Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers*

Jacob Steinhardt[†1], **Moses Charikar**[‡2], **and Gregory Valiant**[§3]

1 **Stanford University, Stanford, USA**
`jsteinha@stanford.edu`
2 **Stanford University, Stanford, USA**
`moses@stanford.edu`
3 **Stanford University, Stanford, USA**
`valiant@stanford.edu`

Abstract

We introduce a criterion, *resilience*, which allows properties of a dataset (such as its mean or best low rank approximation) to be robustly computed, even in the presence of a large fraction of arbitrary additional data. Resilience is a weaker condition than most other properties considered so far in the literature, and yet enables robust estimation in a broader variety of settings. We provide new information-theoretic results on robust distribution learning, robust estimation of stochastic block models, and robust mean estimation under bounded kth moments. We also provide new algorithmic results on robust distribution learning, as well as robust mean estimation in ℓ_p-norms. Among our proof techniques is a method for pruning a high-dimensional distribution with bounded 1st moments to a stable "core" with bounded 2nd moments, which may be of independent interest.

1998 ACM Subject Classification G.3 Probability and Statistics, Robust Estimation

Keywords and phrases robust learning, outliers, stochastic block models, p-norm estimation

Digital Object Identifier 10.4230/LIPIcs.ITCS.2018.45

1 Introduction

What are the fundamental properties that allow one to robustly learn from a dataset, even if some fraction of that dataset consists of arbitrarily corrupted data? While much work has been done in the setting of noisy data, or for restricted families of outliers, it is only recently that provable algorithms for learning in the presence of a large fraction of arbitrary (and potentially adversarial) data have been formulated in high-dimensional settings [14, 25, 5, 16, 24, 3]. In this work, we formulate a conceptually simple criterion that a dataset can satisfy–*resilience*–which guarantees that properties such as the mean of that dataset can be estimated even if a large fraction of additional arbitrary data is inserted.

To illustrate our setting, consider the following game between Alice (the adversary) and Bob. First, a set $S \subseteq \mathbb{R}^d$ of $(1-\epsilon)n$ points is given to Alice. Alice then adds ϵn additional points to S to create a new set $\tilde{S}$, and passes $\tilde{S}$ to Bob. Bob wishes to output a parameter

* A full version of the paper is available at `https://arxiv.org/abs/1703.04940`
† JS was supported by a Fannie & John Hertz Foundation Fellowship, an NSF Graduate Research Fellowship, and a Future of Life Instute grant.
‡ MC was supported by NSF grants CCF-1617577, CCF-1302518 and a Simons Investigator Award.
§ GV was supported by NSF CAREER Award CCF-1351108 and a Sloan Research Fellowship.

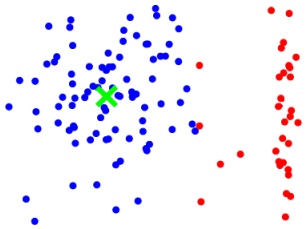

Figure 1 Illustration of the robust mean estimation setting. First, a set of points (blue) is given to Alice, who adds an ϵ fraction of adversarially chosen points (red). Bob's goal is to output the mean of the original set (indicated in green).

$\hat{\mu}$ that is as close as possible to the mean μ of the points in the original set S, with error measured according to some norm $\|\hat{\mu} - \mu\|$. The question is: how well can Bob do, assuming that Alice is an adversary with knowledge of Bob's algorithm?

The above game models mean estimation in the presence of arbitrary outliers; one can easily consider other problems as well (e.g. regression) but we focus on mean estimation here.

With no asumptions on S, Bob will clearly incur arbitrarily large error in the worst case—Alice can add points arbitrarily far away from the true mean μ, and Bob has no way of telling whether those points actually belong to S or were added by Alice. A first pass assumption is to suppose that S has diameter at most ρ; then by discarding points that are very far away from most other points, Bob can obtain error $\mathcal{O}(\epsilon\rho)$. However, in most high-dimensional settings, the diameter ρ grows polynomially with the dimension d (e.g. the d-dimensional hypercube has ℓ_2-diameter $\Theta(\sqrt{d})$). Subtler criteria are therefore needed to obtain dimension-independent bounds in most settings of interest.

Recently, [5] showed that Bob can incur ℓ_2-error $\mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)})$ when the points in S are drawn from a d-dimensional Gaussian, while [16] concurrently showed that Bob can incur ℓ_2-error $\mathcal{O}(\sqrt{\epsilon \log(d)})$ if the points in S are drawn from a distribution with bounded 4th moments. Since then, a considerable amount of additional work has studied high-dimensional estimation in the presence of adversaries, which we discuss in detail below. However, in general, both Bob's strategy and its analysis tend to be quite complex, and specialized to particular distributional assumptions. This raises the question—is it possible to formulate a general and simple-to-understand criterion for the set S under which Bob has a (possibly inefficient) strategy for incurring small error?

In this paper, we provide such a criterion; we identify an assumption–*resilience*–on the set S, under which Bob has a straightforward exponential-time algorithm for estimating μ accurately. This yields new information-theoretic bounds for a number of robust learning problems, including robust learning of stochastic block models, of discrete distributions, and of distributions with bounded kth moments. We also identify additional assumptions under which Bob has an efficient (polynomial-time) strategy for estimating μ, which yields an efficient algorithm for robust learning of discrete distributions, as well as for robust mean estimation in ℓ_p-norms.

The resilience condition is essentially that the mean of every large subset of S must be close to the mean of all of S. More formally, for a norm $\|\cdot\|$, our criterion is as follows:

▶ **Definition 1** (Resilience)**.** A set of points $\{x_i\}_{i\in S}$ lying in $\mathbb{R}^d$ is (σ, ϵ)-*resilient* in a norm $\|\cdot\|$ around a point μ if, for all subsets $T \subseteq S$ of size at least $(1-\epsilon)|S|$, $\left\|\frac{1}{|T|}\sum_{i\in T}(x_i - \mu)\right\| \leq \sigma$.

More generally, a distribution p is said to be (σ, ϵ)-resilient if $\|\mathbb{E}[x - \mu \mid E]\| \leq \sigma$ for every event E of probability at least $1 - \epsilon$.

In the definition above, μ need not equal the mean of S; this distinction is useful in statistical settings where the sample mean of a finite set of points differs slightly from the true mean. However, resilience implies that μ differs from the mean of S by at most σ.

Importantly, Definition 1 is satisfied with high probability by a finite sample in many settings. For instance, samples from a distribution with kth moments bounded by σ will be $(\mathcal{O}(\sigma\epsilon^{1-1/k}), \epsilon)$-resilient in the ℓ_2-norm with high probability. Resilience also holds with high probability under many other natural distributional assumptions, discussed in more detail in Sections 1.1 and 6.

Assuming that the original set S is (σ, ϵ)-resilient, Bob's strategy is actually quite simple—find *any* large (σ, ϵ)-resilient subset S' of the corrupted set $\tilde{S}$, and output the mean of S'. By pigeonhole, S' and S have large intersection, and hence by resilience must have similar means. We establish this formally in Section 2.

Pleasingly, resilience reduces the question of whether Bob can win the game to a purely algorithmic question—that of finding any large resilient set. Rather than wondering whether it is even information-theoretically possible to estimate μ, we can instead focus on efficiently finding resilient subsets of $\tilde{S}$.

We provide one such algorithm in Section 4, assuming that the norm $\|\cdot\|$ is *strongly convex* and that we can approximately solve a certain generalized eigenvalue problem in the dual norm. When specialized to the ℓ_1-norm, our general algorithm yields an efficient procedure for robust learning of discrete distributions.

In the remainder of this section, we will outline our main results, starting with information-theoretic results and then moving on to algorithmic results. In Section 1.1, we show that resilience is indeed information-theoretically sufficient for robust mean estimation. In Section 1.2, we then provide finite-sample bounds showing that resilience holds with high probability for i.i.d. samples from a distribution.

In Section 1.3, we turn our attention to algorithmic bounds. We identity a property–bounded variance in the dual norm–under which efficient algorithms exist. We then show that, as long as the norm is strongly convex, every resilient set has a large subset with bounded variance, thus enabling efficient algorithms. This connection between resilience and bounded variance is the most technically non-trivial component of our results, and may be of independent interest.

Both our information-theoretic and algorithmic bounds yield new results in concrete settings, which we discuss in the corresponding subsections. In Section 1.4, we also discuss an extension of resilience to low-rank matrix approximation, which enables us to derive new bounds in that setting as well. In Section 1.5 we outline the rest of the paper and point to technical highlights, and in Section 1.6 we discuss related work.

1.1 Information-Theoretic Sufficiency

First, we show that resilience is indeed information-theoretically sufficient for robust recovery of the mean μ. Let $\sigma_*(\epsilon)$ denote the smallest σ such that S is (σ, ϵ)-resilient.

▶ **Proposition 2.** *Suppose that $\tilde{S} = \{x_1, \ldots, x_n\}$ contains a set S of size $(1-\epsilon)n$ that is resilient around μ (where S and μ are both unknown). Then if $\epsilon < \frac{1}{2}$, it is possible to recover a $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \leq 2\sigma_*(\frac{\epsilon}{1-\epsilon})$.*

More generally, if $|S| \geq \alpha n$ (even if $\alpha < \frac{1}{2}$), it is possible to output a (random) $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \leq \frac{16}{\alpha}\sigma_*(\frac{\alpha}{4})$ with probability at least $\frac{\alpha}{2}$.

The first part says that robustness to an ϵ fraction of outliers depends on resilience to a $\frac{\epsilon}{1-\epsilon}$ fraction of deletions. Thus, Bob has a good strategy as long as $\sigma_*(\frac{\epsilon}{1-\epsilon})$ is small.

The second part, which is more surprising, says that Bob has a good strategy *even if the majority of $\tilde{S}$ is controlled by Alice.* Here one cannot hope for recovery in the usual sense, because if $\alpha = \frac{1}{2}$ (i.e., Alice controls half the points) then Alice can make $\tilde{S}$ the disjoint union of two identical copies of S (one of which is shifted by a large amount) and Bob has no way of determining which of the two copies is the true S. Nevertheless, in this situation Bob can still identify S (and hence μ) with probability $\frac{1}{2}$; more generally, the second part of Proposition 2 says that if $|S| = \alpha|\tilde{S}|$ then Bob can identify μ with probability at least $\frac{\alpha}{2}$.

The fact that estimation is possible even when $\alpha < \frac{1}{2}$ was first established by [24] in a crowdsourcing setting, and later by [3] in a number of settings including mean estimation. Apart from being interesting due to its unexpectedness, estimation in this regime has immediate implications for robust estimation of mixtures of distributions (by considering each mixture component in turn as the "good" set S) or of planted substructures in random graphs. We refer the reader to [3] for a full elaboration of this point.

The proof of Proposition 2, given in detail in Section 2, is a pigeonhole argument. For the $\epsilon < \frac{1}{2}$ case, we simply search for any large resilient set S' and output its mean; then S and S' must have large overlap, and by resilience their means must both be close to the mean of their intersection, and hence to each other.

For the general case where $|S| = \alpha|\tilde{S}|$ (possibly with $\alpha < \frac{1}{2}$), a similar pigeonhole argument applies but we now need to consider a covering of $\tilde{S}$ by $\frac{2}{\alpha}$ approximately disjoint sets $S'_1, \ldots, S'_{2/\alpha}$. We can show that the true set S must overlap at least one of these sets by a decent amount, and so outputting the mean of one of these sets at random gives a good approximation to the mean of S with probability $\frac{\alpha}{2}$.

1.2 Finite-Sample Concentration

While Proposition 2 provides a deterministic condition under which robust mean estimation is possible, we would also like a way of checking that resilience holds with high probability given samples $x_1, \ldots, x_n$ from a distribution p. First, we provide an alternate characterization of resilience which says that a distribution is resilient if it has *thin tails* in every direction:

▶ **Lemma 3.** *Given a norm $\|\cdot\|$, define the dual norm $\|v\|_* = \sup_{\|x\| \le 1} \langle v, x \rangle$. For a fixed vector v, let $\tau_\epsilon(v)$ denote the ϵ-quantile of $\langle x - \mu, v \rangle$: $\mathbb{P}_{x \sim p}[\langle x - \mu, v \rangle \ge \tau_\epsilon(v)] = \epsilon$. Then, p is (σ, ϵ)-resilient around its mean μ if and only if*

$$\mathbb{E}_p[\langle x - \mu, v \rangle \mid \langle x - \mu, v \rangle \ge \tau_\epsilon(v)] \le \frac{1 - \epsilon}{\epsilon} \sigma \text{ whenever } \|v\|_* \le 1. \tag{1}$$

In other words, if we project onto any unit vector v in the dual norm, the ϵ-tail of $x - \mu$ must have mean at most $\frac{1-\epsilon}{\epsilon}\sigma$. Thus, for instance, a distribution with variance at most σ_0^2 along every unit vector would have $\sigma = \mathcal{O}(\sigma_0\sqrt{\epsilon})$. Note that Lemma 3 requires μ to be the mean, rather than an arbitrary vector as before.

We next provide a meta-result establishing that resilience of a population distribution p very generically transfers to a finite set of samples from that distribution. The number of samples necessary depends on two quantities B and $\log M$ that will be defined in detail later; for now we note that they are ways of measuring the effective dimension of the space.

▶ **Proposition 4.** *Suppose that a distribution p is (σ, ϵ)-resilient around its mean μ with $\epsilon < \frac{1}{2}$. Let B be such that $\mathbb{P}[\|x - \mu\| \ge B] \le \epsilon/2$. Also let M be the covering number of the unit ball in the dual norm $\|\cdot\|_*$.*

Then, given n samples $x_1, \ldots, x_n \sim p$, with probability $1 - \delta - \exp(-\epsilon n/6)$ there is a subset T of $(1-\epsilon)n$ of the x_i that is (σ', ϵ)-resilient with $\sigma' = \mathcal{O}\Big(\sigma \cdot \Big(1 + \sqrt{\frac{\log(M/\delta)}{\epsilon^2 n} + \frac{(B/\sigma)\log(M/\delta)}{n}}\Big)\Big)$.

Note that Proposition 4 only guarantees resilience on a $(1-\epsilon)n$-element subset of the x_i, rather than all of $x_1, \ldots, x_n$. From the perspective of robust estimation, this is sufficient, as we can simply regard the remaining ϵn points as part of the "bad" points controlled by Alice. This weaker requirement seems to be actually necessary to achieve Proposition 4, and was also exploited in [3] to yield improved bounds for a graph partitioning problem. There has been a great deal of recent interest in showing how to "prune" samples to achieve faster rates in random matrix settings citeguedon2014community,le2015concentration,rebrova2015coverings,rebrova2016norms, and we think the general investigation of such pruning results is likely to be fruitful.

We remark that the sample complexity in Proposition 4 is suboptimal in many cases, requiring roughly $d^{1.5}$ samples when d samples would suffice. At the end of the next subsection we discuss a tighter but more specialized bound based on spectral graph sparsification.

Applications. Propositions 2 and 4 together give us a powerful tool for deriving information-theoretic robust recovery results: one needs simply establish resilience for the population distribution p, then use Proposition 4 to obtain finite sample bounds and Proposition 2 to obtain robust recovery guarantees. We do this in three illustrative settings: ℓ_2 mean estimation, learning discrete distributions, and stochastic block models. We outline the results below; formal statements and proofs are deferred to the full version of the paper.

Mean estimation in ℓ_2-norm. Suppose that a distribution on $\mathbb{R}^d$ has bounded kth moments: $\mathbb{E}_{x\sim p}[|\langle x-\mu, v\rangle|^k]^{1/k} \le \sigma\|v\|_2$ for all v for some $k \ge 2$. Then p is $(\mathcal{O}(\sigma\epsilon^{1-1/k}), \epsilon)$-resilient in the ℓ_2-norm. Propositions 4 and 2 then imply that, given $n \ge \frac{d^{1.5}}{\epsilon} + \frac{d}{\epsilon^2}$ samples from p, and an ϵ-fraction of corruptions, it is possible to recover the mean to ℓ_2-error $\mathcal{O}(\sigma\epsilon^{1-1/k})$. Moreover, if only an α-fraction of points are good, the mean can be recovered to error $\mathcal{O}(\sigma\alpha^{-1/k})$ with probability $\Omega(\alpha)$.

The $d^{1.5}/\epsilon$ term in the sample complexity is likely loose, and we believe the true dependence on d is at most $d\log(d)$. This looseness comes from Proposition 4, which uses a naïve covering argument and could potentially be improved with more sophisticated tools. Nevertheless, it is interesting that resilience holds long before the empirical kth moments concentrate, which would require $d^{k/2}$ samples.

Distribution learning. Suppose that we are given k-tuples of independent samples from a discrete distribution: $p = \pi^k$, where π is a distribution on $\{1, \ldots, m\}$. By taking the empirical average of the k samples from π, we can treat a sample from p as an element in the m-dimensional simplex Δ_m. This distribution turns out to be resilient in the ℓ_1-norm with $\sigma = \mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)/k})$, which allows us to estimate p in the ℓ_1-norm (i.e., total variation norm) and recover $\hat{\pi}$ such that $\|\hat{\pi} - \pi\|_{TV} = \mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)/k})$. This reveals a pleasing "error correction" property: if we are given k samples at a time, either all or none of which are good, then our error is $\sqrt{k}$ times smaller than if we only observe the samples individually.

Stochastic block models. Finally, we consider the *semi-random stochastic block model* studied in [3]. For a graph on n vertices, this model posits a subset S of αn "good" vertices, which are connected to each other with probability $\frac{a}{n}$ and to the other ("bad") vertices with probability $\frac{b}{n}$ (where $b < a$); the connections among the bad vertices can be arbitrary. The goal is to recover the set S.

We think of each row of the adjacency matrix as a vector in $\{0,1\}^n$, and show that for the good vertices these vectors are resilient in a truncated ℓ_1-norm $\|x\|$, defined as the sum of the

αn largest coordinates of x (in absolute value). In this case, we have $\sigma = \mathcal{O}(\alpha\sqrt{a\log(2/\alpha)})$ (this requires a separate argument from Proposition 4 to get tight bounds). Applying Proposition 2, we find that we are able to recover (with probability $\frac{\alpha}{2}$) a set $\hat{S}$ with

$$\frac{1}{\alpha n}|S \triangle \hat{S}| = \mathcal{O}\left(\frac{a\log(2/\alpha)}{(a-b)^2\alpha^2}\right). \tag{2}$$

In particular, we get non-trivial guarantees as long as $\frac{(a-b)^2}{a} \gg \frac{\log(2/\alpha)}{\alpha^2}$. [3] derive a weaker (but computationally efficient) bound when $\frac{(a-b)^2}{a} \gg \frac{\log(2/\alpha)}{\alpha^3}$, and remark on the similarity to the famous *Kesten-Stigum threshold* $\frac{(a-b)^2}{a} \gg \frac{1}{\alpha^2}$, which is the conjectured threshold for computationally efficient recovery in the classical stochastic block model (see [4] for the conjecture, and [20, 18] for a proof in the two-block case). Our information-theoretic upper bound matches the Kesten-Stigum threshold up to a $\log(2/\alpha)$ factor. We conjecture that this upper bound is tight; some evidence for this is given in [23], which provides a nearly matching information-theoretic lower bound when $a = 1$, $b = \frac{1}{2}$.

1.3 Strong Convexity, Second Moments, and Efficient Algorithms

Most existing algorithmic results on robust mean estimation rely on analyzing the empirical covariance of the data in some way (see, e.g., [16, 5, 1]). In this section we establish connections between bounded covariance and resilience, and show that in a very general sense, bounded covariance is indeed sufficient to enable robust mean estimation.

Given a norm $\|\cdot\|$, we say that a set of points $x_1, \ldots, x_n$ has *variance bounded by* σ_0^2 in that norm if $\frac{1}{n}\sum_{i=1}^n \langle x_i - \mu, v\rangle^2 \le \sigma_0^2\|v\|_*^2$ (recall $\|\cdot\|_*$ denotes the dual norm). Since this implies a tail bound along every direction, it is easy to see (c.f. Lemma 3) that a set with variance bounded by σ_0^2 is $(\mathcal{O}(\sigma_0\sqrt{\epsilon}), \epsilon)$-resilient around its mean for all $\epsilon < \frac{1}{2}$. Therefore, bounded variance implies resilience.

An important result is that the converse is also true, *provided the norm is strongly convex*. We say that a norm $\|\cdot\|$ is γ-strongly convex if $\|x+y\|^2 + \|x-y\|^2 \ge 2(\|x\|^2 + \gamma\|y\|^2)$ for all $x, y \in \mathbb{R}^d$.[1] As an example, the ℓ_p-norm is $(p-1)$-strongly convex for $p \in (1, 2]$. For strongly convex norms, we show that any resilient set has a "core" with bounded variance:

▶ **Theorem 5.** *If S is $(\sigma, \frac{1}{2})$-resilient in a γ-strongly convex norm $\|\cdot\|$, then S contains a set S_0 of size at least $\frac{1}{2}|S|$ with bounded variance: $\frac{1}{|S_0|}\sum_{i\in S_0}\langle x_i - \mu, v\rangle^2 \le \frac{288\sigma^2}{\gamma}\|v\|_*^2$ for all v.*

Using Lemma 3, we can show that $(\sigma, \frac{1}{2})$-resilience is equivalent to having bounded 1st moments in every direction; Theorem 5 can thus be interpreted as saying that any set with bounded 1st moments can be pruned to have bounded 2nd moments.

We found this result quite striking—the fact that Theorem 5 can hold with no dimension-dependent factors is far from obvious. In fact, if we replace 2nd moments with 3rd moments or take a non-strongly-convex norm then the analog of Theorem 5 is false: we incur polynomial factors in the dimension even if S is the standard basis of $\mathbb{R}^d$ (see the full paper for details). The proof of Theorem 5 involves minimax duality and Khintchine's inequality. We can also strengthen Theorem 5 to yield S_0 of size $(1-\epsilon)|S|$. The proofs of both results are given in Section 3 and may be of independent interest.

[1] In the language of Banach space theory, this is also referred to as having bounded co-type.

Algorithmic results. Given points with bounded variance, we establish algorithmic results assuming that one can solve the "generalized eigenvalue" problem $\max_{\|v\|_* \le 1} v^\top A v$ up to some multiplicative accuracy κ. Specifically, we make the following assumption:

▶ **Assumption 6** (κ-Approximability). *There is a convex set $\mathcal{P}$ of PSD matrices such that*

$$\sup_{\|v\|_* \le 1} v^\top A v \le \sup_{M \in \mathcal{P}} \langle A, M \rangle \le \kappa \sup_{\|v\|_* \le 1} v^\top A v \tag{3}$$

for every PSD matrix A. Moreover, it is possible to optimize linear functions over $\mathcal{P}$ in polynomial time.

A result of [21] implies that this is true with $\kappa = \mathcal{O}(1)$ if $\|\cdot\|_*$ is any "quadratically convex" norm, which includes the ℓ_q-norms for $q \in [2, \infty]$. Also, while we do not use it in this paper, one can sometimes exploit weaker versions of Assumption 6 that only require $\sup_{M \in \mathcal{P}} \langle A, M \rangle$ to be small for certain matrices A; see for instance [17], which obtains an algorithm for robust sparse mean estimation even though Assumption 6 (as well as strong convexity) fails to hold in that setting.

Our main algorithmic result is the following:

▶ **Theorem 7.** *Suppose that $x_1, \dots, x_n$ contains a subset S of size $(1-\epsilon)n$ whose variance around its mean μ is bounded by σ_0^2 in the norm $\|\cdot\|$. Also suppose that Assumption 6 holds for the dual norm $\|\cdot\|_*$. Then, if $\epsilon \le \frac{1}{4}$, there is a polynomial-time algorithm whose output satisfies $\|\hat{\mu} - \mu\| = \mathcal{O}\big(\sigma_0 \sqrt{\kappa \epsilon}\big)$.*

If, in addition, $\|\cdot\|$ is γ-strongly convex, then even if S only has size αn there is a polynomial-time algorithm such that $\|\hat{\mu} - \mu\| = \mathcal{O}\big(\frac{\sqrt{\kappa}\sigma_0}{\sqrt{\gamma}\alpha}\big)$ with probability $\Omega(\alpha)$.

This is essentially a more restrictive, but computationally efficient version of Proposition 2. We note that for the ℓ_2-norm, the algorithm can be implemented as an SVD (singular value decomposition) combined with a filtering step; for more general norms, the SVD is replaced with a semidefinite program.

In the small-ϵ regime, Theorem 7 is in line with existing results which typically achieve errors of $\mathcal{O}(\sqrt{\epsilon})$ in specific norms. While several papers achieve stronger rates of $\mathcal{O}(\epsilon^{3/4})$ citelai2016agnostic or $\tilde{\mathcal{O}}(\epsilon)$ [5, 1], these stronger results rely crucially on specific distributional assumptions such as Gaussianity. At the time of writing of this paper, no results obtained rates better than $\mathcal{O}(\sqrt{\epsilon})$ for any general class of distributions (even under strong assumptions such as sub-Gaussianity). After initial publication of this paper, [15] surpassed $\sqrt{\epsilon}$ and obtained rates of $\epsilon^{1-\gamma}$ for any $\gamma > 0$, for distributions satisfying the Poincaré isoperimetric inequality.

In the small-α regime, Theorem 7 generalizes the mean estimation results of [3] to norms beyond the ℓ_2-norm. That paper achieves a better rate of $1/\sqrt{\alpha}$ (vs. the $1/\alpha$ rate given here). It is likely possible to achieve the $1/\sqrt{\alpha}$ rate here as well, but we leave this for future work.

Applications. Because Assumption 6 holds for ℓ_p-norms, we can perform robust estimation in ℓ_p-norms for any $p \in [1, 2]$, as long as the data have bounded variance in the dual ℓ_q-norm (where $\frac{1}{p} + \frac{1}{q} = 1$). This is the first efficient algorithm for performing robust mean estimation in any ℓ_p-norm with $p \neq 2$. The ℓ_1-norm in particular is often a more meaningful metric than the ℓ_2-norm in discrete settings, allowing us to improve on existing results.

Indeed, as in the previous section, suppose we are given k-tuples of samples from a discrete distribution π on $\{1, \dots, m\}$. Applying Theorem 7 with the ℓ_1-norm yields an

algorithm recovering a $\hat{\pi}$ with $\|\hat{\pi} - \pi\|_{TV} = \tilde{\mathcal{O}}(\sqrt{\epsilon/k})$.[2] In contrast, bounds using the ℓ_2-norm would only yield $\|\hat{\pi} - \pi\|_2 = \mathcal{O}(\sqrt{\epsilon\pi_{\max}/k})$, which is substantially weaker when the maximum probability $\pi_{\max}$ is large. Our result has a similar flavor to that of [5] on robustly estimating binary product distributions, for which directly applying ℓ_2 mean estimation was also insufficient. We discuss our bounds in more detail in the full version of the paper.

Finite-sample bound. To get the best sample complexity for the applications above, we provide an additional finite-sample bound focused on showing that a set of points has bounded variance. This is a simple but useful generalization of Proposition B.1 of [3]; it shows that in a very generic sense, given d samples from a distribution on $\mathbb{R}^d$ with bounded population variance, we can find a subset of samples with bounded variance with high probability. It involves pruning the samples in a non-trivial way based on ideas from graph sparsification [2]. The formal statement is given in Section 6.2.

1.4 Low-Rank Recovery

Finally, to illustrate that the idea behind resilience is quite general and not restricted to mean estimation, we also provide results on recovering a rank-k approximation to the data in the presence of arbitrary outliers. Given a set of points $[x_i]_{i\in S}$, let X_S be the matrix whose columns are the x_i. Our goal is to obtain a low-rank matrix P such that the operator norm $\|(I-P)X_S\|_2$ is not much larger than $\sigma_{k+1}(X_S)$, where σ_{k+1} denotes the $k+1$st singular value; we wish to do this even if S is corrupted to a set $\tilde{S}$ by adding arbitrary outliers.

As before, we start by formulating an appropriate resilience criterion:

▶ **Definition 8** (Rank-resilience)**.** A set of points $[x_i]_{i\in S}$ in $\mathbb{R}^d$ is *δ-rank-resilient* if for all subsets T of size at least $(1-\delta)|S|$, we have $\mathrm{col}(X_T) = \mathrm{col}(X_S)$ and $\|X_T^\dagger X_S\|_2 \le 2$, where $\dagger$ is the pseudoinverse and col denotes column space.

Rank-resilience says that the variation in X should be sufficiently spread out: there should not be a direction of variation that is concentrated in only a δ-fraction of the points. Under rank-resilience, we can perform efficient rank-k recovery even in the presence of a δ-fraction of arbitrary data:

▶ **Theorem 9.** *Let $\delta \le \frac{1}{3}$. If a set of n points contains a set S of size $(1-\delta)n$ that is δ-rank-resilient, then it is possible to efficiently recover a matrix P of rank at most $15k$ such that $\|(I-P)X_S\|_2 = \mathcal{O}(\sigma_{k+1}(X_S))$.*

The power of Theorem 9 comes from the fact that the error depends on σ_{k+1} rather than e.g. σ_2, which is what previous results yielded. This distinction is crucial in practice, since most data have a few (but more than one) large singular values followed by many small singular values. Note that in contrast to Theorem 7, Theorem 9 only holds when S is relatively large: at least $(1-\delta)n \ge \frac{2}{3}n$ in size.

1.5 Summary, Technical Highlights, and Roadmap

In summary, we have provided a deterministic condition on a set of points that enables robust mean estimation, and provided finite-sample bounds showing that this condition holds with

[2] The $\tilde{\mathcal{O}}$ notation suppresses log factors in m and ϵ; the dependence on m can likely be removed with a more careful analysis.

high probability in many concrete settings. This yields new results for distribution learning, stochastic block models, mean estimation under bounded moments, and mean estimation in ℓ_p norms. We also provided an extension of our condition that yields results for robust low-rank recovery.

Beyond the results themselves, the following technical aspects of our work may be particularly interesting: The proof of Proposition 2 (establishing that resilience is indeed sufficient for robust estimation), while simple, is a nice pigeonhole argument that we found to be conceptually illuminating.

In addition, the proof of Theorem 5, on pruning resilient sets to obtain sets with bounded variance, exploits strong convexity in a non-trivial way in conjunction with minimax duality; we think it reveals a fairly non-obvious geometric structure in resilient sets, and also shows how the ability to prune points can yield sets with meaningfully stronger properties.

Finally, in the proof of our algorithmic result (Theorem 7), we establish an interesting generalization of the inequality $\sum_{i,j} X_{ij}^2 \leq \text{rank}(X) \cdot \|X\|_2^2$, which holds not just for the ℓ_2-norm but for any strongly convex norm. This is given as Lemma 18.

Roadmap. The rest of the paper is organized as follows. In Section 2, we prove our information-theoretic recovery result for resilient sets (Proposition 2). In Section 3, we prove Theorem 5 establishing that all resilient sets in strongly convex norms contain large subsets with bounded variance; we also prove a more precise version of Theorem 5 in Section 3.1. In Section 4, we prove our algorithmic results, warming up with the ℓ_2-norm (Section 4.1) and then moving to general norms (Section 4.2). In Section 5, we prove our results on rank-k recovery. In Section 6, we present and prove the finite-sample bounds discussed in Section 1.2. Applications of our results are deferred to the full version of the paper.

1.6 Related Work

A number of authors have recently studied robust estimation and learning in high-dimensional settings: [16] study mean and covariance estimation, while [5] focus on estimating Gaussian and binary product distributions, as well as mixtures thereof; note that this implies mean/-covariance estimation of the corresponding distributions. [3] recently showed that robust estimation is possible even when the fraction α of "good" data is less than $\frac{1}{2}$. We refer to these papers for an overview of the broader robust estimation literature; since those papers, a number of additional results have also been published: [6] provide a case study of various robust estimation methods in a genomic setting, [1] study sparse mean estimation, and others have studied problems including regression, Bayes nets, planted clique, and several other settings [8, 9, 7, 10, 12, 19].

Special cases of the resilience criterion are implicit in some of these earlier works; for instance, ℓ_2-resilience appears in equation (9) in [5], and resilience in a sparsity-inducing norm appears in Theorem 4.5 of [17]. However, these conditions typically appear concurrently with other stronger conditions, and the general sufficiency of resilience for information-theoretic recovery appears to be unappreciated (for instance, [17], despite already having implicitly established resilience, proves its information-theoretic results via reduction to a tournament lemma from [5]).

Low rank estimation was studied by [16], but their bounds depend on the maximum eigenvalue $\|\Sigma\|_2$ of the covariance matrix, while our bound provides robust recovery guarantees in terms of lower singular values of Σ. (Some work, such as [5], shows how to estimate all of Σ in e.g. Frobenius norm, but appears to require the samples to be drawn from a Gaussian.)

2 Resilience and Robustness: Information-Theoretic Sufficiency

Recall the definition of resilience: S is (σ,ϵ)-resilient if $\|\frac{1}{|T|}\sum_{i\in T}(x_i-\mu)\|\le\sigma$ whenever $T\subseteq S$ and $|T|\ge(1-\epsilon)|S|$. Here we establish Proposition 2 showing that, if we ignore computational efficiency, resilience leads directly to an algorithm for robust mean estimation.

Proof (Proposition 2). We prove Proposition 2 via a constructive (albeit exponential-time) algorithm. To prove the first part, suppose that the true set S is $(\sigma,\frac{\epsilon}{1-\epsilon})$-resilient around μ, and let S' be any set of size $(1-\epsilon)n$ that is $(\sigma,\frac{\epsilon}{1-\epsilon})$-resilient (around some potentially different vector μ'). We claim that μ' is sufficiently close to μ.

Indeed, let $T=S\cap S'$, which by the pigeonhole principle has size at least $(1-2\epsilon)n=\frac{1-2\epsilon}{1-\epsilon}|S|=(1-\frac{\epsilon}{1-\epsilon})|S|$. Therefore, by the definition of resilience,

$$\left\|\tfrac{1}{|T|}\textstyle\sum_{i\in T}(x_i-\mu)\right\|\le\sigma. \tag{4}$$

But by the same argument, $\|\frac{1}{|T|}\sum_{i\in T}(x_i-\mu')\|\le\sigma$ as well. By the triangle inequality, $\|\mu-\mu'\|\le 2\sigma$, which completes the first part of the proposition.

For the second part, we need the following lemma relating ϵ-resilience to $(1-\epsilon)$-resilience:

▶ **Lemma 10.** *For any $0<\epsilon<1$, a distribution/set is (σ,ϵ)-resilient around its mean μ if and only if it is $(\frac{1-\epsilon}{\epsilon}\sigma,1-\epsilon)$-resilient. Moreover, even if μ is not the mean, the distribution/set is $(\frac{2-\epsilon}{\epsilon}\sigma,1-\epsilon)$-resilient. In other words, if $\|\frac{1}{|T|}\sum_{i\in T}(x_i-\mu)\|\le\sigma$ for all sets T of size at least $(1-\epsilon)n$, then $\|\frac{1}{|T'|}\sum_{i\in T'}(x_i-\mu)\|\le\frac{2-\epsilon}{\epsilon}\sigma$ for all sets T' of size at least ϵn.*

Given Lemma 10, the second part of Proposition 2 is similar to the first part, but requires us to consider multiple resilient sets S_i rather than a single S'. Suppose S is $(\sigma,\frac{\alpha}{4})$-resilient around μ–and thus also $(\frac{8}{\alpha}\sigma,1-\frac{\alpha}{4})$-resilient by Lemma 10–and let $S_1,\dots,S_m$ be a maximal collection of subsets of $[n]$ such that:

1. $|S_j|\ge\frac{\alpha}{2}n$ for all j.
2. S_j is $(\frac{8}{\alpha}\sigma,1-\frac{\alpha}{2})$-resilient around some point μ_j.
3. $S_j\cap S_{j'}=\emptyset$ for all $j\ne j'$.

Clearly $m\le\frac{2}{\alpha}$. We claim that at least one of the μ_j is close to μ. By maximality of the collection $\{S_j\}_{j=1}^m$, it must be that $S_0=S\backslash(S_1\cup\dots\cup S_m)$ cannot be added to the collection. First suppose that $|S_0|\ge\frac{\alpha}{2}n$. Then S_0 is $(\frac{8}{\alpha}\sigma,1-\frac{\alpha}{2})$-resilient (because any subset of $\frac{\alpha}{2}|S_0|$ points in S_0 is a subset of at least $\frac{\alpha}{4}|S|$ points in S). But this contradicts the maximality of $\{S_j\}_{j=1}^m$, so we must have $|S_0|<\frac{\alpha}{2}n$.

Now, this implies that $|S\cap(S_1\cup\dots\cup S_m)|\ge\frac{\alpha}{2}n$, so by pigeonhole we must have $|S\cap S_j|\ge\frac{\alpha}{2}|S_j|$ for some j. Letting $T=S\cap S_j$ as before, we find that $|T|\ge\frac{\alpha}{2}|S_j|\ge\frac{\alpha}{4}|S|$ and hence by resilience of S_j and S we have $\|\mu-\mu_j\|\le 2\cdot(\frac{8}{\alpha}\sigma)=\frac{16}{\alpha}\sigma$. If we output one of the μ_j at random, we are then within the desired distance of μ with probability $\frac{1}{m}\ge\frac{\alpha}{2}$. ◀

3 Powering up Resilience: Finding a Core with Bounded Variance

In this section we prove Theorem 5, which says that for strongly convex norms, every resilient set contains a core with bounded variance. Recall that this is important for enabling algorithmic applications that depend on a bounded variance condition.

First recall the definition of resilience (Definition 1): a set S is (σ,ϵ)-resilient if for every set $T\subseteq S$ of size $(1-\epsilon)|S|$, we have $\|\frac{1}{|T|}\sum_{i\in T}(x_i-\mu)\|\le\sigma$. For $\epsilon=\frac{1}{2}$, we observe that resilience in a norm is equivalent to having bounded first moments in the dual norm:

▶ **Lemma 11.** *Suppose that S is $(\sigma, \frac{1}{2})$-resilient in a norm $\|\cdot\|$, and let $\|\cdot\|_*$ be the dual norm. Then S has 1st moments bounded by 3σ: $\frac{1}{|S|}\sum_{i\in S}|\langle x_i - \mu, v\rangle| \le 3\sigma\|v\|_*$ for all $v \in \mathbb{R}^d$.*

Conversely, if S has 1st moments bounded by σ, it is $(2\sigma, \frac{1}{2})$-resilient.

The proof is routine and can be found in the full paper. Supposing a set has bounded 1st moments, we will show that it has a large core with bounded second moments. This next result is *not* routine:

▶ **Proposition 12.** *Let S be any set with 1st moments bounded by σ. Then if the norm $\|\cdot\|$ is γ-strongly convex, there exists a core S_0 of size at least $\frac{1}{2}|S|$ with variance bounded by $\frac{32\sigma^2}{\gamma}$. That is, $\frac{1}{|S_0|}\sum_{i\in S_0}|\langle x_i - \mu, v\rangle|^2 \le \frac{32\sigma^2}{\gamma}\|v\|_*^2$ for all $v \in \mathbb{R}^d$.*

The assumptions seem necessary: i.e., such a core does not exist when $\|\cdot\|$ is the ℓ_p-norm with $p > 2$ (which is a non-strongly-convex norm), or with bounded 3rd moments for $p = 2$. The proof of Proposition 12 uses minimax duality and Khintchine's inequality [13]. Note that Lemma 11 and Proposition 12 together imply Theorem 5.

Proof (Proposition 12). Without loss of generality take $\mu = 0$ and suppose that $S = [n]$. We can pose the problem of finding a resilient core as an integer program:

$$\min_{c\in\{0,1\}^n, \|c\|_1 \ge \frac{n}{2}} \max_{\|v\|_* \le 1} \frac{1}{n}\sum_{i=1}^n c_i |\langle x_i, v\rangle|^2. \tag{5}$$

Here the variable c_i indicates whether the point i lies in the core S_0. By taking a continuous relaxation and applying a standard duality argument, we obtain the following:

▶ **Lemma 13.** *Suppose that for all m and all vectors $v_1, \dots, v_m$ satisfying $\sum_{j=1}^m \|v_j\|_*^2 \le 1$, we have*

$$\frac{1}{n}\sum_{i=1}^n \sqrt{\sum_{j=1}^m |\langle x_i, v_j\rangle|^2} \le B. \tag{6}$$

Then the value of (5) is at most $8B^2$.

The proof is straightforward and deferred to the full paper. Now, to bound (6), let $s_1, \dots, s_m \in \{-1, +1\}$ be i.i.d. random sign variables. We have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \sqrt{\sum_{j=1}^m |\langle x_i, v_j\rangle|^2} &\overset{(i)}{\le} \mathbb{E}_{s_{1:m}}\left[\frac{\sqrt{2}}{n}\sum_{i=1}^n \left|\sum_{j=1}^m s_j \langle x_i, v_j\rangle\right|\right] && (7)\\
&= \mathbb{E}_{s_{1:m}}\left[\frac{\sqrt{2}}{n}\sum_{i=1}^n \left|\left\langle x_i, \sum_{j=1}^m s_j v_j\right\rangle\right|\right] && (8)\\
&\overset{(ii)}{\le} \mathbb{E}_{s_{1:m}}\left[\sqrt{2}\sigma\left\|\sum_{j=1}^m s_j v_j\right\|_*\right] && (9)\\
&\le \sqrt{2}\sigma\mathbb{E}_{s_{1:m}}\left[\left\|\sum_{j=1}^m s_j v_j\right\|_*^2\right]^{\frac{1}{2}}. && (10)
\end{aligned}$$

Here (i) is Khintchine's inequality [11] and (ii) is the assumed first moment bound. It remains to bound (10). The key is the following inequality asserting that the dual norm $\|\cdot\|_*$ is strongly smooth whenever $\|\cdot\|$ is strongly convex (c.f. Lemma 17 of [22]):

▶ **Lemma 14.** *If $\|\cdot\|$ is γ-strongly convex, then $\|\cdot\|_*$ is $(1/\gamma)$-strongly smooth: $\frac{1}{2}(\|v+w\|_*^2 + \|v-w\|_*^2) \le \|v\|_*^2 + (1/\gamma)\|w\|_*^2$.*

Applying Lemma 14 inductively to $\mathbb{E}_{s_{1:m}}\left[\left\|\sum_{j=1}^m s_j v_j\right\|_*^2\right]$, we obtain

$$\mathbb{E}_{s_{1:m}}\left[\left\|\sum_{j=1}^m s_j v_j\right\|_*^2\right] \le \frac{1}{\gamma}\sum_{j=1}^m \|v_j\|_*^2 \le \frac{1}{\gamma}. \tag{11}$$

Combining with (10), we have the bound $B \le \sigma\sqrt{2/\gamma}$, which yields the desired result. ◀

3.1 Finding Resilient Cores when $\alpha \approx 1$

Lemma 11 together with Proposition 12 show that a $(\sigma, \frac{1}{2})$-resilient set has a core with bounded 2nd moments. One piece of looseness is that Proposition 12 only exploits resilience for $\epsilon = \frac{1}{2}$, and hence is not sensitive to the degree of (σ,ϵ)-resilience as $\epsilon \to 0$. In particular, it only yields a core S_0 of size $\frac{1}{2}|S|$, while we might hope to find a much larger core of size $(1-\epsilon)|S|$ for some small ϵ.

Here we tighten Proposition 12 to make use of finer-grained resilience information. Recall that we let $\sigma_*(\epsilon)$ denote the resilience over sets of size $(1-\epsilon)|S|$. For a given ϵ, our goal is to construct a core S_0 of size $(1-\epsilon)|S|$ with small second moments. The following key quantity will tell us how small the second moments can be:

$$\widetilde{\sigma}_*(\epsilon) \stackrel{\text{def}}{=} \sqrt{\int_{\epsilon/2}^{1/2} u^{-2}\sigma_*(u)^2 du}. \tag{12}$$

The following proposition says that $\widetilde{\sigma}_*$ controls the 2nd moments of S_0:

▶ **Proposition 15.** *Let S be any resilient set in a γ-strongly-convex norm. Then for any $\epsilon \le \frac{1}{2}$, there exists a core S_0 of size $(1-\epsilon)|S|$ with variance bounded by $\mathcal{O}\left(\widetilde{\sigma}_*^2(\epsilon)/\gamma\right)$.*

The proof is similar to Proposition 12, but requires more careful bookkeeping.

To interpret $\widetilde{\sigma}_*$, suppose that $\sigma_*(\epsilon) = \sigma\epsilon^{1-1/r}$ for some $r \in [1,2)$, which roughly corresponds to having bounded rth moments. Then $\widetilde{\sigma}_*^2(\epsilon) = \sigma^2\int_{\epsilon/2}^{1/2} u^{-2/r}du \le \frac{\sigma^2}{2/r-1}\left(\frac{2}{\epsilon}\right)^{2/r-1}$. If $r=1$ then a core of size $(1-\epsilon)|S|$ might require second moments as large as $\frac{\sigma^2}{\epsilon}$; on the other hand, as $r \to 2$ the second moments can be almost as small as σ^2. In general, $\widetilde{\sigma}_*(\epsilon)$ is $\mathcal{O}\left(\sigma\epsilon^{1/2-1/r}\right)$ if $r \in [1,2)$, is $\mathcal{O}\left(\sigma\sqrt{\log(1/\epsilon)}\right)$ if $r=2$, and is $\mathcal{O}(\sigma)$ if $r>2$.

4 Efficient Recovery Algorithms

We now turn our attention to the question of efficient algorithms. The main point of this section is to prove Theorem 7, which yields efficient robust mean estimation for a general class of norms.

4.1 Warm-Up: Recovery in ℓ_2-norm

We first prove a warm-up to Theorem 7 which focuses on the ℓ_2-norm. Our warm-up is:

▶ **Proposition 16.** *Let $x_1,\dots,x_n \in \mathbb{R}^d$, and let S be a subset of size αn with bounded variance in the ℓ_2-norm: $\lambda_{\max}(\frac{1}{|S|}\sum_{i\in S}(x_i-\mu)(x_i-\mu)^\top) \le \sigma^2$, where μ is the mean of S. Then there is an efficient randomized algorithm (Algorithm 1) which with probability $\Omega(\alpha)$ outputs a parameter $\hat{\mu}$ such that $\|\mu-\hat{\mu}\|_2 = \mathcal{O}\left(\frac{\sigma}{\alpha}\right)$. Moreover, if $\alpha = 1-\epsilon \ge \frac{3}{4}$ then $\|\mu-\hat{\mu}\|_2 = \mathcal{O}(\sigma\sqrt{\epsilon})$ with probability 1.*

Algorithm 1 Algorithm for recovering the mean of a set with bounded variance in ℓ_2-norm.

1: Initialize $c_i = 1$ for all $i = 1, \ldots, n$ and $\mathcal{A} = \{1, \ldots, n\}$.
2: Let $Y \in \mathbb{R}^{d \times d}$ and $W \in \mathbb{R}^{\mathcal{A} \times \mathcal{A}}$ be the maximizer/minimizer of the saddle point problem

$$\max_{\substack{Y \succeq 0, \\ \mathrm{tr}(Y) \leq 1}} \min_{\substack{0 \leq W_{ji} \leq \frac{4-\alpha}{\alpha(2+\alpha)n}, \\ \sum_j W_{ji} = 1}} \sum_{i \in \mathcal{A}} c_i (x_i - X_{\mathcal{A}} w_i)^\top Y (x_i - X_{\mathcal{A}} w_i). \tag{14}$$

3: Let $\tau_i^\star = (x_i - X_{\mathcal{A}} w_i)^\top Y (x_i - X_{\mathcal{A}} w_i)$.
4: **if** $\sum_{i \in \mathcal{A}} c_i \tau_i^\star > 4n\sigma^2$ **then**
5: For $i \in \mathcal{A}$, replace c_i with $\left(1 - \frac{\tau_i^\star}{\tau_{\max}}\right) c_i$, where $\tau_{\max} = \max_{i \in \mathcal{A}} \tau_i^\star$.
6: For all i with $c_i < \frac{1}{2}$, remove i from $\mathcal{A}$.
7: Go back to line 2.
8: **end if**
9: Let W_1 be the result of zeroing out all singular values of W that are greater than 0.9.
10: Let $Z = X_{\mathcal{A}} W_0$, where $W_0 = (W - W_1)(I - W_1)^{-1}$.
11: **if** $\mathrm{rank}(Z) = 1$ **then**
12: Output the average of the columns of $X_{\mathcal{A}}$.
13: **else**
14: Output a column of Z at random.
15: **end if**

At the heart of Algorithm 1 is the following optimization problem:

$$\begin{aligned} &\underset{W \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \|X - XW\|_2^2 \\ &\text{subject to} \quad 0 \leq W_{ji} \leq \frac{1}{\alpha n} \ \forall i, j, \quad \sum_j W_{ji} = 1 \ \forall i. \end{aligned} \tag{13}$$

Here $X \in \mathbb{R}^{d \times n}$ is the data matrix $[x_1 \ \cdots \ x_n]$ and $\|X - XW\|_2$ is the operator norm (maximum singular value) of $X - XW$. Note that (13) can be expressed as a semidefinite program; however, it can actually be solved more efficiently than this, via a singular value decomposition (see the full paper for details).

The idea behind (13) is to re-construct each x_i as an average of αn other x_j. Note that by assumption we can always re-construct each element of S using the mean of S, and have small error. Intuitively, any element that cannot be re-constructed well must not lie in S, and can be safely removed. We do a soft form of removal by maintaining weights c_i on the points x_i (initially all 1), and downweighting points with high reconstruction error. We also maintain an active set $\mathcal{A}$ of points with $c_i \geq \frac{1}{2}$.

Informally, Algorithm 1 for estimating μ takes the following form:

1. Solve the optimization problem (13).
2. If the optimum is $\gg \sigma^2 n$, then find the columns of X that are responsible for the optimum being large, and downweight them.
3. Otherwise, if the optimum is $\mathcal{O}(\sigma^2 n)$, then take a low rank approximation W_0 to W, and return a randomly chosen column of XW_0.

The hope in step 3 is that the low rank projection XW_0 will be close to μ for the columns belonging to S. The choice of operator norm is crucial: it means we can actually expect XW to be close to X (on the order of $\sigma\sqrt{n}$). In contrast, the Frobenius norm scales as $\sigma\sqrt{nd}$.

Finally, we note that $\|X - XW\|_2^2$ is equal to

$$\|X - XW\|_2^2 = \max_{Y \succeq 0, \operatorname{tr}(Y) \leq 1} \sum_{i=1}^{n} (x_i - Xw_i)^\top Y (x_i - Xw_i), \tag{15}$$

which is the form we use in Algorithm 1.

Proof (Proposition 16). We show two things: (1) that the outlier removal step removes many more outliers than good points, and (2) that many columns of XW_0 are close to μ.

Outlier removal. To analyze the outlier removal step (step 2 above, or lines 5-6 of Algorithm 1), we make use of the following general lemma:

▶ **Lemma 17.** *For any scalars τ_i and a, suppose that $\sum_{i \in \mathcal{A}} c_i \tau_i \geq 4a$ while $\sum_{i \in S \cap \mathcal{A}} c_i \tau_i \leq \alpha a$. Then the following invariants are preserved by lines 5-6 of Algorithm 1: (i) $\sum_{i \in S}(1 - c_i) \leq \frac{\alpha}{4} \sum_{i=1}^{n}(1 - c_i)$, and (ii) $|S \cap \mathcal{A}| \geq \frac{\alpha(2+\alpha)}{4-\alpha} n$.*

Lemma 17 says that we downweight points within S at least 4 times slower than we do overall (property i), and in particular we never remove too many points from S (property ii). This type of lemma is not new (cf. Lemma 4.5 of [3]) and its proof is deferred to the full paper.

We will show that we can take $a = n\sigma^2$ in Lemma 17, or in other words that $\sum_{i \in S \cap \mathcal{A}} c_i \tau_i^\star \leq \alpha n \sigma^2$. Let $\tau_i(w) = (x_i - X_{\mathcal{A}} w)^\top (x_i - X_{\mathcal{A}} w)$, and note that $\tau_i^\star = \tau_i(w_i) = \min\{\tau_i(w) \mid 0 \leq w_j \leq \frac{1}{\alpha n}, \sum_j w_j = 1\}$. This is because for a fixed Y, each of the w_i are optimized independently.

We can therefore bound $\tau_i^\star$ by substituting any feasible $\hat{w}_i$. We will choose $\hat{W}_{ji} = \frac{\mathbb{I}[j \in S \cap \mathcal{A}]}{|S \cap \mathcal{A}|}$, in which case $X_{\mathcal{A}} \hat{w}_i = \hat{\mu}$, where $\hat{\mu}$ is the average of x_j over $S \cap \mathcal{A}$. Then we have

$$\begin{aligned}
\sum_{i \in S \cap \mathcal{A}} c_i \tau_i^\star &\leq \sum_{i \in S \cap \mathcal{A}} c_i \tau_i(\hat{w}_i) && (16)\\
&\leq \sum_{i \in S \cap \mathcal{A}} c_i (x_i - \hat{\mu})^\top Y (x_i - \hat{\mu}) && (17)\\
&\overset{(i)}{\leq} \sum_{i \in S \cap \mathcal{A}} c_i (x_i - \mu)^\top Y (x_i - \mu) && (18)\\
&\leq \sum_{i \in S} (x_i - \mu)^\top Y (x_i - \mu) \leq \alpha n \sigma^2 \operatorname{tr}(Y) \leq \alpha n \sigma^2 && (19)
\end{aligned}$$

as desired; (i) is because the covariance around the mean ($\hat{\mu}$) is smaller than around any other point (μ).

Analyzing XW_0. By Lemma 17, we will eventually exit the if statement and obtain $Z = X_{\mathcal{A}} W_0$. It therefore remains to analyze Z; we will show in particular that $\|Z_{\mathcal{A} \cap S} - \mu \mathbb{1}^\top\|_F$ is small, where the subscript indicates restricting to the columns in $S \cap \mathcal{A}$. At a high level, it suffices to show that W_0 has low rank (so that Frobenius norm is close to spectral norm) and that XW_0 and X are close in spectral norm (note that X and $\mu \mathbb{1}^\top$ are close by assumption).

To bound $\operatorname{rank}(W_0)$, note that the constraints in (14) imply that $\|W\|_F^2 \leq \frac{4-\alpha}{\alpha(2+\alpha)}$, and so at most $\frac{4-\alpha}{0.81\alpha(2+\alpha)}$ singular values of W can be greater than 0.9. Importantly, at most 1 singular value can be greater than 0.9 if $\alpha \geq \frac{3}{4}$, and at most $\mathcal{O}(\frac{1}{\alpha})$ can be in general. Therefore, $\operatorname{rank}(W_0) \leq \mathcal{O}(\frac{1}{\alpha})$.

Next, we show that $X_{\mathcal{A}}$ and Z are close in operator norm. Indeed, $X_{\mathcal{A}} - Z = X_{\mathcal{A}}(I - W_0) = X_{\mathcal{A}}(I - W)(I - W_1)^{-1}$, hence:

$$\|X_{\mathcal{A}} - Z\|_2 = \|X_{\mathcal{A}}(I - W)(I - W_1)^{-1}\|_2 \tag{20}$$

$$\leq \|X_{\mathcal{A}}(I - W)\|_2 \|(I - W_1)^{-1}\|_2 \tag{21}$$

$$\overset{(i)}{\leq} 10\|X_{\mathcal{A}}(I - W)\|_2 \tag{22}$$

$$\overset{(ii)}{\leq} 10\sqrt{2}\|X_{\mathcal{A}}(I - W)\operatorname{diag}(c_{\mathcal{A}})^{1/2}\|_2 \overset{(iii)}{\leq} 20\sqrt{2n}\sigma. \tag{23}$$

Here (i) is because all singular values of W_1 are less than 0.9, (ii) is because $\operatorname{diag}(c_{\mathcal{A}})^{1/2} \succeq \frac{1}{\sqrt{2}} I$, and (iii) is by the condition in the if statement (line 4 of Algorithm 1), since the sum on line 4 is equal to $\|X_{\mathcal{A}}(I - W)\operatorname{diag}(c_{\mathcal{A}})^{1/2}\|_2^2$.

Combining the previous two observations, we have

$$\sum_{i \in S \cap \mathcal{A}} \|z_i - \mu\|_2^2 \leq (\operatorname{rank}(Z) + 1)\|[z_i - \mu]_{i \in S \cap \mathcal{A}}\|_2^2 \tag{24}$$

$$\leq (\operatorname{rank}(Z) + 1)\left(\|[z_i - x_i]_{i \in S \cap \mathcal{A}}\|_2 + \|[x_i - \mu]_{i \in S \cap \mathcal{A}}\|_2\right)^2 \tag{25}$$

$$\overset{(i)}{\leq} (\operatorname{rank}(Z) + 1)\left(20\sqrt{2n}\sigma + \sqrt{\alpha n}\sigma\right)^2 = \mathcal{O}\left(\tfrac{\sigma^2}{\alpha} n\right). \tag{26}$$

Here (i) uses the preceding bound on $\|X_{\mathcal{A}} - Z\|_2$, together with the 2nd moment bound $\|[x_i - \mu]_{i \in S}\|_2 \leq \sqrt{\alpha n}\sigma$. Note that $\operatorname{rank}(Z) \leq \operatorname{rank}(W_0) = \mathcal{O}(\frac{1}{\alpha})$.

Since $|S \cap \mathcal{A}| = \Omega(\alpha n)$ by Lemma 17, the average value of $\|z_i - \mu\|_2^2$ over $S \cap \mathcal{A}$ is $\mathcal{O}\big(\frac{\sigma^2}{\alpha^2}\big)$, and hence with probability at least $\frac{|S \cap \mathcal{A}|}{2|\mathcal{A}|} = \Omega(\alpha)$, a randomly chosen z_i will be within distance $\mathcal{O}\left(\frac{\sigma}{\alpha}\right)$ of μ, which completes the first part of Proposition 16.

For the second part, when $\alpha = 1 - \epsilon \geq \frac{3}{4}$, recall that we have $\operatorname{rank}(W_0) = 1$, and that $W_0 = (W - W_1)(I - W_1)^{-1}$. One can then verify that $\mathbb{1}^\top W_0 = \mathbb{1}^\top$. Therefore, $W_0 = u\mathbb{1}^\top$ for some u. Letting $\tilde{\mu} = X_{\mathcal{A}} u$, we have $\|X_{\mathcal{A}} - \tilde{\mu}\mathbb{1}^\top\|_2 \leq 20\sqrt{2n}\sigma$ by (23). In particular, $\mathcal{A}$ is resilient (around its mean) with $\sigma(\epsilon) \leq 20\sigma\sqrt{\frac{2\epsilon}{1-\epsilon}} \leq 40\sigma\sqrt{\epsilon}$ for $\epsilon \leq \frac{1}{2}$. Thus by the proof of Proposition 2 and the fact that $|\mathcal{A}| \geq |S \cap \mathcal{A}| \geq \frac{\alpha(2+\alpha)}{4-\alpha} n \geq (1 - \frac{5}{3}(1 - \alpha))n$, the mean of $\mathcal{A}$ is within $\mathcal{O}(\sigma\sqrt{1 - \alpha})$ of μ, as desired. ◀

4.2 General Case

We are now ready to prove our general algorithmic result, Theorem 7 from Section 1.3. For convenience we recall Theorem 7 here:

▶ **Theorem.** *Suppose that $x_1, \ldots, x_n$ contains a subset S of size $(1 - \epsilon)n$ whose variance around its mean μ is bounded by σ^2 in the norm $\|\cdot\|$. Also suppose that Assumption 6 (κ-approximability) holds for the dual norm $\|\cdot\|_*$. Then, if $\epsilon \leq \frac{1}{4}$, there is a polynomial-time algorithm whose output satisfies $\|\hat{\mu} - \mu\| = \mathcal{O}\big(\sigma\sqrt{\kappa\epsilon}\big)$.*

If, in addition, $\|\cdot\|$ is γ-strongly convex, then even if S only has size αn there is a polynomial-time algorithm such that $\|\hat{\mu} - \mu\| = \mathcal{O}\big(\frac{\sqrt{\kappa}\sigma}{\sqrt{\gamma}\alpha}\big)$ with probability $\Omega(\alpha)$.

Recall that bounded variance means that $\frac{1}{|S|}\sum_{i \in S}\langle x_i - \mu, v\rangle^2 \leq \sigma^2\|v\|_*^2$ for all $v \in \mathbb{R}^d$. There are two equivalent conditions to bounded variance that will be useful. The first is $\sup_{\|v\|_* \leq 1} v^\top \Sigma v \leq \sigma^2$ for all $v \in \mathbb{R}^d$, where $\Sigma = \frac{1}{|S|}\sum_{i \in S}(x_i - \mu)(x_i - \mu)^\top$; this is useful because Assumption 6 allows us to κ-approximate this supremum for any given Σ.

The second equivalent condition re-interprets σ in terms of a matrix norm. Let $\|\cdot\|_\psi$ denote the norm $\|\cdot\|$ above, and for a matrix M define the induced $2 \to \psi$-norm $\|M\|_{2 \to \psi}$ as

$\sup_{\|u\|_2 \le 1} \|Mu\|_\psi$. Then the set S has variance at most σ^2 if and only if $\|[x_i - \mu]_{i \in S}\|_{2\to\psi} \le \sqrt{|S|}\sigma$. This will be useful because induced norms satisfy helpful compositional properties such as $\|AB\|_{2\to\psi} \le \|A\|_{2\to\psi}\|B\|_2$.

The algorithm establishing Theorem 7 is almost identical to Algorithm 1, with two changes. The first change is that on line 4, the quantity $4n\sigma^2$ is replaced with $4\kappa n\sigma^2$, where κ is the approximation factor in Assumption 6. The second change is that in the optimization (14), the constraint $Y \succeq 0, \mathrm{tr}(Y) \le 1$ is replaced with $Y \in \mathcal{P}$, where $\mathcal{P}$ is the feasible set in Assumption 6. In other words, the only difference is that rather than finding the maximum eigenvalue, we κ-approximate the $2 \to \psi$ norm using Assumption 6. We therefore end up solving the saddle point problem

$$\max_{Y \in \mathcal{P}} \min_{W} \Big\{ \sum_{i \in \mathcal{A}} c_i (x_i - X_{\mathcal{A}} w_i)^\top Y (x_i - X_{\mathcal{A}} w_i) \mid 0 \le W_{ji} \le \frac{4-\alpha}{\alpha(2+\alpha)n}, \sum_j W_{ji} = 1 \Big\}. \quad (27)$$

Standard optimization algorithms such as Frank-Wolfe allow us to solve (27) to any given precision with a polynomial number of calls to the linear optimization oracle guaranteed by Assumption 6.

While Algorithm 1 essentially minimizes the quantity $\|X - XW\|_2^2$, this new algorithm can be thought of as minimizing $\|X - XW\|_{2\to\psi}^2$. However, for general norms computing the $2 \to \psi$ norm is NP-hard, and so we rely on a κ-approximate solution by optimizing over $\mathcal{P}$. We are now ready to prove Theorem 7.

Proof (Theorem 7). The proof is similar to Proposition 16, so we only provide a sketch of the differences. First, the condition of Lemma 17 still holds, now with a equal to $\kappa n\sigma^2$ rather than $n\sigma^2$ due to the approximation ratio κ. (This is why we needed to change line 4.)

Next, we need to modify equations (20-23) to hold for the $2 \to \psi$ norm rather than operator norm:

$$\begin{aligned}
\|X_{\mathcal{A}} - Z\|_{2\to\psi} &= \|X_{\mathcal{A}}(I - W)(I - W_1)^{-1}\|_{2\to\psi} && (28)\\
&\overset{(i)}{\le} \|X_{\mathcal{A}}(I - W)\|_{2\to\psi}\|(I - W_1)^{-1}\|_{2\to 2} && (29)\\
&\le 10\|X_{\mathcal{A}}(I - W)\|_{2\to\psi} && (30)\\
&\le 10\sqrt{2}\|X_{\mathcal{A}}(I - W)\,\mathrm{diag}(c_{\mathcal{A}})^{1/2}\|_{2\to\psi} \le 20\sqrt{2\kappa n}\sigma. && (31)
\end{aligned}$$

Here (i) is from the general fact $\|AB\|_{2\to\psi} \le \|A\|_{2\to\psi}\|B\|_{2\to 2}$, and the rest of the inequalities follow for the same reasons as in (20-23).

We next need to modify equations (24-26). This can be done with the following inequality:

▶ **Lemma 18.** *For any matrix $A = [a_1 \ \cdots \ a_n]$ of rank r and any γ-strongly convex norm $\|\cdot\|_\psi$, we have $\sum_{i=1}^n \|a_i\|_\psi^2 \le \frac{r}{\gamma}\|A\|_{2\to\psi}^2$.*

This generalizes the inequality $\|A\|_F^2 \le \mathrm{rank}(A) \cdot \|A\|_2^2$. Using Lemma 18 (proved below), we have

$$\begin{aligned}
\sum_{i \in S \cap \mathcal{A}} \|z_i - \mu\|_\psi^2 &\le \tfrac{\mathrm{rank}(Z)+1}{\gamma}\|[z_i - \mu]_{i \in S\cap\mathcal{A}}\|_{2\to\psi}^2 && (32)\\
&\le \tfrac{\mathrm{rank}(Z)+1}{\gamma}\left(\|[z_i - x_i]_{i \in S\cap\mathcal{A}}\|_{2\to\psi} + \|[x_i - \mu]_{i \in S\cap\mathcal{A}}\|_{2\to\psi}\right)^2 && (33)\\
&= \mathcal{O}\Big(\tfrac{\kappa\sigma^2 n}{\alpha\gamma}\Big). && (34)
\end{aligned}$$

The inequalities again follow for the same reasons as before. If we choose z_i at random, with probability $\Omega(\alpha)$ we will output a z_i with $\|z_i - \mu\|_\psi = \mathcal{O}\left(\frac{\sigma\sqrt{\kappa}}{\alpha\sqrt{\gamma}}\right)$. This completes the first part of the proposition.

For the second part, by the same reasoning as before we obtain $\tilde{\mu}$ with $\|X_{\mathcal{A}} - \tilde{\mu}\mathbb{1}^\top\|_{2\to\psi} = \mathcal{O}(\sqrt{n\kappa}\sigma)$, which implies that $\mathcal{A}$ is resilient with $\sigma_*(\epsilon) = \mathcal{O}(\sigma\sqrt{\kappa\epsilon})$ for $\epsilon \le \frac{1}{2}$. The mean of $\mathcal{A}$ will therefore be within distance $\mathcal{O}(\sigma\sqrt{\kappa\epsilon})$ of μ as before, which completes the proof. ◀

We finish by proving Lemma 18.

Proof (Lemma 18). Let $s \in \{-1, +1\}^n$ be a uniformly random sign vector. We will compare $\mathbb{E}_s[\|As\|_\psi^2]$ in two directions. Let P be the projection onto the span of A. On the one hand, we have $\|As\|_\psi^2 = \|APs\|_\psi^2 \le \|A\|_{2\to\psi}^2\|Ps\|_2^2$, and hence $\mathbb{E}_s[\|As\|_\psi^2] \le \mathbb{E}_s[\|Ps\|_2^2]\|A\|_{2\to\psi}^2 = \mathrm{rank}(A)\|A\|_{2\to\psi}^2$. On the other hand, similarly to (11) we have $\mathbb{E}_s[\|As\|_\psi^2] \ge (1/\gamma)\sum_{i=1}^n \|a_i\|_\psi^2$ by the strong convexity of the norm $\|\cdot\|_\psi$. Combining these yields the desired result. ◀

5 Robust Low-Rank Recovery

In this section we present results on rank-k recovery. We first justify the definition of rank-resilience (Definition 8) by showing that it is information-theoretically sufficient for (approximately) recovering the best rank-k subspace. Then, we provide an algorithm showing that this subspace can be recovered efficiently.

5.1 Information-Theoretic Sufficiency

Let $X_S = [x_i]_{i\in S}$. Recall that δ-rank-resilience asks that $\mathrm{col}(X_T) = \mathrm{col}(X_S)$ and $\|X_T^\dagger X_S\|_2 \le 2$ for $|T| \ge (1-\delta)|S|$. This is justified by the following:

▶ **Proposition 19.** *Let $S \subseteq [n]$ be a set of points of size $(1-\delta)n$ that is $\frac{\delta}{1-\delta}$-rank-resilient. Then it is possible to output a rank-k projection matrix P such that $\|(I-P)X_S\|_2 \le 2\sigma_{k+1}(X_S)$.*

Proof. Find the $\frac{\delta}{1-\delta}$-rank-resilient set S' of size $(1-\delta)n$ such that $\sigma_{k+1}(X'_S)$ is smallest, and let P be the projection onto the top k singular vectors of $X_{S'}$. Then we have $\|(I-P)X_{S'}\|_2 = \sigma_{k+1}(X_{S'}) \le \sigma_{k+1}(X_S)$. Moreover, if we let $T = S \cap S'$, we have $\|(I-P)X_T\|_2 \le \|(I-P)X_{S'}\|_2 \le \sigma_{k+1}(X_S)$ as well. By pigeonhole, $|T| \ge (1-2\delta)n = (1-\frac{\delta}{1-\delta})|S|$. Therefore $\mathrm{col}(X_T) = \mathrm{col}(X_S)$, and $\|(I-P)X_S\|_2 = \|(I-P)X_TX_T^\dagger X_S\|_2 \le \|(I-P)X_T\|_2\|X_T^\dagger X_S\|_2 \le 2\sigma_{k+1}(X_S)$ as claimed. ◀

5.2 Efficient Recovery

We next provide an algorithm for efficient recovery, given as Algorithm 2 below. The proof that it satisfies the guarantees of Theorem 9 is deferred to the full paper.

6 Finite-Sample Concentration

In this section we provide two general finite-sample concentration results that establish resilience with high probability. The first holds for arbitrary resilient distributions but has suboptimal sample complexity, while the latter is specialized to distributions with bounded variance and has near-optimal sample complexity.

Algorithm 2 Algorithm for recovering a rank-k subspace.

1: Initialize $c_i = 1$ for all $i = 1, \dots, n$ and $\mathcal{A} = \{1, \dots, n\}$. Set $\lambda = \frac{(1-\delta)n\sigma^2}{k}$.
2: Let $Y \in \mathbb{R}^{d\times d}$ and $Q \in \mathbb{R}^{\mathcal{A}\times\mathcal{A}}$ be the maximizer/minimizer of the saddle point problem

$$\max_{\substack{Y\succeq 0,\\ \mathrm{tr}(Y)\le 1}} \min_{Q\in\mathbb{R}^{n\times n}} \sum_{i\in\mathcal{A}} c_i[(x_i - X_{\mathcal{A}}q_i)^\top Y(x_i - X_{\mathcal{A}}q_i) + \lambda\|q_i\|_2^2]. \tag{35}$$

3: Let $\tau_i^\star = (x_i - X_{\mathcal{A}}q_i)^\top Y(x_i - X_{\mathcal{A}}q_i) + \lambda\|q_i\|_2^2$.
4: **if** $\sum_{i\in\mathcal{A}} c_i\tau_i^\star > 8n\sigma^2$ **then**
5: For $i \in \mathcal{A}$, replace c_i with $(1 - \frac{\tau_i^\star}{\tau_{\max}})c_i$, where $\tau_{\max} = \max_{i\in\mathcal{A}} \tau_i^\star$.
6: For all i with $c_i < \frac{1}{2}$, remove i from $\mathcal{A}$.
7: Go back to line 3.
8: **end if**
9: Let Q_1 be the result of zeroing out all singular values of Q greater than 0.9.
10: Output $P = X_{\mathcal{A}}Q_0X_{\mathcal{A}}^\dagger$, where $Q_0 = (Q - Q_1)(I - Q_1)^{-1}$.

6.1 Concentration for Resilient Distributions

Our first result, stated as Proposition 4 in Section 1.2, applies to any (σ, ϵ)-resilient distribution p; recall that p is (σ, ϵ)-resilient iff $\|\mathbb{E}[x \mid E] - \mu\| \le \sigma$ for any event E of probability $1 - \epsilon$.

We define the *covering number* of the unit ball in a norm $\|\cdot\|_*$ to be the minimum M for which there are vectors $v_1, \dots, v_M$, each with $\|v_j\|_* \le 1$, such that $\max_{j=1}^M \langle x, v_j\rangle \ge \frac{1}{2}\sup_{\|v\|_*\le 1}\langle x, v\rangle$ for all vectors $v \in \mathbb{R}^d$. Note that $\log M$ is a measure of the effective dimension of the unit ball, i.e. $\log M = \Theta(d)$ if $\|\cdot\|_*$ is the ℓ_∞ or ℓ_2 norm, while $\log M = \Theta(\log d)$ for the ℓ_1 norm.

We recall Proposition 4 for convenience:

▶ **Proposition.** *Suppose that a distribution p is (σ, ϵ)-resilient around its mean μ with $\epsilon < \frac{1}{2}$. Let B be such that $\mathbb{P}[\|x - \mu\| \ge B] \le \epsilon/2$. Also let M be the covering number of the unit ball in the dual norm $\|\cdot\|_*$.*

Then, given n samples $x_1, \dots, x_n \sim p$, with probability $1 - \delta - \exp(-\epsilon n/6)$ there is a subset T of $(1-\epsilon)n$ of the x_i that is (σ', ϵ)-resilient with $\sigma' = \mathcal{O}\Big(\sigma\Big(1 + \sqrt{\frac{\log(M/\delta)}{\epsilon^2 n}} + \frac{(B/\sigma)\log(M/\delta)}{n}\Big)\Big)$.

Proof. Let p' be the distribution of samples from p conditioned on $\|x - \mu\| \le B$. Note that p' is $(\sigma, \frac{\epsilon}{2})$-resilient since every event with probability $1 - \epsilon/2$ in p' is an event of probability $(1 - \epsilon/2)^2 \ge 1 - \epsilon$ in p. Moreover, with probability $1 - \exp(-\epsilon n/6)$, at least $(1 - \epsilon)n$ of the samples from p will come from p'. Therefore, we can focus on establishing resilience of the $n' = (1 - \epsilon)n$ samples from p'.

With a slight abuse of notation, let $x_1, \dots, x_{n'}$ be the samples from p'. Then to check resilience we need to bound $\|\frac{1}{|T|}\sum_{i\in T}(x_i - \mu)\|$ for all sets T of size at least $(1 - \epsilon)n'$. We will first use the covering $v_1, \dots, v_M$ to obtain

$$\Big\|\frac{1}{|T|}\sum_{i\in T}(x_i - \mu)\Big\| \le 2\max_{j=1}^M \frac{1}{|T|}\sum_{i\in T}\langle x_i - \mu, v_j\rangle. \tag{36}$$

The idea will be to analyze the sum over $\langle x_i - \mu, v_j\rangle$ for a fixed v_j and then union bound over the M possibilities. For a fixed v_j, we will split the sum into two components: those with small magnitude (roughly σ/ϵ) and those with large magnitude (between σ/ϵ and B). We can then bound the former with Hoeffding's inequality, and using resilience we will be

able to upper-bound the second moment of the latter, after which we can use Bernstein's inequality.

More formally, let $\tau = \frac{1-\epsilon}{\epsilon/4}\sigma$ and define

$$y_i = \langle x_i - \mu, v_j\rangle \mathbb{I}[|\langle x_i - \mu, v_j\rangle| < \tau], \tag{37}$$

$$z_i = \langle x_i - \mu, v_j\rangle \mathbb{I}[|\langle x_i - \mu, v_j\rangle| \geq \tau]. \tag{38}$$

Clearly $y_i + z_i = \langle x_i - \mu, v_j\rangle$. Also, we have $|y_i| \leq \tau$ almost surely, and $|z_i| \leq B$ almost surely (because $x_i \sim p'$ and hence $\langle x_i - \mu, v_j\rangle \leq \|x_i - \mu\| \leq B$). The threshold τ is chosen so that z_i is non-zero with probability at most $\epsilon/2$ under p (see Lemma 3).

Now, for any set T of size at least $(1-\epsilon)n'$, we have

$$\frac{1}{|T|}\sum_{i\in T}\langle x_i - \mu, v_j\rangle = \frac{1}{|T|}\sum_{i\in T} y_i + z_i \tag{39}$$

$$\leq \Big|\frac{1}{|T|}\sum_{i\in T} y_i\Big| + \frac{1}{|T|}\sum_{i\in T}|z_i| \tag{40}$$

$$\leq \Big|\frac{1}{|T|}\sum_{i=1}^{n'} y_i\Big| + \Big|\frac{1}{|T|}\sum_{i\notin T} y_i\Big| + \frac{1}{|T|}\sum_{i=1}^{n'}|z_i| \tag{41}$$

$$\leq \frac{1}{1-\epsilon}\Big|\frac{1}{n'}\sum_{i=1}^{n'} y_i\Big| + \frac{\epsilon}{1-\epsilon}\tau + \frac{1}{(1-\epsilon)n'}\sum_{i=1}^{n'}|z_i|. \tag{42}$$

The last step uses the fact that $|y_i| \leq \tau$ for all i. It thus suffices to bound $|\frac{1}{n'}\sum_{i=1}^{n'} y_i|$ as well as $\frac{1}{n'}\sum_{i=1}^{n'}|z_i|$.

For the y_i term, note that by resilience $\|\mathbb{E}[y_i]\| \leq \sigma$ (since y_i is sampled from p conditioned on $|\langle x_i - \mu, v_j\rangle| < \tau$ and $\|x_i - \mu\| \leq B$, which each occur with probabiliy at least $1 - \epsilon/2$). Then by Hoeffding's inequality, $|\frac{1}{n'}\sum_{i=1}^{n'} y_i| = \mathcal{O}(\sigma + \tau\sqrt{\log(2/\delta)/n'})$ with probability $1-\delta$.

For the z_i term, we note that $\mathbb{E}[|z_i|] = \mathbb{E}[\max(z_i, 0)] + \mathbb{E}[\max(-z_i, 0)]$. Let τ' be the ϵ-quantile of $\langle x_i - \mu, v_j\rangle$ under p, which by Lemma 3 is at most τ. Then we have

$$\mathbb{E}_p[\max(z_i, 0)] = \mathbb{E}_p[\langle x_i - \mu, v_j\rangle \mathbb{I}[\langle x_i - \mu, v_j\rangle \geq \tau]] \tag{43}$$

$$\leq \mathbb{E}_p[\langle x_i - \mu, v_j\rangle \mathbb{I}[\langle x_i - \mu, v_j\rangle \geq \tau']] \tag{44}$$

$$\overset{(i)}{\leq} \epsilon \cdot \frac{1-\epsilon}{\epsilon}\sigma = (1-\epsilon)\sigma, \tag{45}$$

where (i) is again Lemma 3. Then we have $\mathbb{E}_{p'}[\max(z_i, 0)] \leq \frac{1}{1-\epsilon}\mathbb{E}_p[\max(z_i, 0)] \leq \sigma$, and hence $\mathbb{E}_{p'}[|z_i|] \leq 2\sigma$ (as $\mathbb{E}[\max(-z_i, 0)] \leq \sigma$ by the same argument as above). Since $|z_i| \leq B$, we then have $\mathbb{E}[|z_i|^2] \leq 2B\sigma$.

Therefore, by Bernstein's inequality, with probability $1-\delta$ we have

$$\frac{1}{n'}\sum_{i=1}^{n'}|z_i| \leq \mathcal{O}\Big(\sigma + \sqrt{\frac{\sigma B\log(2/\delta)}{n'}} + \frac{B\log(2/\delta)}{n'}\Big) = \mathcal{O}\Big(\sigma + \frac{B\log(2/\delta)}{n'}\Big). \tag{46}$$

Taking a union bound over the v_j for both y and z, and plugging back into (42), we get that $\frac{1}{|T|}\sum_{i\in T}\langle x_i - \mu, v_j\rangle \leq \mathcal{O}\big(\sigma + \frac{\sigma}{\epsilon}\sqrt{\frac{\log(2M/\delta)}{n}} + \frac{B\log(2M/\delta)}{n}\big)$ for all T and v_j with probability $1-\delta$. Plugging back into (36), we get that $\|\frac{1}{|T|}\sum_{i\in T}(x_i - \mu)\| \leq \mathcal{O}\big(\sigma + \frac{\sigma}{\epsilon}\sqrt{\frac{\log(2M/\delta)}{n}} + \frac{B\log(2M/\delta)}{n}\big)$, as was to be shown. ◀

6.2 Concentration Under Bounded Covariance

In this section we state a stronger but more restrictive finite-sample bound giving conditions under which samples have bounded variance. It is a straightforward extension of Proposition B.1 of [3], so we defer the proof to the full paper.

▶ **Proposition 20.** *Suppose that a distribution p has bounded variance in a norm $\|\cdot\|_*$: $\mathbb{E}_{x\sim p}[\langle x-\mu, v\rangle^2] \le \sigma^2\|v\|_*^2$ for all $v \in \mathbb{R}^d$. Then, given n samples $x_1, \ldots, x_n \sim p$, with probability $1-\exp(-\epsilon^2 n/16)$ there is a subset T of $(1-\epsilon)n$ of the points such that*

$$\frac{1}{|T|}\sum_{i\in T}\langle x_i-\mu, v\rangle^2 \le (\sigma')^2\|v\|_*^2 \text{ for all } v\in\mathbb{R}^d, \text{ where } (\sigma')^2 = \frac{4\sigma^2}{\epsilon}\left(1+\frac{d}{(1-\epsilon)n}\right). \quad (47)$$

This that whenever a distribution p on $\mathbb{R}^d$ has bounded variance, if $n \ge d$ samples x_i are drawn from p then some large subset of the samples will have bounded variance as well.

References

1 S. Balakrishnan, S. S. Du, J. Li, and A. Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory (COLT)*, pages 169–212, 2017.

2 J. Batson, D. A. Spielman, and N. Srivastava. Twice-Ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

3 M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *STOC*, 2017.

4 A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6), 2011.

5 I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *FOCS*, 2016.

6 I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. *arXiv*, 2017.

7 I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robustly learning a Gaussian: Getting optimal error, efficiently. *arXiv*, 2017.

8 I. Diakonikolas, D. Kane, and A. Stewart. Robust learning of fixed-structure Bayesian networks. *arXiv*, 2016.

9 I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. *arXiv*, 2016.

10 I. Diakonikolas, D. M. Kane, and A. Stewart. Learning geometric concepts with nasty noise. *arXiv*, 2017.

11 U. Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

12 D. Kane, S. Karmalkar, and E. Price. Robust polynomial regression up to the information theoretic limit. *arXiv*, 2017.

13 A. Khintchine. Über dyadische brüche. *Mathematische Zeitschrift*, 18:109–116, 1923.

14 A. R. Klivans, P. M. Long, and R. A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009.

15 P. Kothari and J. Steinhardt. Better agnostic clustering via tensor norms. *arXiv*, 2017.

16 K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *FOCS*, 2016.

17 J. Li. Robust sparse estimation tasks in high dimensions. *arXiv*, 2017.

18 L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *STOC*, pages 694–703, 2014.

19 M. Meister and G. Valiant. A data prism: Semi-verified learning in the small-alpha regime. *arXiv*, 2017.

20 E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arXiv*, 2013.

21 Y. Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization methods and software*, 9:141–160, 1998.

22 S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

23 J. Steinhardt. Does robustness imply tractability? A lower bound for planted clique in the semi-random model. *arXiv*, 2017.

24 J. Steinhardt, G. Valiant, and M. Charikar. Avoiding imposters and delinquents: Adversarial crowdsourcing and peer prediction. In *NIPS*, 2016.

25 H. Xu, C. Caramanis, and S. Mannor. Principal component analysis with contaminated data: The high dimensional case. *arXiv*, 2010.