Fact 36 (Cauchy-Schwarz for pseudoexpectations) *Let f, g be polynomials of degree at most t. Then, for any degree-$2t$ pseudoexpectation $\widetilde{\mathbf{E}}$, $\widetilde{\mathbf{E}}[fg] \leq \sqrt{\widetilde{\mathbf{E}}[f^2]}\sqrt{\widetilde{\mathbf{E}}[g^2]}$. Consequently, for every squared polynomial p of degree t, and k a power of two, $\widetilde{\mathbf{E}}[p^k] \geq (\widetilde{\mathbf{E}}[p])^k$ for every $\widetilde{\mathbf{E}}$ of degree-$2tk$.*

Fact 37 (SoS triangle inequality) *If k is a power of two, $\Big|\frac{a_1,a_2,\ldots,a_n}{k} \left\{(\sum_i a_i)^k \leq n^k \left(\sum_i a_i^k\right)\right\}$.*

Fact 38 (SoS Cauchy-Schwartz and Hölder) *Let $f_1, g_1, \ldots, f_n, g_n$ be indeterminates over $\mathbb{R}$. Then,*

$$\Big|\frac{f_1,\ldots,f_n,g_1,\ldots,g_n}{2} \left\{ \left(\frac{1}{n}\sum_{i=1}^n f_i g_i\right)^2 \leq \left(\frac{1}{n}\sum_{i=1}^n f_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n g_i^2\right) \right\} .$$

Moveover, if $p_1, \ldots, p_n$ are indeterminates, for any $t \in \mathbb{Z}_+$ that is a power of 2, we have that

$$\{w_i^2 = w_i \mid i \in [n]\} \Big|\frac{p_1,\ldots,p_n}{O(t)} \left(\sum_i w_i p_i\right)^t \leq \left(\sum_{i\in[n]} w_i\right)^{t-1} \cdot \sum_{i\in[n]} p_i^t \quad \textit{and}$$

$$\{w_i^2 = w_i \mid i \in [n]\} \Big|\frac{p_1,\ldots,p_n}{O(t)} \left(\sum_i w_i p_i\right)^t \leq \left(\sum_{i\in[n]} w_i\right)^{t-1} \cdot \sum_{i\in[n]} w_i p_i^t .$$

Definition 39 (Certifiably bounded moments) *For an even $t \in \mathbb{N}$, we say a distribution P with mean μ_P over $\mathbb{R}^d$ has (t, M_t)-certifiably bounded moments if the polynomial inequality $p(v) \geq 0$ for $p(v) := M_t^t - \mathbf{E}_{X\sim P}[\langle v, X - \mu_P\rangle^t]$ has an SoS proof of degree $O(t)$ under the assumption $\|v\|_2^2 = 1$. If S is a set of points in $\mathbb{R}^d$, we say that S has (t, M_t)-certifiably bounded moments if the uniform distribution over S satisfies the previous definition.*

Many distribution families, such as rotationally invariant distributions, t-wise product distributions with bounded moments, and Poincare distributions are known to be certifiably bounded (Kothari et al., 2018). Furthermore, all subgaussian distributions also have certifiably bounded moments (Diakonikolas et al., 2024).

Theorem 40 *Let $\epsilon \in (0, c)$ for a sufficiently small absolute constant c. Let S be a set of n points in $\mathbb{R}^d$ with (unknown) mean μ. Further assume that the uniform distribution on S has (t, M_t)-certifiably bounded moments for t being a power of 2. Let T be the version of the dataset S after introducing ϵ-fraction of global outliers and ρ-strong local corruptions (as per Global+StrongLC Model). Then, there exists an algorithm that takes as input T, ρ, ϵ, M_t, and t, runs in time $\mathrm{poly}(n^t, d^{t^2})$, and returns $\widehat{\mu}$ such that with probability at least 0.9, it holds $\|\widehat{\mu} - \mu\|_2 \lesssim M_t \epsilon^{1-1/t} + \rho$.*

Let $S = \{x_1, \dots, x_n\}$ be the original dataset of inliers. Let $S' = \{x'_1, \dots, x'_n\}$ with $x'_i = x_i + z_i$ such that $\forall v \in \mathcal{S}^{d-1}$: $\mathbf{E}_{i\sim[n]}[|\langle v, z_i\rangle|] \leq \rho$ be the dataset after the local corruptions (we use the notation $\mathbf{E}_{i\sim[n]}$ to denote taking the average over $i \in [n]$, for example, $\mathbf{E}_{i\sim[n]}[x_i] = \frac{1}{n}\sum_{i\in[n]} x_i$). Finally, let $T = \{y_1, \dots, y_n\}$ be the final dataset after the global corruptions, i.e., T is such that for all but ϵn of the points we have $x'_i = y_i$. Let $\mathcal{I} \subset [n]$ denote the set of indices such that $x'_i = y_i$.

The algorithm is the following: First, it finds a pseudoexpectation $\widetilde{\mathbf{E}}$ over (i) d-dimensional variables $(\widetilde{y}_i)_{i=1}^n, (\widetilde{x}_i)_{i=1}^n, (\widetilde{z}_i)_{i=1}^n, \widetilde{\mu}$, (ii) scalar variables $(\widetilde{w}_i)_{i=1}^n$, and (iii) appropriate auxiliary variables,[19] under the constraints that $\widetilde{\mathbf{E}}$ satisfies the following set of polynomial (in)equalities for a large enough absolute constant C:

(I.i) For all $i \in [n]$: $\widetilde{w}_i^2 = \widetilde{w}_i$.

(I.ii) For all $i \in [n]$: $\widetilde{w}_i \widetilde{y}_i = \widetilde{w}_i y_i$.

(I.iii) $\sum_{i=1}^n \widetilde{w}_i \geq (1 - 2\epsilon)n$.

(I.iv) For all $i \in [n]$: $\widetilde{y}_i = \widetilde{x}_i + \widetilde{z}_i$.

(I.v) $\widetilde{\mu} = \frac{1}{n}\sum_{i=1}^n \widetilde{y}_i$.

(I.vi) There exists an SoS proof in the variable v of the inequality $\mathbf{E}_{i\sim[n]}[\langle v, \widetilde{x}_i - \widetilde{\mu}\rangle^t] \leq C^t \left(M_t^t + \rho^t\right)$ under the constraints $\|v\|_2^2 = 1$.

(I.vii) There exists an SoS proof in the variable v of the inequality $\mathbf{E}_{i\sim[n]}[\langle v, \widetilde{z}_i\rangle^2] \leq C\rho^2/\epsilon$ under the constraints $\|v\|_2^2 = 1$.

Finally, the algorithm outputs $\widehat{\mu} = \widetilde{\mathbf{E}}[\widetilde{\mu}]$.

I.1. Proof of Theorem 40

If $z_i := x'_i - x_i$ denote the local perturbations, then by Proposition 8, we know that there exists a subset of indices $\mathcal{I}' \subset [n]$ with $|\mathcal{I}'| \geq (1 - \epsilon)n$ such that $\mathbf{E}_{i\sim\mathcal{I}'}[\langle v, z_i\rangle^2] \lesssim \rho^2/\epsilon$. Without loss of generality, we can treat the remaining points $i \in [n] \setminus \mathcal{I}'$ as outliers. This is why we use $1 - 2\epsilon$ in the right hand side of Constraint (I.iii). Thus, throughout this proof, we assume that $\mathbf{E}_{i\sim[n]}[\langle v, z_i\rangle^2] \lesssim \rho^2/\epsilon$ and that we have 2ϵ of outliers (i.e., the set $\mathcal{I}$ of indices $i \in [n]$ with $x'_i = y_i$ has size $|\mathcal{I}| \geq (1 - 2\epsilon)n$).

[19] These are needed for encoding the constraints Constraints (I.vi) and (I.vii). We refer the reader to Hopkins and Li (2018); Kothari and Steurer (2017) for further details on how to encode these constraints using auxiliary variables.

Satisfiability We first argue that the system of polynomial inequalities above is satisfiable. Recall the notation for $S, x_i, x_i', z_i, y_i, T$ provided after the statement of Theorem 40. Let $\mu = \frac{1}{n}\sum_{i\in[n]} x_i$. To show satisfiability, we make the following choice of the variables: $\widetilde{x}_i = x_i$ for $i \in \mathcal{I}$ and $\widetilde{x}_i = \mu$ otherwise, $\widetilde{y}_i = y_i$ for $i \in \mathcal{I}$ and $y_i = \mu$ otherwise, $\widetilde{w}_i = \mathbb{1}_{i\in\mathcal{I}}$, and we choose $\widetilde{z}_i = \widetilde{y}_i - \widetilde{x}_i$ (recall that $\mathcal{I}$ is the set of indices i such that $x_i' = y_i$).

Under these choices, Constraints (I.ii) and (I.iv) are satisfied trivially. Moreover, the $\widetilde{w}_i$'s satisfy the Constraints (I.i) and (I.iii) because $|\mathcal{I}| \geq (1-2\epsilon)n$.

We now argue that the Constraint (I.vi) is also satisfiable. First, define $\widetilde{\mu} = \frac{1}{n}\sum_{i\in[n]} \widetilde{y}_i$ and $\widetilde{\mu}' = \frac{1}{n}\sum_{i\in[n]} \widetilde{x}_i$, and observe that under the above choices of $\widetilde{y}_i$ and $\widetilde{x}_i$, we see that $\|\widetilde{\mu} - \widetilde{\mu}'\|_2 \lesssim \rho$; this is because $\|\widetilde{\mu} - \widetilde{\mu}'\|_2 \leq \|\sum_{i\in[n]} z_i/n\| \leq \rho$. Moreover, by SoS Cauchy-Schwarz inequality, there exists an $O(t)$-degree SoS proof of the following inequality in the variable v under the constraint $\|v\|_2^2 = 1$: $\langle v, \widetilde{\mu}' - \widetilde{\mu}\rangle^t \lesssim \|\widetilde{\mu}' - \widetilde{\mu}\|_2^t \leq \rho^t$. Applying the SoS triangle inequality Fact 37, we obtain that the following inequality has an $O(t)$-degree sum of squares proof in the variable v:

$$\mathop{\mathbf{E}}_{i\sim[n]}[\langle v, \widetilde{x}_i - \widetilde{\mu}\rangle^t] \leq 2^t \mathop{\mathbf{E}}_{i\sim[n]}[\langle v, \widetilde{x}_i - \widetilde{\mu}'\rangle^t] + 2^t\langle v, \widetilde{\mu}' - \widetilde{\mu}\rangle^t \lesssim 2^t(M_t^t + \rho^t)\ .$$

Thus, Constraint (I.vi) is satisfiable.

By Proposition 8, $\widetilde{z}_i$ have bounded covariance, which is equivalent to a degree two polynomial inequality in the variable v, and since it is a degree-two polynomial, it also has a sum of squares proof in the variable v, satisfying Constraint (I.vii).[20] Therefore, all the constraints in our program are satisfied by this construction.

Correctness Fix a direction $v \in \mathcal{S}^{d-1}$. Let $\mu = \frac{1}{n}\sum_{i\in[n]} x_i$. We will show that $\langle \widehat{\mu} - \mu, v\rangle = \widetilde{\mathbf{E}}[\langle \widetilde{\mu} - \mu, v\rangle] \leq \tau$ for $\tau = O(M_t\epsilon^{1-1/t})$ for any pseudoexpectation $\widetilde{\mathbf{E}}$ satisfying the program Constraints (I.i) to (I.vii). By duality between pseudoexpectations and SoS proofs, it suffices to show that there is an SoS proof of $\langle \widetilde{\mu} - \mu, v\rangle \leq \tau$ under the polynomial constraints above.

Let $r_i = \mathbb{1}_{i\in\mathcal{I}}$ be the locally corrupted inliers and $W_i = \widetilde{w}_i r_i$ be the variables corresponding to the surviving inliers "selected" by the program. Let $W_i' = (1 - W_i)$. Then there is an SoS proof of $(W_i')^2 = W_i'$ and $\sum_i W_i' \leq 3\epsilon n$ (with proof similar to Claim 4.3 in Diakonikolas et al. (2022a)). Therefore, under the constraints above we have the following (recall that we use the notation $\mathbf{E}_{i\sim[n]}[x_i]$ to denote the average $\frac{1}{n}\sum_{i\in[n]} x_i$):

$$\begin{aligned}
\langle v, \widetilde{\mu} - \mu\rangle^{2t} &= \left(\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v, \widetilde{y}_i - x_i\rangle\right]\right)^{2t} = \left(\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v, \widetilde{y}_i - x_i'\rangle\right] + \mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v, z_i\rangle\right]\right)^{2t} \\
&\leq 2^{2t}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v, \widetilde{y}_i - x_i'\rangle\right]\right)^{2t} + 2^{2t}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v, z_i\rangle\right]\right)^{2t} \\
&= 2^{2t}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[W_i'\langle v, \widetilde{y}_i - x_i'\rangle\right]\right)^{2t} + 2^{2t}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v, z_i\rangle\right]\right)^{2t}\ .
\end{aligned}$$

where we used the SoS triangle inequality (Fact 37) and that $\widetilde{y}_i W_i = x_i' W_i$. The last term, which does not depend on the program variables, is at most $(2\rho)^{2t}$ by assumption. We now focus on the

[20] Formally, we can replace this constraint by an equivalent constraint, $\mathbf{E}_{i\sim[n]}\, \widetilde{z}_i\widetilde{z}_i^\top = (\rho^2/\epsilon)\mathbf{I} - \mathbf{B}\mathbf{B}^\top$, for some auxiliary matrix variable $\mathbf{B} \in \mathbb{R}^{d\times d}$.

first term. By using Constraint (I.iv) and another application of the SoS triangle inequality:

$$\begin{aligned}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[W_i'\langle v,\widetilde{y}_i-x_i'\rangle\right]\right)^{2t} &= \left(\mathop{\mathbf{E}}_{i\sim[n]}\left[W_i'\langle v,\widetilde{x}_i+\widetilde{z}_i-x_i'\rangle\right]\right)^{2t} \\ &\le 2^{2t}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[W_i'\langle v,\widetilde{x}_i-x_i\rangle\right]\right)^{2t} + 2^{2t}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[W_i'\langle v,\widetilde{z}_i-z_i\rangle\right]\right)^{2t} \;.\end{aligned} \tag{28}$$

We shall use different assumptions on $\widetilde{x}_i$ and $\widetilde{z}_i$ to handle these terms differently. For the first term, we use the SoS Hölder inequality (Fact 38) and the constraints that $W_i'^2 = W_i$ with $\mathbf{E}_{i\sim[n]} W_i' \le 3\epsilon$ (within the SoS proof system) to get the following:

$$\begin{aligned}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[W_i'\langle v,\widetilde{x}_i-x_i\rangle\right]\right)^{2t} &\le \left(\mathop{\mathbf{E}}_{i\in[n]}[W_i']\right)^{2t-2}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v,\widetilde{x}_i-x_i\rangle^t\right]\right)^2 \\ &\le (3\epsilon)^{2t-2}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v,\widetilde{x}_i-x_i\rangle^t\right]\right)^2 \;.\end{aligned} \tag{29}$$

To control the right hand side above, we further have the following inequalities (in the SoS proof system): Starting with the SoS triangle inequality, we obtain

$$\begin{aligned}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v,\widetilde{x}_i-x_i\rangle^t\right]\right)^2 &\lesssim 2^t\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v,\widetilde{x}_i-\widetilde{\mu}\rangle^t\right]^2 + \langle v,\widetilde{\mu}-\mu\rangle^{2t} + \mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v,x_i-\mu\rangle^t\right]^2\right) \\ &\lesssim 2^t\left(M_t^{2t}+\rho^{2t}+\langle v,\widetilde{\mu}-\mu\rangle^{2t}+M_t^{2t}\right) \;,\end{aligned} \tag{30}$$

where we used Constraint (I.vi) and the moment assumption on the inliers. Combining (29) and (30) we get that $\mathbf{E}_{i\sim[n]}\left[(W_i'\langle v,\widetilde{x}_i-x_i\rangle)^{2t}\right]^2 \lesssim (C\epsilon)^{2t-2}\langle v,\widetilde{\mu}-\mu\rangle^{2t} + (C\epsilon)^{2t}M^{2t} + (C\rho)^{2t}$.

We now move to the second term in (28). We apply the SoS Hölder inequality to get

$$\begin{aligned}\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[W_i'\langle v,\widetilde{z}_i-z_i\rangle\right]\right)^2 &\le \left(\mathop{\mathbf{E}}_{i\sim[n]}[W_i']\right)\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v,\widetilde{z}_i-z_i\rangle^2\right] \\ &\lesssim \epsilon\left(\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v,z_i\rangle^2\right]+\mathop{\mathbf{E}}_{i\sim[n]}\left[\langle v,\widetilde{z}_i\rangle^2\right]\right) \\ &\lesssim \epsilon(\rho^2/\epsilon+\rho^2/\epsilon) = O(\rho^2) \;.\end{aligned}$$

where the last line uses Constraint (I.vii) and our assumption for the inliers. Combining everything, we have shown an SoS proof of

$$\langle v,\widetilde{\mu}-\mu\rangle^{2t} \lesssim (C\epsilon)^{2t-2}\langle v,\widetilde{\mu}-\mu\rangle^{2t} + (C\epsilon)^{2t-2}M_t^{2t} + (C\rho)^{2t} \;.$$

for some constant absolute C. By solving for $\langle v,\widetilde{\mu}-\mu\rangle^{2t}$, and using that $\epsilon < c$ for a sufficiently small constant, we get $\langle v,\widetilde{\mu}-\mu\rangle^{2t} \lesssim (C\epsilon)^{2t-2}M_t^{2t} + (C\rho)^{2t}$. Finally, by SoS Hölder inequality, we have that $\langle v,\widetilde{\mu}-\mu\rangle \lesssim (C\epsilon)^{1-1/t}M_t + \rho$.

Appendix J. Auxiliary facts

Lemma 41 *Let $y_1,\dots,y_n$ be n vectors in $\mathbb{R}^d$ and a $w \in \Delta_{n,\epsilon}$. Let $\mu \in \mathbb{R}^d$ be fixed and define $\mu_w := \sum_{i=1}^n w_i y_i$ and $\overline{\mathbf{\Sigma}}_w := \sum_{i=1}^n w_i(y_i-\mu)(y_i-\mu)^\top$. Then there exists a set $\mathcal{I} \subset [n]$ satisfying (i) $|\mathcal{I}| \ge (1-2\epsilon)n$ and (ii) the set $S := \{y_i\}_{i\in\mathcal{I}}$ satisfying* $\max_{\mathbf{V}\in\mathcal{V}_k}\langle\mathbf{V},\overline{\mathbf{\Sigma}}_S\rangle \lesssim \max_{\mathbf{V}\in\mathcal{V}_k}\langle\mathbf{V},\overline{\mathbf{\Sigma}}_w\rangle$.

Proof Without loss of generality, let $w_1 \geq w_2 \geq \cdots \geq w_n$. Define the set $\mathcal{I}$ to be the set $[(1-2\epsilon)n]$. Then for each $i \in \mathcal{I}$, we have $w_i \gtrsim \frac{1}{n}$. Let $S = (y_i)_{i\in\mathcal{I}}$. Since $|S| \geq (1-2\epsilon)n$, it can be shown that $\frac{1}{|S|} \lesssim w_i$ for all $i \in \mathcal{I}$ (see Lemma D.2 in Diakonikolas et al. (2020)). Defining $z_i := y_i - \mu$, for any $\mathbf{V} \in \mathcal{V}_k$, we have

$$\langle \overline{\mathbf{\Sigma}}_S, \mathbf{V} \rangle = \frac{1}{|S|} \sum_{i\in\mathcal{I}} \|z_i\|_{\mathbf{V}}^2 \lesssim \sum_{i\in\mathcal{I}} w_i \|z_i\|_{\mathbf{V}}^2 \lesssim \sum_{i\in[n]} w_i \|z_i\|_{\mathbf{V}}^2 = \langle \overline{\mathbf{\Sigma}}_w, \mathbf{V} \rangle \, .$$

■