Robust Sub-Gaussian Principal Component Analysis and Width-Independent Schatten Packing

Arun Jambulapati
Stanford University
jmblpati@stanford.edu

Jerry Li
Microsoft Research
jerrl@microsoft.com

Kevin Tian
Stanford University
kjtian@stanford.edu

Abstract

We develop two methods for the following fundamental statistical task: given an ϵ-corrupted set of n samples from a d-dimensional sub-Gaussian distribution, return an approximate top eigenvector of the covariance matrix. Our first robust PCA algorithm runs in polynomial time, returns a $1 - O(\epsilon \log \epsilon^{-1})$-approximate top eigenvector, and is based on a simple iterative filtering approach. Our second, which attains a slightly worse approximation factor, runs in nearly-linear time and sample complexity under a mild spectral gap assumption. These are the first polynomial-time algorithms yielding non-trivial information about the covariance of a corrupted sub-Gaussian distribution without requiring additional algebraic structure of moments. As a key technical tool, we develop the first width-independent solvers for Schatten-p norm packing semidefinite programs, giving a $(1+\epsilon)$-approximate solution in $O(p \log(\frac{nd}{\epsilon})\epsilon^{-1})$ input-sparsity time iterations (where n, d are problem dimensions).

1 Introduction

We study two natural, but seemingly unrelated, problems in high dimensional robust statistics and continuous optimization respectively. As we will see, these problems have an intimate connection.

Problem 1: Robust sub-Gaussian principal component analysis. We consider the following statistical task, which we call *robust sub-Gaussian principal component analysis* (PCA). Given samples $X_1, \ldots, X_n$ from sub-Gaussian[1] distribution $\mathcal{D}$ with covariance $\mathbf{\Sigma}$, an ϵ fraction of which are arbitrarily corrupted, the task asks to output unit vector u with $u^\top \mathbf{\Sigma} u \geq (1-\gamma) \left\|\mathbf{\Sigma}\right\|_\infty$[2] for tolerance γ. Ergo, the goal is to robustly return a $(1-\gamma)$-approximate top eigenvector of the covariance of sub-Gaussian $\mathcal{D}$. This is the natural extension of PCA to the robust statistics setting.

There has been a flurry of recent work on efficient algorithms for robust statistical tasks, e.g. covariance estimation and PCA. From an information-theoretic perspective, sub-Gaussian concentration suffices for robust covariance estimation. Nonetheless, to date all polynomial-time algorithms achieving nontrivial guarantees on covariance estimation (including PCA specifically) in the presence of adversarial noise require additional algebraic structure. For instance, sum-of-squares certifiably bounded moments have been leveraged in polynomial time covariance estimation algorithms [HL18, KSS18]; however, this is a stronger assumption than sub-Gaussianity.

In many applications (see discussion in [DKK+17]), the end goal of covariance estimation is PCA. Thus, a natural question which relaxes robust covariance estimation is: can we robustly estimate the top eigenvector of the covariance $\mathbf{\Sigma}$, assuming only sub-Gaussian concentration? Our work answers this question affirmatively via two incomparable algorithms. The first achieves $\gamma = O(\epsilon \log \epsilon^{-1})$ in

[1] See Section 2 for a formal definition.

[2] Throughout we use $\left\|\mathbf{M}\right\|_p$ to denote the Schatten p-norm (cf. Section 2 for more details).

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

polynomial time; the second achieves $\gamma = O(\sqrt{\epsilon \log \epsilon^{-1} \log d})$, in nearly-linear time under a mild gap assumption on $\mathbf{\Sigma}$. Moreover, both methods have nearly-optimal sample complexity.

Problem 2: Width-independent Schatten packing. We consider a natural generalization of packing semidefinite programs (SDPs) which we call *Schatten packing*. Given symmetric positive semidefinite $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and parameter $p \geq 1$, a Schatten packing SDP asks to solve the optimization problem

$$\min \left\| \sum_{i \in [n]} w_i \mathbf{A}_i \right\|_p \text{ subject to } w \in \Delta^n. \tag{1}$$

Here, $\|\mathbf{M}\|_p$ is the Schatten-p norm of matrix $\mathbf{M}$ and Δ^n is the probability simplex (see Section 2). When $p = \infty$, (1) is the well-studied (standard) packing SDP objective [JY11, ALO16, PTZ16], which asks to find the most spectrally bounded convex combination of packing matrices. For smaller p, the objective encourages combinations more (spectrally) uniformly distributed over directions.

The specialization of (1) to diagonal matrices is a smooth generalization of packing linear programs, previously studied in the context of fair resource allocation [MSZ16, DFO18]. For the ℓ_∞ case of (1), packing SDPs have the desirable property of admitting "width-independent" approximation algorithms via exploiting positivity structure. Specifically, width-independent solvers obtain multiplicative approximations with runtimes independent or logarithmically dependent on size parameters of the problem. This is a strengthening of additive notions of approximation typically used for approximate semidefinite programming. Our work gives the first width-independent solver for Schatten packing.

1.1 Previous work

Learning with adversarial outliers. The study of estimators robust to a small fraction of adversarial outliers dates back to foundational work, e.g. [Hub64, Tuk75]. Following more recent work [LRV16, DKK+19], there has been significant interest in efficient, robust algorithms for statistical tasks in high-dimensional settings. We focus on methods robustly estimating covariance properties here, and defer a thorough discussion of the (extensive) robust statistics literature to [Ste18, Li18, DK19].

There has been quite a bit of work in understanding and giving guarantees for robust covariance estimation where the uncorrupted distribution is exactly Gaussian [DKK+17, DKK+18, DKK+19, CDGW19]. These algorithms strongly use relationships between higher-order moments of Gaussian distributions via Isserlis' theorem. Departing from the Gaussian setting, work of [LRV16] showed that if the distribution is an affine transformation of a 4-wise independent distribution, robust covariance estimation is possible. This was extended by [KSS18], which also assumed nontrivial structure in the moments of the distribution, namely that sub-Gaussianity was certifiable via the sum-of-squares proof system. To the best of our knowledge it has remained open to give nontrivial guarantees for robust estimation of any covariance properties under minimal assumptions, i.e. sub-Gaussian concentration.

All aforementioned algorithms also yield guarantees for robust PCA, by applying a top eigenvector method to the learned covariance. However, performing robust PCA via the intermediate covariance estimation step is lossy, both statistically and computationally. From a statistical perspective, $\Omega(d^2)$ samples are necessary to learn the covariance of a d-dimensional Gaussian in Frobenius norm (and for known efficient algorithms for spectral norm error [DKS17]); in contrast, $O(d)$ samples suffice for (non-robust) PCA. Computationally, even when the underlying distrubution is exactly Gaussian, the best-known covariance estimation algorithms run in time $\Omega(d^{3.25})$; algorithms working in more general settings based on the sum-of-squares approach require much more time. In contrast, the power method for PCA in a $d \times d$ matrix takes time $\widetilde{O}(d^2)$[3]. Motivated by this, our work initiates the direct study of robust PCA, which is often independently interesting in applications.

We remark there is another problem termed "robust PCA" in the literature, e.g. [CLMW11], under a different generative model. We defer a detailed discussion to [DKK+17], which experimentally shows that algorithms from that line of work do not transfer well to our corruption model.

Width-independent iterative methods. Semidefinite programming (SDP) and its linear programming specialization are fundamental computational tasks, with myriad applications in learning, operations research, and computer science. Though general-purpose polynomial time algorithms exist for

[3]We say $g = \widetilde{O}(f)$ if $g = O(f \log^c f)$ for some constant $c > 0$.

SDPs ([NN94]), in practical settings in high dimensions, approximations depending linearly on input size and polynomially on error ϵ are sometimes desirable. To this end, approximation algorithms based on entropic mirror descent have been intensely studied [WK06, AK16, GHM15, AL17, CDST19], obtaining ϵ additive approximations to the objective with runtimes depending polynomially on ρ/ϵ, where ρ is the "width", the largest spectral norm of a constraint.

For structured SDPs, stronger guarantees can be obtained in terms of width. Specifically, several algorithms developed for packing SDPs ((1) with $p = \infty$) yield $(1+\epsilon)$-*multiplicative* approximations to the objective, with *logarithmic* dependence on width [JY11, PTZ16, ALO16, JLL$^+$20]. As ρ upper bounds objective value in this setting, in the worst case runtimes of width-dependent solvers yielding $\epsilon\rho$-additive approximations have similar dependences as width-independent counterparts. Width-independent solvers simultaneously yield stronger multiplicative bounds at all scales of objective value, making them desirable in suitable applications. In particular, ℓ_∞ packing SDPs have found great utility in robust statistics algorithm design [CG18, CDG19, CDGW19, DL19].

Beyond ℓ_∞ packing, width-independent guarantees in the SDP literature are few and far between; to our knowledge, other than the covering and mixed solvers of [JLL$^+$20], ours is the first such guarantee for a broader family of objectives[4]. Our method complements analogous ℓ_p extensions in the width-dependent setting, e.g. [ALO15], as well as width-independent solvers for ℓ_p packing linear programs [MSZ16, DFO18]. We highlight the fair packing solvers of [MSZ16, DFO18], motivated by problems in equitable resource allocation, which further solved ℓ_p packing variants for $p \notin [1, \infty)$. We find analogous problems in semidefinite settings interesting, and defer to future work.

Concurrent work. Concurrent work by Kong et al. [KSKO20] also develops a PCA algorithm tolerant to a bounded fraction of adversarial corruption. Their method is similar to our algorithm based on soft downweighting (Algorithm 6), is analyzed under a fourth moment bound assumption (as opposed to sub-Gaussianity as in this paper), and also generalizes to top-k eigenvector estimation. To our knowledge, our fast algorithm (Algorithm 4) is the first in the literature which robustly solves the 1-PCA problem in near-linear time (for gapped covariances), at the cost of weaker error bounds.

1.2 Our results

Robust sub-Gaussian principal component analysis. We give two algorithms for robust sub-Gaussian PCA[5]. Both are sample optimal, polynomial-time, and assume only sub-Gaussianity. The first is via a simple filtering approach, as summarized in the following (and developed in Section 3).

Theorem 1. *Under Assumption 1, let $\delta \in [0, 1]$, and $n = \Omega\left(\frac{d+\log \delta^{-1}}{(\epsilon \log \epsilon^{-1})^2}\right)$. Algorithm 6 runs in time $O(\frac{nd^2}{\epsilon} \log \frac{n}{\delta\epsilon} \log \frac{n}{\delta})$, and outputs u with $u^\top \mathbf{\Sigma} u > (1 - C^\star \epsilon \log \epsilon^{-1}) \|\mathbf{\Sigma}\|_\infty$, for $C^\star$ a fixed multiple of parameter c in Assumption 1, with probability at least $1 - \delta$.*

Our second algorithm is more efficient under mild conditions, but yields a worse approximation $1 - \gamma$ for $\gamma = O(\sqrt{\epsilon \log \epsilon^{-1} \log d})$. Specifically, if there are few eigenvalues of $\mathbf{\Sigma}$ larger than $1 - \gamma$, our algorithm runs in nearly-linear time. Note that if there are many eigenvalues above this threshold, then the PCA problem itself is not very well-posed; our algorithm is very efficient in the interesting setting where the approximate top eigenvector is identifiable. We state our main algorithmic guarantee here, and defer details to Section 5.

Theorem 2. *Under Assumption 1, let $\delta \in [0, 1]$, $n = \Omega\left(\frac{d+\log \delta^{-1}}{(\epsilon \log \epsilon^{-1})^2}\right)$, $\gamma = C\sqrt{\epsilon \log \epsilon^{-1} \log d}$, for C a fixed multiple of parameter c from Assumption 1, and let $t \in [d]$ satisfy $\mathbf{\Sigma}_{t+1} < (1 - \gamma) \|\mathbf{\Sigma}\|_\infty$. Algorithm 4 outputs a unit vector $u \in \mathbb{R}^d$ with $u^\top \mathbf{\Sigma} u \geq (1 - \gamma)\|\mathbf{\Sigma}\|_\infty$ in time $\widetilde{O}(\frac{nd}{\epsilon^{4.5}} + \frac{ndt}{\epsilon^{1.5}})$.*

Since $\Omega(d\epsilon^{-2})$ samples are necessary for a $(1 - \epsilon)$-approximation to the top eigenvector of $\mathbf{\Sigma}$ via uncorrupted samples, our first method is sample-optimal, as is our second up to a $\widetilde{O}(\epsilon^{-1})$ factor.

Width-independent Schatten packing. Our second method crucially requires an efficient solver for Schatten packing SDPs. We demonstrate that Schatten packing, i.e. (1) for arbitrary p, admits width-independent solvers. We state an informal guarantee, and defer details to Section 4.

[4] In concurrent and independent work, [CMY20] develops width-independent solvers for Ky-Fan packing objectives, a different notion of generalization than the Schatten packing objectives we consider.

[5] We follow the distribution and corruption model described in Assumption 1.

Theorem 3. *Let $\{\mathbf{A}_i\}_{i\in[n]} \in \mathbb{S}^d_{\geq 0}$, and $\epsilon > 0$. There is an algorithm taking $O(\frac{p\log(\frac{nd}{\epsilon})}{\epsilon})$ iterations, returning a $1+\epsilon$ multiplicative approximation to the problem* (1)*. For odd p, each iteration can be implemented in time nearly-linear in the number of nonzeros amongst all $\{\mathbf{A}_i\}_{i\in[n]}$.*

2 Preliminaries

General notation. $[n]$ denotes the set $1 \leq i \leq n$. The operation $\circ$ applied to two vectors of equal dimension is their entrywise product. Applied to a vector, $\|\cdot\|_p$ is the ℓ_p norm; applied to a symmetric matrix, $\|\cdot\|_p$ is the Schatten-p norm, i.e. the ℓ_p norm of the spectrum. The *dual norm* of ℓ_p is ℓ_q for $q = \frac{p}{p-1}$; when $p = \infty$, $q = 1$. Δ^n is the n-dimensional simplex (subset of positive orthant with ℓ_1-norm 1) and we define $\mathfrak{S}^n_\varepsilon \subseteq \Delta^n$ to be the truncated simplex:

$$\mathfrak{S}^n_\varepsilon := \left\{ w \in \mathbb{R}^n_{\geq 0} \;\middle|\; \|w\|_1 = 1,\ w \leq \frac{1}{n(1-\varepsilon)} \text{ entrywise} \right\}. \tag{2}$$

Matrices. $\mathbb{S}^d$ is $d \times d$ symmetric matrices, and $\mathbb{S}^d_{\geq 0}$ is the positive semidefinite subset. $\mathbf{I}$ is the identity of appropriate dimension. $\lambda_{\max}$, $\lambda_{\min}$, and Tr are the largest and smallest eigenvalues and trace of a symmetric matrix. For $\mathbf{M}, \mathbf{N} \in \mathbb{S}^d$, $\langle \mathbf{M}, \mathbf{N} \rangle := \mathrm{Tr}\,(\mathbf{MN})$ and we use the Loewner order $\preceq$, ($\mathbf{M} \preceq \mathbf{N}$ iff $\mathbf{N} - \mathbf{M} \in \mathbb{S}^d_{\geq 0}$). The seminorm of $\mathbf{M} \succeq 0$ is $\|v\|_{\mathbf{M}} := \sqrt{v^\top \mathbf{M} v}$.

Fact 1. *For $\mathbf{A}$, $\mathbf{B}$ with compatible dimension,* $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$. *For $\mathbf{M}, \mathbf{N} \in \mathbb{S}^d_{\geq 0}$, $\langle \mathbf{M}, \mathbf{N} \rangle \geq 0$.*

Fact 2. *We have the following characterization of the Schatten-p norm: for $\mathbf{M} \in \mathbb{S}^d$, and $q = \frac{p}{p-1}$,*

$$\|\mathbf{M}\|_p = \sup_{\mathbf{N}\in\mathbb{S}^d,\ \|\mathbf{N}\|_q = 1} \langle \mathbf{N}, \mathbf{M} \rangle .$$

For $\mathbf{M} = \sum_{j\in[d]} \lambda_i v_i v_i^\top$, the satisfying $\mathbf{N}$ is $\frac{\sum_{j\in[d]} \pm\lambda_i^{p-1} v_i v_i^\top}{\|\mathbf{M}\|_p^{p-1}}$, so $\mathbf{NM}$ has spectrum $\frac{|\lambda|^p}{\|\mathbf{M}\|_p^{p-1}}$.

Distributions. We denote drawing vector X from distribution $\mathcal{D}$ by $X \sim \mathcal{D}$, and the covariance $\mathbf{\Sigma}$ of $\mathcal{D}$ is $\mathbb{E}_{X\sim\mathcal{D}}\left[XX^\top\right]$. We say scalar distribution $\mathcal{D}$ is γ^2-sub-Gaussian if $\mathbb{E}_{X\sim\mathcal{D}}[X] = 0$ and

$$\mathbb{E}_{X\sim\mathcal{D}}\left[\exp\left(tX\right)\right] \leq \exp\left(\frac{t^2\gamma^2}{2}\right)\ \forall t \in \mathbb{R}.$$

Multivariate $\mathcal{D}$ has sub-Gaussian proxy $\mathbf{\Gamma}$ if its restriction to any unit v is $\|v\|^2_{\mathbf{\Gamma}}$-sub-Gaussian, i.e.

$$\mathbb{E}_{X\sim\mathcal{D}}\left[\exp\left(tX^\top v\right)\right] \leq \exp\left(\frac{t^2\|v\|^2_{\mathbf{\Gamma}}}{2}\right) \text{ for all } \|v\|_2 = 1,\ t \in \mathbb{R}. \tag{3}$$

We consider the following standard model for gross corruption with respect to a distribution $\mathcal{D}$.

Assumption 1 (Corruption model, see [DKK+19])**.** *Let $\mathcal{D}$ be a mean-zero distribution on $\mathbb{R}^d$ with covariance $\mathbf{\Sigma}$ and sub-Gaussian proxy $\mathbf{\Gamma} \preceq c\mathbf{\Sigma}$ for a constant c. Denote by index set G' with $|G'| = n$ a set of (uncorrupted) samples $\{X_i\}_{i\in G'} \sim \mathcal{D}$. An adversary arbitrarily replaces ϵn points in G'; we denote the new index set by $[n] = B \cup G$, where B is the (unknown) set of points added by an adversary, and $G \subseteq G'$ is the set of points from G' that were not changed.*

As we only estimate covariance properties, the assumption that $\mathcal{D}$ is mean-zero only loses constants in problem parameters, by pairing samples and subtracting them (cf. [DKK+19], Section 4.5.1).

3 Robust sub-Gaussian PCA via filtering

In this section, we sketch the proof of Theorem 1, which gives guarantees on our filtering algorithm for robust sub-Gaussian PCA. This algorithm obtains stronger statistical guarantees than Theorem 2, at the cost of super-linear runtime; the algorithm is given as Algorithm 6. Our analysis stems largely from concentration facts about sub-Gaussian distributions, as well as the following (folklore) fact regarding estimation of variance along any particular direction.

Lemma 1. *Under Assumption 1, let $\delta \in [0,1]$, $n = \Omega\left(\frac{\log \delta^{-1}}{(\epsilon \log \epsilon^{-1})^2}\right)$, and $u \in \mathbb{R}^d$ be a fixed unit vector. Algorithm 5,* 1DRobustVariance*, takes input $\{X_i\}_{i\in[n]}$, u, and ϵ, and outputs σ_u^2 with $|u^\top \mathbf{\Sigma} u - \sigma_u^2| < C u^\top \mathbf{\Sigma} u \cdot \epsilon \log \epsilon^{-1}$ with probability at least $1-\delta$, and runs in time $O(nd + n \log n)$, for C a fixed multiple of the parameter c in Assumption 1.*

In other words, we show that using corrupted samples, we can efficiently estimate a $1 + O(\epsilon \log \epsilon^{-1})$-multiplicative approximation of the variance of $\mathcal{D}$ in any unit direction[6]. This proof is deferred to Appendix B for completeness. Algorithm 6 combines this key insight with a soft filtering approach which has found many applications in the recent robust statistics literature, suggested by the following known structural fact found in previous work (e.g. Lemma A.1 of [DHL19], see also [SCV17, Ste18]).

Lemma 2. *Let $\{a_i\}_{i\in[m]}$, $\{w_i\}_{i\in[m]}$ be sets of nonnegative reals, and $a_{\max} = \max_{i\in[m]} a_i$. Define $w_i' = \left(1 - \frac{a_i}{a_{\max}}\right) w_i$, for all $i \in [m]$. Consider any disjoint partition I_B, I_G of $[m]$ with $\sum_{i\in I_B} w_i a_i > \sum_{i\in I_G} w_i a_i$. Then, $\sum_{i\in I_B} w_i - w_i' > \frac{1}{2a_{\max}} \sum_{i\in[m]} w_i a_i > \sum_{i\in I_G} w_i - w_i'$.*

Our Algorithm 6, PCAFilter, takes as input a set of corrupted samples $\{X_i\}_{i\in[n]}$ following Assumption 1 and the corruption parameter ϵ. At a high level, it initializes a uniform weight vector $w^{(0)}$, and iteratively operates as follows (we denote by $\mathbf{M}(w)$ the empirical covariance $\sum_{i\in[n]} w_i X_i X_i^\top$).

1. $u_t \leftarrow$ approximate top eigenvector of $\mathbf{M}(w^{(t-1)})$ via power iteration.
2. Compute $\sigma_t^2 \leftarrow$ 1DRobustVariance$(\{X_i\}_{i\in[n]}, u_t, \epsilon)$.
3. If $\sigma_t^2 > (1 - O(\epsilon \log \epsilon^{-1})) \cdot u_t^\top \mathbf{M}(w^{(t-1)}) u_t$, then terminate and return u_t.
4. Else:
   (a) Sort indices $i \in [n]$ by $a_i \leftarrow \langle u_t, X_i \rangle^2$, with a_1 smallest.
   (b) Let $\ell \le i \le n$ be the smallest set for which $\sum_{i=\ell}^n w_i \ge 2\epsilon$, and apply the downweighting procedure of Lemma 2 to this subset of indices.

The analysis of Algorithm 6 then proceeds in two stages.

Monotonicity of downweighting. We show the invariant criteria for Lemma 2 (namely, that for the set $\ell \le i \le n$ in every iteration, there is more spectral mass on bad points than good) holds inductively for our algorithm. Specifically, lack of termination implies $\mathbf{M}(w^{(t-1)})$ puts significant mass on bad directions, which combined with concentration of good directions yields the invariant. The details of this argument can be found as Lemma 11.

Roughly uniform weightings imply approximation quality. As Lemma 2 then applies, the procedure always removes more mass from bad points than good, and thus can only remove at most 2ϵ mass total by the corruption model. Thus, the weights $w^{(t)}$ are always roughly uniform (in $\mathfrak{S}_{O(\epsilon)}^n$), which by standard concentration facts (see Appendix A) imply the quality of the approximate top eigenvector is good. Moreover, the iteration count is bounded by roughly d because whenever the algorithm does not terminate, enough mass is removed from large spectral directions. Combining with the termination criteria imply that when a vector is returned, it is a close approximation to the top direction of $\mathbf{\Sigma}$. Details can be found as Lemma 13 and in the proof of Theorem 1.

4 Schatten packing

For our second robust PCA algorithm, developed in Section 5, we require a key technical tool which we now develop in this section. The tool, Schatten-norm packing semidefinite programs (and hybrid-norm extensions), is a smoothed generalization of the classical packing semidefinite program, which may be of independent interest in other applications. At a high level, the reason Schatten packing solvers are useful for the robust PCA problem is because while an adversary can fool a PCA algorithm based on operator-norm semidefinite programs by "promoting" a single other eigenvector to have a larger variance, a p-norm-based semidefinite program forces a tradeoff between the number of directions promoted and the amount of variance introduced.

[6]Corollary 4 gives a slightly stronger guarantee that reusing samples does not break dependencies of u.

4.1 Mirror descent interpretation of [MRWZ16]

We begin by reinterpreting the [MRWZ16] solver, which achieves the state-of-the-art parallel runtime for packing LPs[7]. An (ℓ_∞) packing LP algorithm solves the following decision problem.[8].

Problem 1 (ℓ_∞ packing linear program)**.** *Given entrywise nonnegative $\mathbf{A} \in \mathbb{R}^{d\times n}_{\geq 0}$, either find primal solution $x \in \Delta^n$ with $\|\mathbf{A}x\|_\infty \leq 1+\epsilon$ or dual solution $y \in \Delta^d$ with $\mathbf{A}^\top y \geq (1-\epsilon)\mathbf{1}$.*

Algorithm 1 PackingLP$(\mathbf{A}, \epsilon)$

Input: $\mathbf{A} \in \mathbb{R}^{d\times n}_{\geq 0}$, $\epsilon \in [0, \frac{1}{2}]$
$K \leftarrow \frac{3\log(d)}{\epsilon}$, $\eta \leftarrow K^{-1}$, $T \leftarrow \frac{4\log(d)\log(nd/\epsilon)}{\epsilon^2}$
$[w_0]_i \leftarrow \frac{\epsilon}{n^2 d}$ for all $i \in [n]$, $z \leftarrow \mathbf{0}$, $t \leftarrow 0$
while $\mathbf{A}w_t \leq K\mathbf{1}$, $\|w_t\|_1 \leq K$ **do**
  $v_t \leftarrow \frac{\exp(\mathbf{A}w_t)}{\|\exp(\mathbf{A}w_t)\|_1}$
  $g_t \leftarrow \max(0, \mathbf{1} - \mathbf{A}^\top v_t)$ entrywise
  $w_{t+1} \leftarrow w_t \circ (1+\eta g_t)$, $z \leftarrow z + v_t$, $t \leftarrow t+1$
  if $t \geq T$ **then**
    return $y \leftarrow \frac{1}{T} z$
  end if
end while
return $x \leftarrow \frac{w_t}{\|w_t\|_1}$

The following result is shown in [MRWZ16].

Proposition 1. PackingLP *(Algorithm 1) solves Problem 1 in $O(\text{nnz}(\mathbf{A}) \cdot \frac{\log(d)\log(nd/\epsilon)}{\epsilon^2})$ time.*

Our interpretation of the analysis of [MRWZ16] combines two ingredients: a potential argument and mirror descent (alternatively known as the "multiplicative weights" framework), which yields a dual feasible point if $\|w_t\|_1$ did not grow sufficiently.

Potential argument. The potential used by [MRWZ16] is $\log(\sum_{j\in[d]} \exp([\mathbf{A}w_t]_j)) - \|w_t\|_1$, well-known to be a $O(\log d)$-additive approximation of $\|\mathbf{A}w_t\|_\infty - \|w_t\|_1$. As soon as $\|\mathbf{A}w_t\|_\infty$ or $\|w_t\|_1$ reaches the scale $O(\frac{\log d}{\epsilon})$, by nonnegativity this becomes a multiplicative guarantee, motivating the setting of threshold K. To prove the potential is monotone, [MRWZ16] uses step size K^{-1} and a Taylor approximation; combining with the termination condition yields the desired claim.

Mirror descent. To certify that w_t grows sufficiently (e.g. the method terminates in few iterations, else dual feasibility holds), we interpret the step $w_{t+1} \leftarrow w_t \circ (1+\eta g_t)$ as approximate entropic mirror descent. Specifically, we track the quantity $\sum_{0\leq t<T} \langle \eta g_t, u\rangle$, and show that if $\|w_t\|_1$ has not grown sufficiently, then it must be bounded for every $u \in \Delta^n$, certifying dual feasibility. Formally, for any g_t sequence and $u \in \Delta^n$, we show

$$O(\log(nd/\epsilon)) + \log\left(\frac{\|w_T\|_1}{\|w_0\|_1}\right) \geq \sum_{0\leq t<T} \langle \eta g_t, u\rangle \geq \eta \sum_{0\leq t<T} \left\langle \mathbf{1} - \mathbf{A}^\top v_t, u\right\rangle.$$

The last inequality followed by g_t being an upwards truncation. If $\|w_T\|_1$ is bounded (else, we have primal feasibility), we show the entire above expression is bounded $O(\log \frac{nd}{\epsilon})$ for any u. Thus, by setting $T = O(\frac{\log(nd/\epsilon)}{\eta\epsilon})$ and choosing u to be each coordinate indicator, it follows that the average of all v_t is coordinatewise at least $1-\epsilon$, and solves Problem 1 as a dual solution.

Our g_t is the (truncated) gradient of the function used in the potential analysis, so its form allows us to interpret dual feasibility (e.g. v_t has ℓ_1 norm 1 and is a valid dual point). Our analysis patterns standard mirror descent, complemented by side information which says that lack of a primal solution can transform a regret guarantee into a feasibility bound. We apply this framework to analyze ℓ_p

[7]The [MRWZ16] solver also generalizes to covering and mixed objectives; we focus on packing in this work.

[8]Packing linear programs are sometimes expressed as the optimization problem $\max_{x\geq 0, \mathbf{A}x\leq \mathbf{1}} \|x\|_1$, similarly to (1); these problems are equivalent up to a standard binary search, see e.g. discussion in [JLL+20].

variants of Problem 1, via different potentials; our proofs are quite straightforward upon adopting this perspective, and we believe it may yield new insights for instances with positivity structure.

4.2 ℓ_p-norm packing linear programs

In this section, we give an example of the framework proposed in Section 4.1, for approximately solving ℓ_p norm packing linear programs. Specifically, we now consider the generalization of Problem 1 to ℓ_p norms; throughout, $q = \frac{p}{p-1}$ is the dual norm.

Problem 2 (ℓ_p packing linear program)**.** *Given entrywise nonnegative $\mathbf{A} \in \mathbb{R}^{d \times n}_{\geq 0}$, either find primal solution $x \in \Delta^n$ with $\|\mathbf{A}x\|_p \leq 1 + \epsilon$ or dual solution $y \in \mathbb{R}^d_{\geq 0}$, $\|y\|_q = 1$ with $\mathbf{A}^\top y \geq (1 - \epsilon)\mathbf{1}$.*

For $p = \frac{\log d}{\epsilon}$, Problem 2 recovers Problem 1 up to constants as ℓ_p multiplicatively approximates ℓ_∞ by $1 + \epsilon$. We now state our method for solving Problem 2 as Algorithm 2.

Algorithm 2 PNormPacking$(\mathbf{A}, \epsilon, p)$

1: **Input:** $\mathbf{A} \in \mathbb{R}^{d \times n}_{\geq 0}, \epsilon \in [0, \frac{1}{2}], p \geq 2$
2: $\eta \leftarrow p^{-1}, T \leftarrow \frac{4p \log(\frac{nd}{\epsilon})}{\epsilon}$
3: $[w_0]_i \leftarrow \frac{\epsilon}{n^2 d}$ for all $i \in [n]$, $z \leftarrow \mathbf{0}$, $t \leftarrow 0$
4: **while** $\|w_t\|_1 \leq \epsilon^{-1}$ **do**
5: $\quad g_t \leftarrow \max(0, \mathbf{1} - \mathbf{A}^\top (v_t)^{p-1})$ entrywise, for $v_t \leftarrow \frac{\mathbf{A}w_t}{\|\mathbf{A}w_t\|_p}$
6: $\quad w_{t+1} \leftarrow w_t \circ (1 + \eta g_t)$, $z \leftarrow z + (v_t)^{p-1}$, $t \leftarrow t + 1$
7: $\quad$ **if** $t \geq T$ **then**
8: $\quad\quad$ **return** $y = \frac{z}{\|z\|_q}$
9: $\quad$ **end if**
10: **end while**
11: **return** $x = \frac{w_t}{\|w_t\|_1}$

Other than changing parameters, the only difference from Algorithm 1 is that v is a point with unit ℓ_q norm induced by the gradient of our potential Φ_t. We state our main potential fact, whose proof is based straightforwardly on Taylor expanding $\|\cdot\|_p$, and deferred to Appendix C for brevity.

Lemma 3. *In all iterations t of Algorithm 2, defining $\Phi_t := \|\mathbf{A}w_t\|_p - \|w_t\|_1$, $\Phi_{t+1} \leq \Phi_t$.*

We now state our main result, which leverages the potential bound following the framework of Section 4.1. A proof can be found in Appendix C.

Theorem 4. *Algorithm 2 runs in time $O(\text{nnz}(\mathbf{A}) \cdot \frac{p \log(nd/\epsilon)}{\epsilon})$. Further, its output solves Problem 2.*

4.3 Schatten-norm packing semidefinite programs

We generalize Algorithm 2 to solve Schatten packing semidefinite programs, which we now define.

Problem 3. *Given $\{\mathbf{A}_i\}_{i \in [n]} \in \mathbb{S}^d_{\geq 0}$, either find primal solution $x \in \Delta^n$ with $\left\|\sum_{i \in [n]} x_i \mathbf{A}_i\right\|_p \leq 1 + \epsilon$ or dual solution $\mathbf{Y} \in \mathbb{S}^d_{\geq 0}$, $\|\mathbf{Y}\|_q = 1$ with $\langle \mathbf{A}_i, \mathbf{Y} \rangle \geq 1 - \epsilon$ for all $i \in [n]$.*

We assume that p is an odd integer for simplicity (sufficient for our applications), and leave for interesting future work the cases when p is even or noninteger. The potential used in the analysis and an overall guarantee are stated here, and deferred to Appendix C. The proofs are simple modifications of Lemma 3 and Theorem 4 using trace inequalities (similar to those in [JLL+20]) in place of scalar inequalities, as well as efficient approximation of quantities in Line 5 via the standard technique of Johnson-Lindestrauss projections.

Lemma 4. *In all iterations t of Algorithm 3, defining $\Phi_t := \left\|\sum_{i \in [n]} [w_t]_i \mathbf{A}_i\right\|_p - \|w_t\|_1$, $\Phi_{t+1} \leq \Phi_t$.*

Algorithm 3 SchattenPacking$(\{\mathbf{A}_i\}_{i\in[n]}, \epsilon, p)$

Input: $\{\mathbf{A}_i\}_{i\in[n]} \in \mathbb{S}^d_{\geq 0}, \epsilon \in [0, \frac{1}{2}], p \geq 2$
$\eta \leftarrow p^{-1}, T \leftarrow \frac{4p\log(\frac{nd}{\epsilon})}{\epsilon}$
$[w_0]_i \leftarrow \frac{\epsilon}{n^2 d}$ for all $i \in [n]$, $z \leftarrow 0$
while $\|w_t\|_1 \leq \epsilon^{-1}$ **do**
  $g_t \leftarrow \max\left(0, \mathbf{1} - \left\langle \mathbf{A}_i, \mathbf{V}_t^{p-1} \right\rangle\right)$ entrywise, for $\mathbf{V}_t \leftarrow \frac{\sum_{i\in[n]}[w_t]_i \mathbf{A}_i}{\left\|\sum_{i\in[n]}[w_t]_i \mathbf{A}_i\right\|_p}$
  $w_{t+1} \leftarrow w_t \circ (1 + \eta g_t)$, $\mathbf{Z} \leftarrow \mathbf{Z} + (\mathbf{V}_t)^{p-1}$, $t \leftarrow t + 1$
  if $t \geq T$ **then**
    return $\mathbf{Y} = \frac{\mathbf{Z}}{\|\mathbf{Z}\|_q}$
  end if
end while
return $x = \frac{w_t}{\|w_t\|_1}$

Theorem 5. *Let p be odd. Algorithm 3 runs in $O(\frac{p\log(nd/\epsilon)}{\epsilon})$ iterations, and its output solves Problem 3. Each iteration is implementable in $O(\text{nnz} \cdot \frac{p\log(nd/\epsilon)}{\epsilon^2})$, where* nnz *is the number of nonzero entries amongst all $\{\mathbf{A}_i\}_{i\in[n]}$, losing $O(\epsilon)$ in the quality of Problem 3 with probability $1 - \text{poly}((nd/\epsilon)^{-1})$.*

4.4 Schatten packing with a ℓ_∞ constraint

We remark that the framework outlined in Section 4.1 is flexible enough to handle mixed-norm packing problems. Specifically, developments in Section 5 require the following guarantee.

Proposition 2. *Following Theorem 5's notation, let p be odd, $\{\mathbf{A}_i\}_{i\in[n]} \in \mathbb{S}^d_{\geq 0}$, $0 < \epsilon = O(\alpha)$, and*

$$\min_{\substack{x\in\Delta^n \\ \|x\|_\infty \leq \frac{1+\alpha}{n}}} \|\mathcal{A}(x)\|_p = \text{OPT}. \tag{4}$$

for $\mathcal{A}(x) := \sum_{i\in[n]} x_i \mathbf{A}_i$. Given estimate of OPT *exponentially bounded in $\frac{nd}{\epsilon}$, there is a procedure calling Algorithm 7 $O(\log \frac{nd}{\epsilon})$ times giving $x \in \Delta^n$ with $\|x\|_\infty \leq \frac{(1+\alpha)(1+\epsilon)}{n}$, $\|\mathcal{A}(x)\|_p \leq (1 + \epsilon)$*OPT*. Algorithm 7 runs in $O(\frac{\log(nd/\epsilon)\log n}{\epsilon^2})$ iterations, each requiring time $O(\text{nnz} \cdot \frac{p\log(nd/\epsilon)}{\epsilon^2})$.*

Our method, found in Appendix C, approximately solves (4) by first applying a standard binary search to place $\mathcal{A}(x)$ on the right scale, for which it suffices to solve an approximate decision problem. Then, we apply a truncated mirror descent procedure on the potential $\Phi(w) = \log(\exp(\|\mathcal{A}(w)\|_p) + \exp(\frac{n}{1+\alpha}\|w\|_\infty)) - \|w\|_1$, and prove correctness for solving the decision problem following the framework we outlined in Section 4.1.

5 Robust sub-Gaussian PCA in nearly-linear time

We give our nearly-linear time robust PCA method, leveraging developments of Section 4. Throughout, we will be operating under Assumption 1, for some corruption parameter ϵ with $\epsilon \log \epsilon^{-1} \log d = O(1)$; $\epsilon = O(\frac{1}{\log d \log\log d})$ suffices. We now develop tools to prove Theorem 2.

Algorithm 4 uses three subroutines: our earlier 1DRobustVariance method (Lemma 1), an application of our earlier Proposition 2 to approximate the solution to

$$\min_{w\in\mathfrak{S}^n_\epsilon} \left\| \sum_{i\in[n]} w_i X_i X_i^\top \right\|_p, \text{ for } p = \Theta\left(\sqrt{\frac{\log d}{\epsilon \log \epsilon^{-1}}}\right), \tag{5}$$

and a method for computing approximate eigenvectors by [MM15] (discussed in Appendix D).

Proposition 3. *There is an algorithm* Power *(Algorithm 1, [MM15]), parameterized by $t \in [d]$, tolerance $\tilde{\epsilon} > 0$, $p \geq 1$, and $\mathbf{A} \in \mathbb{S}^d_{\geq 0}$, which outputs orthonormal $\{z_j\}_{j\in[t]}$ with the guarantee*

$$\left.\begin{aligned} \left|z_j^\top \mathbf{A}^p z_j - \lambda_j^p(\mathbf{A})\right| &\leq \tilde{\epsilon}\lambda_j^p(\mathbf{A}) \\ \left|z_j^\top \mathbf{A}^{p-1} z_j - \lambda_j^{p-1}(\mathbf{A})\right| &\leq \tilde{\epsilon}\lambda_j^{p-1}(\mathbf{A}) \end{aligned}\right\} \text{ for all } j \in [t]. \tag{6}$$

Here, $\lambda_j(\mathbf{A})$ is the j^{th} largest eigenvalue of $\mathbf{A}$. The total time required by the method is $O(\text{nnz}(\mathbf{A})\frac{tp\log d}{\varepsilon})$.

Algorithm 4 RobustPCA$(\{X_i\}_{i\in[n]}, \epsilon, t)$

1: **Input:** $\{X_i\}_{i\in[n]}$ $\epsilon = O(\frac{1}{\log d \log\log d})$, $t \in [d]$ with $\mathbf{\Sigma}_{t+1} \leq (1-\gamma)\mathbf{\Sigma}$ for γ in Theorem 2
2: $w \leftarrow$ BoxedSchattenPacking (Proposition 2) on $\{\mathbf{A}_i = X_i X_i^\top\}_{i\in[n]}$, $\alpha \leftarrow \epsilon$, p as in (5)
3: $\mathbf{M} = \sum_{i\in[n]} w_i X_i X_i^\top$
4: $\{z_j\}_{j\in[t]} =$ Power$(t, \epsilon, p, \mathbf{M})$
5: $\alpha_j \leftarrow$ 1DRobustVariance$(\{X_i\}_{i\in[n]}, \mathbf{M}^{\frac{p-1}{2}} z_j / \|\mathbf{M}^{\frac{p-1}{2}} z_j\|_2, \epsilon)$ for all $j \in [t]$
6: **return** z_{j^*} for $j^* = \text{argmax}_{j\in[t]}\alpha_j$

Algorithm 4 is computationally bottlenecked by the application of Proposition 2 on Line 2 and the t calls to 1DRobustVariance on Line 5, from which the runtime guarantee of Theorem 2 follows straightforwardly. To demonstrate correctness, we first certify the quality of the solution to (5).

Lemma 5. *Let $n = \Omega\left(\frac{d+\log\delta^{-1}}{(\epsilon\log\epsilon^{-1})^2}\right)$. With probability $1-\frac{\delta}{2}$, the uniform distribution over G attains value $(1+\frac{\tilde{\epsilon}}{2})\|\mathbf{\Sigma}\|_p$ for objective* (5)*, where $\tilde{\epsilon} = C'\epsilon\log\epsilon^{-1}$ for a universal constant $C' > 0$.*

The proof of this is similar to results in e.g. [DKK+19, Li18], and combines concentration guarantees with a union bound over all possible corruption sets B. This implies the following immediately, upon applying the guarantees of Proposition 2.

Corollary 1. *Let w be the output of Line 2 of* RobustPCA*. Then, we have $\|w\|_\infty \leq \frac{1}{(1-2\epsilon)n}$, and $\left\|\sum_{i\in[n]} w_i X_i X_i^\top\right\|_p \leq (1+\tilde{\epsilon})\|\mathbf{\Sigma}\|_p$ under the guarantee of Lemma 5.*

Let w be the output of the solver. Recall that $\mathbf{M} = \sum_{i=1}^n w_i X_i X_i^\top$. Additionally, define

$$\mathbf{M}_G := \sum_{i\in G} w_i X_i X_i^\top,\ w_G := \sum_{i\in G} w_i,\ \mathbf{M}_B := \sum_{i\in B} w_i X_i X_i^\top,\ w_B := \sum_{i\in G} w_i\ . \tag{7}$$

Notice in particular that $\mathbf{M} = \mathbf{M}_G + \mathbf{M}_B$, and that all these matrices are PSD. We next prove the second, crucial fact, which says that $\mathbf{M}_G$ is a good approximator to $\mathbf{\Sigma}$ in Loewner ordering:

Lemma 6. *Let $n = \Omega\left(\frac{d+\log\delta^{-1}}{(\epsilon\log\epsilon^{-1})^2}\right)$. With probability at least $1-\frac{\delta}{2}$, $(1+\tilde{\epsilon})\mathbf{\Sigma} \succeq \mathbf{M}_G \succeq (1-\tilde{\epsilon})\mathbf{\Sigma}$.*

The proof combines the strategy in Lemma 5 with the SDP solver guarantee. Perhaps surprisingly, Corollary 1 and Lemma 6 are the only two properties about $\mathbf{M}$ that our final analysis of Theorem 2 will need. In particular, we have the following key geometric proposition, which carefully combines trace inequalities to argue that the corrupted points cannot create too many new large eigendirections.

Proposition 4. *Let $\mathbf{M} = \mathbf{M}_G + \mathbf{M}_B$ be so that $\|\mathbf{M}\|_p \leq (1+\tilde{\epsilon})\|\mathbf{\Sigma}\|_p$, $\mathbf{M}_G \succeq 0$ and $\mathbf{M}_B \succeq 0$, and so that $(1+\tilde{\epsilon})\mathbf{\Sigma} \succeq \mathbf{M}_G \succeq (1-\tilde{\epsilon})\mathbf{\Sigma}$. Following notation of Algorithm 4, let*

$$\mathbf{M} = \sum_{j\in[d]} \lambda_j v_j v_j^\top,\ \mathbf{\Sigma} = \sum_{j\in[d]} \sigma_j u_j u_j^\top \tag{8}$$

be sorted eigendecompositions of $\mathbf{M}$ and $\mathbf{\Sigma}$, so $\lambda_1 \geq \ldots \geq \lambda_d$, and $\sigma_1 \geq \ldots \geq \sigma_d$. Let γ be as in Theorem 2, and assume $\sigma_{t+1} < (1-\gamma)\sigma_1$. Then,

$$\max_{j\in[t]} v_j^\top \mathbf{\Sigma} v_j \geq (1-\gamma)\|\mathbf{\Sigma}\|_\infty .$$

With Proposition 4 in place, the recovery bound of Theorem 2 follows from an exact SVD. We show in Appendix D that the method is robust to approximations of the form (6), yielding our final claim.

Broader Impact

Our work provides frameworks for learning properties about the covariance of sub-Gaussian distributions which have been corrupted under noise. As a key subroutine, we develop solvers for smoothed positive linear and semidefinite programs. We believe these results are interesting from an academic perspective, e.g. our techniques may be applicable generally for robust statistics and convex optimization researchers. Moreover, because our primary results concern robustness of models to arbitrarily corrupted data, we believe our methods may have practical implications for downstream tasks where protection against a malicious adversary is warranted. Similarly, as our main subroutine is a solver attaining strong computational guarantees for a wider variety of objectives than was previously known, it is possible that our methods can be leveraged to broaden the types of downstream tasks that can be performed. Namely, as ℓ_p norm packing linear program solvers have found applications in fair resource allocation, our hope is that our smoothed and mixed-norm guarantee semidefinite solvers can find similar applications in learning algorithms for objectives designed with fairness or privacy in mind.

References

[AK16] Sanjeev Arora and Satyen Kale. A combinatorial, primal-dual approach to semidefinite programs. *J. ACM*, 63(2):12:1–12:35, 2016.

[AL17] Zeyuan Allen-Zhu and Yuanzhi Li. Follow the compressed leader: Faster online learning of eigenvectors and faster MMWU. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 116–125, 2017.

[ALO15] Zeyuan Allen Zhu, Zhenyu Liao, and Lorenzo Orecchia. Spectral sparsification and regret minimization beyond matrix multiplicative updates. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 237–245, 2015.

[ALO16] Zeyuan Allen Zhu, Yin Tat Lee, and Lorenzo Orecchia. Using optimization to obtain a width-independent, parallel, simpler, and faster positive SDP solver. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1824–1831, 2016.

[CDG19] Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2755–2771, 2019.

[CDGW19] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David Woodruff. Faster algorithms for high-dimensional robust covariance estimation. *arXiv preprint arXiv:1906.04661*, 2019.

[CDST19] Yair Carmon, John C. Duchi, Aaron Sidford, and Kevin Tian. A rank-1 sketch for matrix multiplicative weights. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 589–623, 2019.

[CG18] Yu Cheng and Rong Ge. Non-convex matrix completion against a semi-random adversary. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pages 1362–1394, 2018.

[CLMW11] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[CMY20] Yeshwanth Cherapanamjeri, Sidhanth Mohanty, and Morris Yau. List decodable mean estimation in nearly linear time. *CoRR*, abs/2005.09796, 2020.

[DFO18] Jelena Diakonikolas, Maryam Fazel, and Lorenzo Orecchia. Width-independence beyond linear objectives: Distributed fair packing and covering algorithms. *CoRR*, abs/1808.02517, 2018.

[DG03] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

[DHL19] Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In *Advances in Neural Information Processing Systems*, pages 6065–6075, 2019.

[DK19] Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

[DKK+17] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 999–1008. JMLR. org, 2017.

[DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM, 2018.

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.*, 48(2):742–864, 2019.

[DKS17] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

[DL19] Jules Despersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *CoRR*, abs/1906.03058, 2019.

[GHM15] Dan Garber, Elad Hazan, and Tengyu Ma. Online learning of eigenvectors. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 560–568, 2015.

[HL18] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

[Hub64] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[JLL+20] Arun Jambulapati, Yin Tat Lee, Jerry Li, Swati Padmanabhan, and Kevin Tian. Positive semidefinite programming: Mixed, parallel, and width-independent. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 2020.

[JY11] Rahul Jain and Penghui Yao. A parallel approximation algorithm for positive semidefinite programming. In *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 463–471, 2011.

[KSKO20] Weihao Kong, Raghav Somani, Sham M. Kakade, and Sewoong Oh. Robust meta-learning for mixed linear regression with small batches. *CoRR*, abs/2006.09702, 2020.

[KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

[Li18] Jerry Zheng Li. *Principled approaches to robust machine learning and beyond*. PhD thesis, Massachusetts Institute of Technology, 2018.

[LRV16] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

[MM15] Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1396–1404, 2015.

[MRWZ16] Michael W. Mahoney, Satish Rao, Di Wang, and Peng Zhang. Approximating the solution to mixed packing and covering lps in parallel o~(epsilonˆ{-3}) time. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 52:1–52:14, 2016.

[MSZ16] Jelena Marasevic, Clifford Stein, and Gil Zussman. A fast distributed stateless algorithm for alpha-fair packing problems. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 54:1–54:15, 2016.

[NN94] Yurii Nesterov and Arkadi Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

[PTZ16] Richard Peng, Kanat Tangwongsan, and Peng Zhang. Faster and simpler width-independent parallel algorithms for positive semidefinite programming. *CoRR*, abs/1201.5135, 2016.

[RH17] Philippe Rigollet and Jan-Christian Hütter. *High-Dimensional Statistics*. 2017.

[SCV17] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *arXiv preprint arXiv:1703.04940*, 2017.

[Ste18] Jacob Steinhardt. *Robust Learning: Information Theory and Algorithms*. PhD thesis, Stanford University, 2018.

[Tuk75] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

[Ver16] Roman Vershynin. *High-Dimensional Probability, An Introduction with Applications in Data Science*. 2016.

[WK06] Manfred K. Warmuth and Dima Kuzmin. Randomized PCA algorithms with regret bounds that are logarithmic in the dimension. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 1481–1488, 2006.

A Concentration

A.1 Sub-Gaussian concentration

We use the following concentration facts on sub-Gaussian distributions following from standard techniques, and give an application bounding Schatten-norm deviations.

Lemma 7. *Under Assumption 1, there are universal constants C_1, C_2 such that*

$$\Pr\left[\sup_{\substack{v\in\mathbb{R}^d\\ \|v\|_2=1}}\left|v^\top\left(\frac{1}{n}\sum_{i\in G'}X_iX_i^\top-\mathbf{\Sigma}\right)v\right|-tv^\top\mathbf{\Sigma}v>0\right]\le\exp\left(C_1d-C_2n\min(t,t^2)\right).$$

Proof. By observing (3), it is clear that the random vector $\widetilde{X}=\mathbf{\Sigma}^{-\frac{1}{2}}X$ for $X\sim\mathcal{D}$ has covariance $\mathbf{I}$ and sub-Gaussian proxy $c\mathbf{I}$. For any fixed unit vector u, by Lemma 1.12 of [RH17], the random variable $(u^\top\widetilde{X})^2-1$ is sub-exponential with parameter $16c$, so by Bernstein's inequality (Theorem 1.13, [RH17]), defining $\widetilde{X}_i=\mathbf{\Sigma}^{-\frac{1}{2}}X_i$ for each $X_i\sim\mathcal{D}$,

$$\Pr\left[\left|u^\top\left(\frac{1}{n}\sum_{i\in G'}\widetilde{X}_i\widetilde{X}_i^\top-\mathbf{I}\right)u\right|>\frac{t}{2}\right]\le\exp\left(-\frac{n}{2^{11}c^2}\min(t,t^2)\right).$$

For shorthand define $\mathbf{M}:=\frac{1}{n}\sum_{i\in G'}\widetilde{X}_i\widetilde{X}_i^\top$, and let $\mathcal{N}$ be a maximal $\frac{1}{4}$-net of the unit ball (as measured in ℓ_2 distance). By Lemma 1.18 of [RH17], $|\mathcal{N}|\le 12^d$, so by a union bound,

$$\Pr\left[\sup_{u\in\mathcal{N}}\left|u^\top(\mathbf{M}-\mathbf{I})u\right|>\frac{t}{2}\right]\le\exp\left(3d-\frac{n}{2^{11}c^2}\min(t,t^2)\right).$$

Next, by a standard application of the triangle inequality (see e.g. Exercise 4.3.3, [Ver16])

$$\sup_{\substack{v\in\mathbb{R}^d\\ \|v\|_2=1}}\left|v^\top(\mathbf{M}-\mathbf{I})v\right|\le 2\sup_{u\in\mathcal{N}}\left|u^\top(\mathbf{M}-\mathbf{I})u\right|\le t$$

with probability at least $1-\exp\left(C_1d-C_2n\min(t,t^2)\right)$ for appropriate C_1, C_2. The conclusion follows since its statement is scale invariant, so it suffices to show as we have that

$$\Pr\left[\sup_{\substack{v\in\mathbb{R}^d\\ \|v\|_{\mathbf{\Sigma}}=1}}\left|v^\top\left(\frac{1}{n}\sum_{i\in G'}X_iX_i^\top-\mathbf{\Sigma}\right)v\right|-tv^\top\mathbf{\Sigma}v>0\right]\le\exp\left(C_1d-C_2n\min(t,t^2)\right).$$

□

Corollary 2. *Let $p\ge 2$. Under Assumption 1, there are universal constants C_1, C_2 with*

$$\Pr\left[\left\|\frac{1}{n}\sum_{i\in G'}X_iX_i^\top-\mathbf{\Sigma}\right\|_p>t\left\|\mathbf{\Sigma}\right\|_p\right]\le\exp\left(C_1d-C_2n\min(t,t^2)\right).$$

Proof. Suppose the event in Lemma 7 does not hold, which happens with probability at least $1-\exp(C_1d-C_2n\min(t,t^2))$. Define for shorthand $\mathbf{M}:=\frac{1}{n}\sum_{i\in G'}X_iX_i^\top-\mathbf{\Sigma}$ and let its spectral decomposition be $\sum_{j\in[d]}\lambda_jv_jv_j^\top$. By the triangle inequality and Fact 2,

$$\begin{aligned}\|\mathbf{M}\|_p&\le\sum_{j\in[d]}\frac{|\lambda_j|^{p-1}}{\|\mathbf{M}\|_p^{p-1}}\left|v_j^\top\left(\frac{1}{n}\sum_{i\in G'}X_iX_i^\top-\mathbf{\Sigma}\right)v_j\right|\\&\le t\sum_{j\in[d]}\frac{|\lambda_j|^{p-1}}{\|\mathbf{M}\|_p^{p-1}}v_j^\top\mathbf{\Sigma}v_j=t\left\langle\sum_{j\in[d]}\frac{|\lambda_j|^{p-1}}{\|\mathbf{M}\|_p^{p-1}}v_jv_j^\top,\mathbf{\Sigma}\right\rangle\le t\left\|\mathbf{\Sigma}\right\|_p.\end{aligned}$$

In the last inequality, we used that $\sum_{j\in[d]}\frac{|\lambda_j|^{p-1}}{\|\mathbf{M}\|_p^{p-1}}v_jv_j^\top$ has unit ℓ_q norm, and applied Fact 2. □

A.2 Concentration under weightings in $\mathfrak{S}_\epsilon^n$

We consider concentration of the empirical covariance under weightings which are not far from uniform, in spectral and Schatten senses.

Lemma 8. *Under Assumption 1, let $\delta \in [0,1]$, $p \geq 2$, and $n = \Omega\left(\frac{d+\log \delta^{-1}}{(\epsilon \log \epsilon^{-1})^2}\right)$ for a sufficiently large constant. Then for a universal constant C_3,*

$$\Pr\left[\exists w \in \mathfrak{S}_\epsilon^n \;\middle|\; \left\|\sum_{i\in G'} w_i X_i X_i^\top - \mathbf{\Sigma}\right\|_p > C_3 \cdot \epsilon \log \epsilon^{-1} \|\mathbf{\Sigma}\|_p\right] \leq \frac{\delta}{2}.$$

Proof. Because the vertices of $\mathfrak{S}_\epsilon^n$ are uniform over sets $S \subseteq G'$ with $|S| = (1-\epsilon)n$ (see e.g. Section 4.1, [DKK+19]), by convexity of the Schatten-p norm it suffices to prove

$$\Pr\left[\exists S \text{ with } |S| = (1-\epsilon)n \;\middle|\; \left\|\frac{1}{(1-\epsilon)n}\sum_{i\in S} X_i X_i^\top - \mathbf{\Sigma}\right\|_p > C_3 \cdot \epsilon \log \epsilon^{-1} \|\mathbf{\Sigma}\|_p\right] \leq \frac{\delta}{4}.$$

For any fixed S, and recalling $|S^c| = \epsilon n$, we can decompose this sum as

$$\frac{1}{(1-\epsilon)n}\sum_{i\in S} X_i X_i^\top = \frac{1}{1-\epsilon}\left(\frac{1}{n}\sum_{i\in G'} X_i X_i^\top\right) - \frac{\epsilon}{1-\epsilon}\left(\frac{1}{|S^c|}\sum_{i\in S^c} X_i X_i^\top\right). \tag{9}$$

By applying Corollary 2, it follows that by setting $t = \frac{1-\epsilon}{2} \cdot \epsilon \log \epsilon^{-1}$ and our choice of n that

$$\Pr\left[\left\|\frac{1}{n}\sum_{i\in G'} X_i X_i^\top - \mathbf{\Sigma}\right\|_p > \frac{1-\epsilon}{2} \cdot \epsilon \log \epsilon^{-1} \|\mathbf{\Sigma}\|_p\right] \leq \frac{\delta}{4}. \tag{10}$$

Moreover, for any fixed S^c, setting $t = \frac{1-\epsilon}{2} \cdot C_3 \log \epsilon^{-1}$ where C_3 is a sufficiently large constant, so that for sufficiently small ϵ, $t = \min(t, t^2)$,

$$\begin{aligned}\Pr\left[\left\|\frac{1}{\epsilon n}\sum_{i\in S^c} X_i X_i^\top - \mathbf{\Sigma}\right\|_p > \frac{1-\epsilon}{2} \cdot C_3 \cdot \log \epsilon^{-1} \|\mathbf{\Sigma}\|_p\right] &\leq \exp\left(C_1 d - C_2 \epsilon n t\right) \\ &\leq \exp\left(-\left(\log \delta^{-1} + n\epsilon \log \epsilon^{-1}\right)\right) \\ &\leq \frac{\delta}{4\binom{n}{\epsilon n}}.\end{aligned} \tag{11}$$

Here, we used that $\log \binom{n}{\epsilon n} = O\left(n\epsilon \log \epsilon^{-1}\right)$. Finally, union bounding over all possible sets S^c imply that with probability at least $1 - \frac{\delta}{2}$, the following events hold:

$$\left\|\frac{1}{n}\sum_{i\in G'} X_i X_i^\top - \mathbf{\Sigma}\right\|_p < \frac{1-\epsilon}{2} \cdot \epsilon \log \epsilon^{-1} \|\mathbf{\Sigma}\|_p,$$

$$\left\|\frac{1}{|S^c|}\sum_{i\in S^c} X_i X_i^\top - \mathbf{\Sigma}\right\|_p < \frac{1-\epsilon}{2} \cdot C_3 \cdot \log \epsilon^{-1} \|\mathbf{\Sigma}\|_p \text{ for all } S \text{ with } |S| = (1-\epsilon)n.$$

Combining these bounds in the context of (9) after applying the triangle inequality, we have with probability at least $1 - \frac{\delta}{2}$ for all S the desired conclusion,

$$\left\|\frac{1}{(1-\epsilon)n}\sum_{i\in S} X_i X_i^\top - \mathbf{\Sigma}\right\|_p < C_3 \cdot \epsilon \log \epsilon^{-1} \|\mathbf{\Sigma}\|_p.$$

□

Corollary 3. *Under Assumption 1, let $n = \Omega\left(\frac{d+\log\delta^{-1}}{(\epsilon\log\epsilon^{-1})^2}\right)$ for a sufficiently large constant. For universal C_3 and all $w \in \mathfrak{S}_\epsilon^n$, with probability at least $1-\frac{\delta}{2}$,*

$$C_3 \cdot \epsilon\log\epsilon^{-1}\mathbf{\Sigma} \succeq \sum_{i\in G'} w_i X_i X_i^\top - \mathbf{\Sigma} \succeq -C_3\cdot\epsilon\log\epsilon^{-1}\mathbf{\Sigma}.$$

Proof. Consider any unit vector $v \in \mathbb{R}^d$. By similar arguments as in (10), (11), and applying a union bound over all S with $|S| = (1-\epsilon)n$, with probability at least $1-\frac{\delta}{2}$, it follows from Lemma 7 that

$$\left| v^\top \left(\frac{1}{n}\sum_{i\in G'} X_i X_i^\top - \mathbf{\Sigma}\right) v\right| < \frac{1-\epsilon}{2}\cdot\epsilon\log\epsilon^{-1} v^\top\mathbf{\Sigma}v, \tag{12}$$

$$\left| v^\top \left(\frac{1}{|S^c|}\sum_{i\in S^c} X_i X_i^\top - \mathbf{\Sigma}\right) v\right| < \frac{1-\epsilon}{2}\cdot C_3\cdot\log\epsilon^{-1} v^\top\mathbf{\Sigma}v\ . \tag{13}$$

Therefore, again using the formula (9) and the triangle inequality yields the desired conclusion for all directions v, which is equivalent to the spectral bound of the lemma statement. □

B Deferred proofs from Section 3

B.1 Robust univariate variance estimation

In this section, we prove Lemma 1, which allows us to robustly estimate the quadratic form of a vector in the covariance of a sub-Gaussian distribution from corrupted samples. Algorithm 5 is folklore, and intuitively very simple; it projects all samples onto u, throws away the 2ϵ fraction of points with largest magnitude in this direction, and takes the mean of the remaining set.

Algorithm 5 Univariate variance estimation: 1DRobustVariance$(\{X_i\}_{i\in[n]}, \epsilon, u)$

Input: $\{X_i\}_{i\in[n]}$, $\epsilon > 0$, and a unit vector u
Let $a_i = \langle X_i, u\rangle^2$ for $i = 1,\dots,n$
Sort the a_i in increasing order. WLOG assume $a_1 \le a_2 \le \ldots \le a_n$.
return $\sigma_u^2 = \frac{1}{(1-2\epsilon)n}\sum_{i=1}^{(1-2\epsilon)n} a_i$

We require the following helper fact.

Fact 3. *Let Z be a sub-exponential random variable with parameter at most λ[9], and let $\epsilon \in [0,1]$. Then, for any event E with $\Pr[Z\in E] \le \epsilon$, $|\mathbb{E}\left[Z\cdot\mathbf{1}[Z\in E]\right]| \le 2\lambda\epsilon\log\epsilon^{-1}$.*

Proof. We have by Hölder's inequality that for any $p, q \ge 1$ with $p^{-1}+q^{-1} = 1$,

$$|\mathbb{E}\left[Z\cdot\mathbf{1}[Z\in E]\right]| \le \mathbb{E}[|Z|^p]^{1/p}\cdot\epsilon^{1/q} \le 2\lambda p\cdot\epsilon^{1/q}.$$

The second inequality is Lemma 1.10 [RH17]. Setting $p = \log\epsilon^{-1}$ yields the result. □

Lemma 1. *Under Assumption 1, let $\delta \in [0,1]$, $n = \Omega\left(\frac{\log\delta^{-1}}{(\epsilon\log\epsilon^{-1})^2}\right)$, and $u \in \mathbb{R}^d$ be a fixed unit vector. Algorithm 5,* 1DRobustVariance*, takes input $\{X_i\}_{i\in[n]}$, u, and ϵ, and outputs σ_u^2 with $|u^\top\mathbf{\Sigma}u - \sigma_u^2| < Cu^\top\mathbf{\Sigma}u\cdot\epsilon\log\epsilon^{-1}$ with probability at least $1-\delta$, and runs in time $O(nd + n\log n)$, for C a fixed multiple of the parameter c in Assumption 1.*

Proof. The runtime claim is immediate; we now turn our attention to correctness. We follow notation of Assumption 1, and in a slight abuse of notation, also define $a_i = \langle X_i, u\rangle^2$ for $i \in G'$. First, for $X\sim\mathcal{D}$, then $\langle u, X\rangle^2 - u^\top\mathbf{\Sigma}u$ is sub-exponential with parameter at most $16cu^\top\mathbf{\Sigma}u$ (Lemma 1.12, [RH17]). By Bernstein's inequality, we have that if $X\sim\mathcal{D}$, then for all $t \ge 1$,

$$\Pr\left[\langle X,u\rangle^2 > 32ctu^\top\mathbf{\Sigma}u\right] \le \exp(-t)\ . \tag{14}$$

[9] We say mean-zero Z is sub-exponential with parameter λ if $\forall|s| \le \lambda^{-1}$, $\mathbb{E}[\exp(sZ)] \le \exp(\frac{s^2\lambda^2}{2})$.

Using this in a standard Chernoff bound, we have that with probability $1 - \frac{\delta}{2}$,

$$\frac{\left|\{i \in G' : a_i > 64c \log \epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u\}\right|}{n} \leq \epsilon \,. \tag{15}$$

Let $T = 64c \log \epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u$, and let Y be distributed as $(\langle u, X\rangle^2 - u^\top \mathbf{\Sigma} u) \cdot \mathbf{1}[\langle u, X\rangle^2 \leq T]$, where $X \sim \mathcal{D}$. We observe $Y - \mathbb{E}[Y]$ is also sub-exponential with parameter $16cu^\top \mathbf{\Sigma} u$, and that by Fact 3,

$$|\mathbb{E}[Y]| \leq 32cu^\top \mathbf{\Sigma} u \epsilon \log \epsilon^{-1}. \tag{16}$$

Define the interval $I = [0, T]$ and let S be the set of points in $[n]$ that survive the truncation procedure, so that $\sigma_u^2 = \frac{1}{|S|}\sum_{i \in S} a_i$. Given event (15), $a_i \in I$ for all $i \in S$, since there are at most ϵn points in G outside I, and $|B| \leq \epsilon n$. We decompose the deviation as follows:

$$\begin{aligned}
\sum_{i \in S} a_i - |S| u^\top \mathbf{\Sigma} u &= \sum_{i \in G \cap S} (a_i - u^\top \mathbf{\Sigma} u) + \sum_{i \in B \cap S} (a_i - u^\top \mathbf{\Sigma} u) \\
&= \sum_{i \in G' \cap I} (a_i - u^\top \mathbf{\Sigma} u) + \sum_{i \in B \cap S} (a_i - u^\top \mathbf{\Sigma} u) \\
&\quad - \sum_{i \in (G' \setminus G) \cap I} (a_i - u^\top \mathbf{\Sigma} u) - \sum_{i \in (G \cap I) \setminus S} (a_i - u^\top \mathbf{\Sigma} u).
\end{aligned} \tag{17}$$

Here we overloaded $i \in I$ to mean that a_i lies in the interval I, and conditioned on S lying entirely in I. We bound each of these terms individually. First, for all $i \in G' \cap I$, conditioning on (15) (i.e. all $a_i \in I$), $a_i - u^\top \mathbf{\Sigma} u$ is an independent sample from Y. Thus, by (16) and Bernstein's inequality,

$$\begin{aligned}
\left| \frac{1}{|G' \cap I|} \sum_{i \in G' \cap I} (a_i - u^\top \mathbf{\Sigma} u) \right| &\leq \left| \frac{1}{|G' \cap I|} \sum_{i \in G' \cap I} (a_i - u^\top \mathbf{\Sigma} u) - \mathbb{E}[Y] \right| + 32cu^\top \mathbf{\Sigma} u \epsilon \log \epsilon^{-1} \\
&\leq 64c \cdot u^\top \mathbf{\Sigma} u \epsilon \log \epsilon^{-1},
\end{aligned} \tag{18}$$

with (conditional) probability at least $1 - \frac{\delta}{2}$. By a union bound, both events occur with probability at least $1 - \delta$; condition on this for the remainder of the proof. Under this assumption, we control the other three terms of (17). Observe that $|B \cap S| \leq \epsilon n$, $|(G' \setminus G) \cap I| \leq \epsilon n$, and $|(G \cap I) \setminus S| \leq \epsilon n$. Further, by definition of I, every summand is at most $64c \log \epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u$. Thus,

$$\left| \sum_{i \in B \cap S} (a_i - u^\top \mathbf{\Sigma} u) \right| \leq 64c\epsilon n \log \epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u, \tag{19}$$

$$\left| \sum_{i \in (G' \setminus G) \cap I} (a_i - u^\top \mathbf{\Sigma} u) \right| \leq 64c\epsilon n \log \epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u. \tag{20}$$

$$\left| \sum_{i \in (G' \cap I) \setminus S} (a_i - u^\top \mathbf{\Sigma} u) \right| \leq 64c\epsilon n \log \epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u. \tag{21}$$

Combining (18), (19), (20), and (21) in derivation (17) and dividing by $|S|$ yields the claim. □

Finally, we also give an alternative set of conditions under which we can certify correctness of 1DRobustVariance. Specifically, this assumption will be useful in lifting indpendence assumptions between u and our samples $\{X_i\}_{i \in [n]}$ in repeated calls within Algorithm 6.

Assumption 2. *Under Assumption 1, let the following conditions hold for universal constant C_4:*

$$C_4 \epsilon \log \epsilon^{-1} \cdot \mathbf{\Sigma} \succeq \frac{1}{n} \sum_{i \in G'} X_i X_i^\top - \mathbf{\Sigma} \succeq -C_4 \epsilon \log \epsilon^{-1} \cdot \mathbf{\Sigma}, \tag{22}$$

$$C_4 \log \epsilon^{-1} \cdot \mathbf{\Sigma} \succeq \sum_{i \in G'} w_i \left(X_i X_i^\top - \mathbf{\Sigma} \right) \succeq -C_4 \log \epsilon^{-1} \cdot \mathbf{\Sigma} \textit{ for all } w \in \mathfrak{S}_{1-\epsilon}^n. \tag{23}$$

Note that (23) is a factor ϵ weaker in its guarantee than Corollary 3, and is over weights in a different set $\mathfrak{S}_{1-\epsilon}^n$. Standard sub-Gaussian concentration (i.e. an unweighted variant of Corollary 3) and modifying the proof of Corollary 3 to take the constraint set $\mathfrak{S}_{1-\epsilon}^n$ and normalizing over vertex sets of size ϵn yield the following conclusion.

Lemma 9. *Let* $n = \Omega\left(\frac{d+\log\delta^{-1}}{(\epsilon\log\epsilon^{-1})^2}\right)$ *for a sufficiently large constant. Assumption 2 holds with probability at least* $1 - \frac{\delta}{2}$.

We give a variant of Lemma 1 with slightly stronger guarantees for 1DRobustVariance; specifically, it holds for all u simultaneously for a fixed set of samples satisfying Assumption 2.

Corollary 4. *Under Assumption 2, Algorithm 5 outputs* σ_u^2 *with* $|u^\top\mathbf{\Sigma}u - \sigma_u^2| < Cu^\top\mathbf{\Sigma}u \cdot \epsilon\log\epsilon^{-1}$, *for* C *a fixed multiple of the parameter* c *in Assumption 1, and runs in time* $O(nd + n\log n)$.

Proof. We discuss how to modify the derivations from Lemma 1 appropriately in the absence of applications of Bernstein's inequality. First, note that appropriately combining (22) and (23) in a derivation such as (9) yields the following bound (deterministically under Assumption 2):

$$C_4\epsilon\log\epsilon^{-1}\cdot\mathbf{\Sigma} \succeq \sum_{i\in G'} w_i\left(X_iX_i^\top - \mathbf{\Sigma}\right) \succeq -C_4\epsilon\log\epsilon^{-1}\mathbf{\Sigma} \text{ for all } w \in \mathfrak{S}_{3\epsilon}^n. \tag{24}$$

Now, consider the decomposition (17). We claim first that similarly to (19), (20), (21) we can bound each summand in the latter three terms by $O(u^\top\mathbf{\Sigma}u\log\epsilon^{-1})$; to prove this, it suffices to show that at least one filtered a_i attains this bound, as then by definition of the algorithm, each non-filtered a_i will as well. Note that a fraction between ϵ and 2ϵ of points in $G \subset G'$ is filtered (since there are only ϵn points from B). The assumption (23) then implies precisely the desired bound on some filtered a_i by placing uniform mass on filtered points from G, and applying pigeonhole. So, all non-filtered a_i are bounded by $O(u^\top\mathbf{\Sigma}u\log\epsilon^{-1})$, yielding analogous statements to (19), (20), (21).

Finally, an analogous derivation to (18) follows via an application of the bound (24), where we place uniform mass on the set $G' \cap I$ and adjust constants appropriately, since the above argument shows that under the assumption (23), we have that at most $2\epsilon n$ indices $i \in G'$ have $a_i \notin I$. □

B.2 Preliminaries

For convenience, we give the following preliminaries before embarking on our proof of Theorem 1 and giving guarantees on Algorithm 6. First, we state a set of assumptions which augments Assumption 2 with one additional condition, used in bounding the iteration count of our algorithm.

Assumption 3. *Under Assumption 1, let Assumption 2 hold, as well as the following additional condition for the same universal constant* C_4*:*

$$\|X_i\|_2^2 \le C_4\log\frac{n}{\delta}\cdot\mathrm{Tr}(\mathbf{\Sigma}) \text{ for all } i \in G. \tag{25}$$

Standard sub-Gaussian concentration inequalities and a union bound, combined with our earlier claim Lemma 9, then yield the following guarantee.

Lemma 10. *Let* $n = \Omega\left(\frac{d+\log\delta^{-1}}{(\epsilon\log\epsilon^{-1})^2}\right)$ *for a sufficiently large constant. Assumption 3 holds with probability at least* $1 - \delta$.

B.3 Analysis of PCAFilter

For this section, for any nonnegative weights w, define $\mathbf{M}(w) := \sum_{i\in[n]} w_iX_iX_i^\top$. We now state our algorithm, PCAFilter. At all iterations t, it maintains a current nonnegative weight vector $w^{(t)}$ (initialized to be the uniform distribution on $[n]$), preserving the following invariants for all t:

$$w_i^{(t-1)} \ge w_i^{(t)} \text{ for all } i \in [n],\ \sum_{i\in B} w_i^{(t-1)} - w_i^{(t)} \ge \sum_{i\in G} w_i^{(t-1)} - w_i^{(t)}. \tag{26}$$

We now state our method as Algorithm 6; note that the update to $w^{(t)}$ is of the form in Lemma 2.

Algorithm 6 $\mathsf{PCAFilter}(\{X_i\}_{i\in[n]}, \epsilon)$

1: Remove all points $i \in [n]$ with $\|X_i\|_2^2 > c\log(\frac{n}{\delta}) \cdot \mathrm{Tr}(\mathbf{\Sigma})$
2: $w_i^{(0)} \leftarrow \frac{1}{n}$ for all $i \in [n]$, $t \leftarrow 1$
3: $u_1 \leftarrow$ approximate top eigenvector of $\mathbf{M}(w^{(0)})$
4: $\sigma_1^2 \leftarrow \mathsf{1DRobustVariance}(\{X_i\}_{i\in[n]}, \epsilon, u_1)$
5: **while** $u_t^\top \mathbf{M}(w^{(t-1)})u_t > (1 + 5C_5\epsilon\log\epsilon^{-1})\sigma_t^2$, where $C_5 = \max(C, C_4)$ from constants in Assumption 2, Corollary 4 **do**
6: $\quad a_i \leftarrow \langle u_t, X_i\rangle^2$ for all $i \in [n]$
7: $\quad$ Sort (permute) the indices $[n]$ so the a_i are in increasing order (with a_1 smallest, a_n largest)
8: $\quad$ Let ℓ be the largest index with $\sum_{i=\ell}^n w_i \geq 2\epsilon$
9: $\quad$ Define

$$w_i^{(t)} \leftarrow \begin{cases} \left(1 - \frac{a_i}{a_n}\right) w_i^{(t-1)} & \ell \leq i \leq n \\ w_i^{(t-1)} & i < \ell \end{cases}$$

10: $\quad u_t \leftarrow$ approximate top eigenvector of $\mathbf{M}(w^{(t)})$
11: $\quad \sigma_t^2 \leftarrow \mathsf{1DRobustVariance}(\{X_i\}_{i\in[n]}, u_t, \epsilon)$
12: $\quad t \leftarrow t + 1$
13: **end while**
14: **return** u_t

We assume that in Line 8, we also have $\sum_{i=\ell}^n w_i \leq 3\epsilon$, as we can assume at least one point is corrupted i.e. $\epsilon \geq \frac{1}{n}$ (else standard algorithms suffice for our setting), so adding an additional w_i can only change the sum by ϵ. We first prove invariants (26) are preserved; at a high level, we simply demonstrate that Lemma 2 holds via concentration on G and lack of termination.

Lemma 11. *Under Assumption 2, for any iteration t of Algorithm 6, suppose* (26) *held for all iterations $t' \leq t-1$. Then,* (26) *holds at iteration t.*

Proof. The first part of (26) is immediate by observing the update in Line 9, so we show the second. We drop subscripts and superscripts for conciseness and focus on a single iteration t. Let $I_B = \{\ell, \ldots, n\} \cap B$, and $I_G = \{\ell, \ldots, n\} \cap G$. By Lemma 2, it suffices to demonstrate that

$$\sum_{i\in I_B} w_i a_i > \sum_{i\in I_G} w_i a_i. \tag{27}$$

First, $\sum_{i\in I_B} w_i \leq \epsilon$, so by definition of index ℓ, we have $\epsilon \leq \sum_{i\in I_G} w_i \leq 2\epsilon$. Define $\widetilde{w}_i = \frac{w_i}{\sum_{i\in I_G} w_i}$ if $i \in I_G$, and 0 otherwise, and observe $\widetilde{w} \in \mathfrak{S}_{1-2\epsilon}^n$. By modifying constants appropriately from (23), it follows from definition of $a_i = u^\top X_i X_i^\top u$ that

$$\sum_{i\in I_G} w_i a_i \leq \left(\sum_{i\in I_G} w_i\right) \cdot C_4 \log\epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u \leq 2C_4\epsilon\log\epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u. \tag{28}$$

On the other hand, by (24) we know that the total quadratic form over G is bounded as

$$\sum_{i\in G} w_i a_i < \left(\sum_{i\in G} w_i\right)\left(1 + C_4\epsilon\log\epsilon^{-1}\right) u^\top \mathbf{\Sigma} u < \left(1 + C_4\epsilon\log\epsilon^{-1}\right) u^\top \mathbf{\Sigma} u. \tag{29}$$

Here, we applied the observation that the normalized w_i restricted to G are in $\mathfrak{S}_{1-3\epsilon}^n$ (e.g. using Lemma 12 inductively). However, since we did not terminate (Line 5), we must have by u_t being a top eigenvector and Corollary 4 (we defer discussions of inexactness to Theorem 1) that

$$\sum_{i\in[n]} w_i a_i \geq (1 + 5C_5\epsilon\log\epsilon^{-1})\sigma_t^2 \geq (1 + 4C_4\epsilon\log\epsilon^{-1}) \cdot u^\top \mathbf{\Sigma} u$$

$$\implies \sum_{i\in B} w_i a_i > 3C_4\epsilon\log\epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u.$$

To obtain the last conclusion, we used (29). Finally, note that for all $i \in B \setminus I_B$,

$$a_i \le a_\ell \le \sum_{i \in I_G} \widetilde{w}_i a_i \le C_4 \log \epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u$$

by rearranging (28). This implies that

$$\sum_{i \in B \setminus I_B} w_i a_i \le \left(\sum_{i \in B \setminus I_B} w_i \right) \cdot C_4 \log \epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u \le C_4 \epsilon \log \epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u.$$

Thus, the desired inequality (27) follows from combining the above derivations, e.g. using (28) and

$$\sum_{i \in I_B} w_i a_i = \sum_{i \in B} w_i a_i - \sum_{i \in B \setminus I_B} w_i a_i > 2C_4 \epsilon \log \epsilon^{-1} \cdot u^\top \mathbf{\Sigma} u.$$

□

Lemma 11 yields for all t that $\sum_{i \in B} w_i^{(0)} - w_i^{(t)} \ge \sum_{i \in G} w_i^{(0)} - w_i^{(t)}$ by telescoping. Note that we can only remove at most 2ϵ mass from w total, as $\sum_{i \in B} w_i^{(0)} - w_i^{(t)} \le \epsilon$. Denote for shorthand normalized weights $v^{(t)} := \frac{w^{(t)}}{\left\| w^{(t)} \right\|_1}$. Then, the following is immediate by $\left\| w^{(t)} \right\|_1 \ge 1 - 2\epsilon$.

Lemma 12. *Under Assumption 2, in all iterations t of Algorithm 6, $v^{(t)} \in \mathfrak{S}_{2\epsilon}^n$.*

Using Lemma 12, we show that the output has the desired quality of being a large eigenvector.

Lemma 13. *Under Assumption 2, let the output of Algorithm 6 be u_T. Then for a universal constant $C^\star$, $u_T^\top \mathbf{\Sigma} u_T \ge (1 - C^\star \epsilon \log \epsilon^{-1}) \|\mathbf{\Sigma}\|_\infty$.*

Proof. We assume for now that u_T is an exact top eigenvector, and discuss inexactness while proving Theorem 1. By (24) and Lemma 12, as then the normalized restriction of $w^{(T)}$ to G is in $\mathfrak{S}_{3\epsilon}^n$,

$$\mathbf{M}(w^{(T)}) \succeq \sum_{i \in G} w_i^{(T)} X_i X_i^\top \succeq \left(1 - 2C_4 \epsilon \log \epsilon^{-1}\right) \mathbf{\Sigma}$$
$$\implies u_T^\top \mathbf{M}(w^{(T)}) u_T \ge \left(1 - 2C_4 \epsilon \log \epsilon^{-1}\right) \|\mathbf{\Sigma}\|_\infty .$$

We used the Courant-Fischer characterization of eigenvalues, and that u_T is a top eigenvector of $\mathbf{M}(w^{(T)})$. Moreover, by termination conditions and Corollary 4 (correctness of 1DRobustVariance),

$$(1 + C\epsilon \log \epsilon^{-1}) u_T^\top \mathbf{\Sigma} u_T \ge \sigma_T^2 \ge (1 + 5C_5 \epsilon \log \epsilon^{-1})^{-1} u_T^\top \mathbf{M}(w^{(T)}) u_T.$$

Combining these two bounds and rescaling yields the conclusion. □

Finally, we prove our main guarantee about Algorithm 6.

Theorem 1. *Under Assumption 1, let $\delta \in [0, 1]$, and $n = \Omega\left(\frac{d + \log \delta^{-1}}{(\epsilon \log \epsilon^{-1})^2}\right)$. Algorithm 6 runs in time $O(\frac{nd^2}{\epsilon} \log \frac{n}{\delta\epsilon} \log \frac{n}{\delta})$, and outputs u with $u^\top \mathbf{\Sigma} u > (1 - C^\star \epsilon \log \epsilon^{-1}) \|\mathbf{\Sigma}\|_\infty$, for $C^\star$ a fixed multiple of parameter c in Assumption 1, with probability at least $1 - \delta$.*

Proof. First, we will operate under Assumption 3, which holds with probability at least $1 - \delta$. It is clear that the analyses of Lemma 11 and 13 hold with $1 - \Theta(\epsilon \log \epsilon^{-1})$ multiplicative approximations of top eigenvector computation, which the power method approximates with high probability. Thus, each iteration takes time $O\left(\frac{nd}{\epsilon} \log \frac{n}{\delta\epsilon}\right)$, where we will union bound over the number of iterations.

We now give an iteration bound: in any iteration where we do not terminate, Lemma 2 implies

$$\begin{aligned}\sum_{i=1}^{n} w_i^{(t-1)} - w_i^{(t)} &\geq \frac{1}{2\max_{i\in[n]} \langle u_t, X_i\rangle^2} \sum_{i=\ell}^{n} w_i a_i \\ &\geq \frac{1}{2C_4 \log\frac{n}{\delta} \cdot \mathrm{Tr}(\mathbf{\Sigma})} \sum_{i=\ell}^{n} w_i a_i \\ &\geq \frac{1}{2C_4 \log\frac{n}{\delta} \cdot \mathrm{Tr}(\mathbf{\Sigma})} \left(\frac{\sum_{i=\ell}^{n} w_i}{\sum_{i\in[n]} w_i}\right) \sum_{i\in[n]} w_i a_i \\ &= \Omega\left(\epsilon \cdot \frac{\|\mathbf{\Sigma}\|_\infty}{\log\frac{n}{\delta} \cdot \mathrm{Tr}(\mathbf{\Sigma})}\right) = \Omega\left(\frac{\epsilon}{d\log\frac{n}{\delta}}\right).\end{aligned}$$

Here, the second line used Assumption 3, the third used that the a_i are in sorted order, and the last used the definition of ℓ as well as the derivations of Lemma 13 (specifically, that $\mathbf{M}(w)$ spectrally dominates $(1 - O(\epsilon\log\epsilon^{-1}))\mathbf{\Sigma}$ for roughly uniform w). The conclusion follows since there can be at most $O(d\log\frac{n}{\delta})$ iterations, as the algorithm terminates when a 2ϵ fraction of the mass is removed, giving the overall runtime claim. □

C Deferred proofs from Section 4

C.1 Proofs from Section 4.2

Since our notion of approximation is multiplicative, we can assume without more than constant loss that $\mathbf{A}$ has bounded entries. This observation is standard, and formalized in the following lemma.

Lemma 14 (Entrywise bounds on $\mathbf{A}$)**.** *Feasibility of Problem 2 is unaffected (up to constants in ϵ) by removing columns of $\mathbf{A}$ with entries larger than $n\epsilon^{-1}$.*

Proof. If $\mathbf{A}_{ji} > n\epsilon^{-1}$ for any entry, then $x_i \leq \frac{\epsilon(1+\epsilon)}{n}$, else $\|\mathbf{A}x\|_p$ is already larger than $1+\epsilon$. Ignoring all such entries of x and rescaling can only change the objective by a $1 + O(\epsilon)$ factor. □

Lemma 3. *In all iterations t of Algorithm 2, defining $\Phi_t := \|\mathbf{A}w_t\|_p - \|w_t\|_1$, $\Phi_{t+1} \leq \Phi_t$.*

Proof. Fix an iteration t. Define $\delta = \eta g_t$, and note $w_{t+1} = w_t + \delta \circ w_t$; henceforth in this proof, we will drop subscripts t when clear. Observe that

$$\|\mathbf{A}w_{t+1}\|_p = \|\mathbf{A}((1+\delta)\circ w)\|_p = \left(\sum_{j\in[d]} [\mathbf{A}w]_j^p \left(1 + \frac{[\mathbf{A}(\delta\circ w_t)]_j}{[\mathbf{A}w_t]_j}\right)^p\right)^{1/p}.$$

As $g \leq \mathbf{1} \implies \delta \leq p^{-1}\mathbf{1}$, $\frac{\mathbf{A}(\delta\circ w_t)}{\mathbf{A}w_t} \leq p^{-1}$ entrywise. Via $(1+x)^p \leq \exp(px) \leq 1 + px + p^2x^2$ for $x \leq p^{-1}$, it follows that

$$\|\mathbf{A}((1+\delta)\circ w)\|_p \leq \left(\sum_{j\in[d]} [\mathbf{A}w]_j^p \left(1 + \frac{p[\mathbf{A}(\delta\circ w)]_j}{[\mathbf{A}w]_j} + \left(\frac{p[\mathbf{A}(\delta\circ w)]_j}{[\mathbf{A}w]_j}\right)^2\right)\right)^{1/p}.$$

By direct manipulation of the above quantity, and recalling we defined $v = \frac{\mathbf{A}w}{\|\mathbf{A}w\|_p}$,

$$\left(\sum_{j\in[d]}\left([\mathbf{A}w]_j^p + p[\mathbf{A}w]_j^{p-1}[\mathbf{A}(\delta\circ w)]_j + p^2[\mathbf{A}w]_j^{p-2}[\mathbf{A}(\delta\circ w)]_j^2\right)\right)^{1/p}$$
$$= \left(\|\mathbf{A}w\|_p^p \sum_{j\in[d]}\left(v_j^p + pv_j^{p-1}\frac{[\mathbf{A}(\delta\circ w)]_j}{\|\mathbf{A}w\|_p} + p^2v_j^{p-2}\left(\frac{[\mathbf{A}(\delta\circ w)]_j}{\|\mathbf{A}w\|_p}\right)^2\right)\right)^{1/p}$$
$$= \|\mathbf{A}w\|_p\left(1 + \sum_{j\in[d]}\left(pv_j^{p-1}\frac{[\mathbf{A}(\delta\circ w)]_j}{\|\mathbf{A}w\|_p} + p^2v_j^{p-2}\left(\frac{[\mathbf{A}(\delta\circ w)]_j}{\|\mathbf{A}w\|_p}\right)\right)^2\right)^{1/p}.$$

Using $(1+x)^p > 1+px$, i.e. $(1+px)^{1/p} < 1+x$, we thus obtain

$$\|\mathbf{A}((1+\delta)\circ w)\|_p \le \|\mathbf{A}w\|_p + \left\langle v^{p-1}, \mathbf{A}(\delta\circ w)\right\rangle + p\left\langle v^{p-1}, \frac{(\mathbf{A}(\delta\circ w))^2}{\mathbf{A}w}\right\rangle.$$

Cauchy-Schwarz yields that $[\mathbf{A}(\delta\circ w)]_j^2 \le [\mathbf{A}(\delta^2\circ w)]_j[\mathbf{A}w]_j, \forall j\in[d]$. Substituting into the above,

$$\begin{aligned}\|\mathbf{A}((1+\delta)\circ w)\|_p &\le \|\mathbf{A}w\|_p + \left\langle v^{p-1}, \mathbf{A}(\delta\circ w)\right\rangle + p\left\langle v^{p-1}, \mathbf{A}(\delta^2\circ w)\right\rangle \\ &= \|\mathbf{A}w\|_p + \sum_{j\in[d]}\left[\mathbf{A}^\top v^{p-1}\right]_j\delta_j w_j(1+p\delta_j). \end{aligned} \tag{30}$$

Finally, to bound this latter quantity, since $\delta = \eta g$, we observe that for all j either $\delta_j = 0$ or $1+p\delta_j = 1+g_j = 2-[\mathbf{A}^\top v^{p-1}]_j$, in which case

$$\left[\mathbf{A}^\top v^{p-1}\right]_j(1+p\delta_j) = \left[\mathbf{A}^\top v^{p-1}\right]_j\left(2-\left[\mathbf{A}^\top v^{p-1}\right]_j\right) \le 1.$$

Thus, plugging this bound into (30) entrywise,

$$\|\mathbf{A}((1+\delta)\circ w)\|_p - \|\mathbf{A}w\|_p \le \sum_{j\in[d]}\delta_j w_j\left[\mathbf{A}^\top v^{p-1}\right]_j(1+p\delta_j) \le \sum_{j\in[d]}\delta_j w_j = \|w_{t+1}\|_1 - \|w_t\|_1.$$

Rearranging yields the desired claim. □

Theorem 4. *Algorithm 2 runs in time $O(\mathrm{nnz}(\mathbf{A})\cdot\frac{p\log(nd/\epsilon)}{\epsilon})$. Further, its output solves Problem 2.*

Proof. The runtime follows from Line 7 (each iteration cost is dominated by multiplication through $\mathbf{A}$), so we prove correctness. Define potential Φ_t as in Lemma 3, and note that as $w_0 = \frac{\epsilon}{n^2d}\mathbf{1}$,

$$\Phi_0 \le \|\mathbf{A}w_0\|_p \le \frac{1}{n}\|\mathbf{1}\|_p \le 1.$$

The second inequality followed from our assumption on $\mathbf{A}$ entry sizes (Lemma 14). If Algorithm 2 breaks out of the while loop of Line 4, we have by Lemma 3 that for x returned on Line 11,

$$\|\mathbf{A}w_t\|_p - \|w_t\|_1 \le 1 \implies \|\mathbf{A}x\|_p \le \frac{1+\|w_t\|_1}{\|w_t\|_1} \le 1+\epsilon.$$

Thus, primal feasibility is always correct. We now prove correctness of dual feasibility. First, let $V_x(u) = \sum_{i\in[n]}u_i\log(\frac{u_i}{x_i})$ be the Kullback-Leibler divergence from x to u, for $x, u\in\Delta^d$. Define the normalized points $x_t = \frac{w_t}{\|w_t\|_1}$ in each iteration. Expanding definitions,

$$\begin{aligned}V_{x_{t+1}}(u) - V_{x_t}(u) &= \sum_{i\in[n]}u_i\log\frac{[x_t]_i}{[x_{t+1}]_i} \\ &= \sum_{i\in[n]}u_i\left(\log\left(\frac{\|w_{t+1}\|_1}{\|w_t\|_1}\right) + \log\left(\frac{1}{1+\eta[g_t]_i}\right)\right) \\ &\le -\eta(1-\eta)\langle g_t, u\rangle + \log\left(\frac{\|w_{t+1}\|_1}{\|w_t\|_1}\right).\end{aligned} \tag{31}$$

The only inequality used the bounds, for $g, \eta \in [0, 1]$,

$$\log\left(\frac{1}{1+\eta g}\right) \leq g \log\left(\frac{1}{1+\eta}\right) \leq -\eta(1-\eta)g.$$

Telescoping (31) over all T iterations, and using $V_{x_0}(u) \leq \log n$ for all $u \in \Delta^n$ since x_0 is uniform, we have that whenever Line 4 is not satisfied before the check on Line 7 (i.e. $t \geq T$),

$$\eta(1-\eta) \sum_{0 \leq t < T} \langle g_t, u \rangle \leq \log\left(\frac{\|w_T\|_1}{\|w_0\|_1}\right) + V_{x_0}(u) \leq \log\left(\frac{nd}{\epsilon^2}\right) + \log n \leq 2\log\left(\frac{nd}{\epsilon}\right). \quad (32)$$

The last inequality used $\|w_T\|_1 \leq \epsilon^{-1}$ by assumption, and $\|w_0\|_1 = \frac{\epsilon}{nd}$. Next, since each $g_t \geq \mathbf{1} - \mathbf{A}^\top (v_t)^{p-1}$ entrywise, defining $\bar{z} = \frac{z}{T}$,

$$\sum_{0 \leq t < T} \langle g_t, u \rangle \geq \sum_{0 \leq t < T} \left\langle \mathbf{1} - \mathbf{A}^\top (v_t)^{p-1}, u \right\rangle = T \left\langle \mathbf{1} - \mathbf{A}^\top \bar{z}, u \right\rangle, \text{ for all } u \in \Delta^n. \quad (33)$$

Combining (32) and (33), and rearranging, yields by definition of T,

$$\left\langle \mathbf{1} - \mathbf{A}^\top \bar{z}, u \right\rangle \leq \frac{2\log(\frac{nd}{\epsilon})}{T\eta(1-\eta)} \leq \frac{4p\log(\frac{nd}{\epsilon})}{T} \leq \epsilon \implies \mathbf{A}^\top \bar{z} \geq 1 - \epsilon \text{ entrywise.}$$

The last claim follows by setting u to be each coordinate-sparse simplex vector. Finally, since $\frac{\bar{z}}{\|\bar{z}\|_q} = \frac{z}{\|z\|_q}$, to show that y is a correct dual solution to Problem 2 it suffices to show $\|\bar{z}\|_q \leq 1$. This follows as $\bar{z}$ is an average of the $(v_t)^{p-1}$, convexity of ℓ_q norms, and that for all t,

$$\left\|(v_t)^{p-1}\right\|_q^q = \sum_{j \in [d]} v_t^p = \sum_{j \in [d]} \frac{[\mathbf{A}w_t]_j^p}{\|\mathbf{A}w_t\|_p^p} = 1.$$

□

C.2 Proofs from Section 4.3

Our analysis of Algorithm 3 will use the following helper fact.

Lemma 15 (Spectral bounds on $\{\mathbf{A}_i\}_{i \in [n]}$)**.** *Feasibility of Problem 3 is unaffected (up to constants in ϵ) by removing matrices $\mathbf{A}_i$ with an eigenvalue larger than $n\epsilon^{-1}$.*

Proof. The proof is identical to Lemma 14; we also require the additional fact that the Schatten norm $\|\cdot\|_p$ is monotone in the Loewner order, forcing the constraint $x_i \leq \frac{\epsilon(1+\epsilon)}{n}$. □

We remark that we can perform this preprocessing procedure via power iteration on each $\mathbf{A}_i$.

Lemma 4. *In all iterations t of Algorithm 3, defining $\Phi_t := \left\|\sum_{i \in [n]} [w_t]_i \mathbf{A}_i\right\|_p - \|w_t\|_1$, $\Phi_{t+1} \leq \Phi_t$.*

Proof. Drop t and define $\delta = \eta g$. For simplicity, define the matrices

$$\mathbf{M}_0 := \sum_{i \in [n]} w_i \mathbf{A}_i, \ \mathbf{M}_1 := \sum_{i \in [n]} \delta_i w_i \mathbf{A}_i, \ \mathbf{M}_2 := \sum_{i \in [n]} \delta_i^2 w_i \mathbf{A}_i.$$

We recall the Lieb-Thirring inequality $\text{Tr}((\mathbf{ABA})^p) \leq \text{Tr}(\mathbf{A}^{2p}\mathbf{B}^p)$. Applying this, we have

$$\|\mathbf{M}_0 + \mathbf{M}_1\|_p^p = \text{Tr}\left((\mathbf{M}_0 + \mathbf{M}_1)^p\right) \leq \text{Tr}\left(\mathbf{M}_0^p \left(\mathbf{I} + \mathbf{M}_0^{-\frac{1}{2}} \mathbf{M}_1 \mathbf{M}_0^{-\frac{1}{2}}\right)^p\right).$$

As $g \leq \mathbf{1}$, we have $\mathbf{M}_0^{-\frac{1}{2}} \mathbf{M}_1 \mathbf{M}_0^{-\frac{1}{2}} \preceq p^{-1}\mathbf{I}$. Applying the bounds $(\mathbf{I} + \mathbf{M})^p \preceq \exp(p\mathbf{M}) \preceq \mathbf{I} + p\mathbf{M} + p^2\mathbf{M}^2$ for $\mathbf{M} = \mathbf{M}_0^{-\frac{1}{2}} \mathbf{M}_1 \mathbf{M}_0^{-\frac{1}{2}}$, where we use that $\mathbf{I}$ commutes with all $\mathbf{M}$, it follows that

$$\|\mathbf{M}_0 + \mathbf{M}_1\|_p^p \leq \text{Tr}\left(\mathbf{M}_0^p + p\mathbf{M}_0^{p-1}\mathbf{M}_1 + p^2 \mathbf{M}_0^{p-1}\mathbf{M}_1\mathbf{M}_0^{-1}\mathbf{M}_1\right).$$

Definitions of $\mathbf{M}_0$, $\mathbf{M}_1$, $\mathbf{M}_2$, and preservation of positiveness under Schur complements imply

$$\begin{pmatrix} \mathbf{M}_0 & \mathbf{M}_1 \\ \mathbf{M}_1 & \mathbf{M}_2 \end{pmatrix} \succeq 0 \implies \mathbf{M}_2 - \mathbf{M}_1 \mathbf{M}_0^{-1} \mathbf{M}_1 \succeq 0.$$

Thus, $\mathbf{M}_1 \mathbf{M}_0^{-1} \mathbf{M}_1 \preceq \mathbf{M}_2$. Applying this and recalling $\mathbf{V} = \frac{\mathbf{M}_0}{\|\mathbf{M}_0\|_p}$,

$$\begin{aligned} \|\mathbf{M}_0 + \mathbf{M}_1\|_p^p &\leq \mathrm{Tr}\left(\mathbf{M}_0^p + p\mathbf{M}_0^{p-1}\mathbf{M}_1 + p^2\mathbf{M}_0^{p-1}\mathbf{M}_2\right) \\ &= \|\mathbf{M}_0\|_p^p \left(1 + p\left\langle \mathbf{V}^{p-1}, \frac{\mathbf{M}_1}{\|\mathbf{M}_0\|_p} + \frac{p\mathbf{M}_2}{\|\mathbf{M}_0\|_p} \right\rangle\right). \end{aligned}$$

By $(1+px)^{1/p} < 1 + x$, taking p^{th} roots we thus have

$$\|\mathbf{M}_0 + \mathbf{M}_1\|_p \leq \|\mathbf{M}_0\|_p + \left\langle \mathbf{V}^{p-1}, \mathbf{M}_1 + p\mathbf{M}_2 \right\rangle.$$

Finally, the conclusion follows as in Lemma 3; by linearity of trace and $g = p\delta$,

$$\left\langle \mathbf{V}^{p-1}, \mathbf{M}_1 + p\mathbf{M}_2 \right\rangle = \sum_{i\in[n]} \left\langle \mathbf{V}^{p-1}, \mathbf{A}_i \right\rangle \delta_i w_i (1 + p\delta_i) \leq \sum_{i\in[n]} \delta_i w_i.$$

Here, we used the inequality for all nonzero g_i,

$$\left\langle \mathbf{V}^{p-1}, \mathbf{A}_i \right\rangle (1 + p\delta_i) = \left\langle \mathbf{V}^{p-1}, \mathbf{A}_i \right\rangle \left(2 - \left\langle \mathbf{V}^{p-1}, \mathbf{A}_i \right\rangle\right) \leq 1.$$

□

Theorem 5. *Let p be odd. Algorithm 3 runs in $O(\frac{p\log(nd/\epsilon)}{\epsilon})$ iterations, and its output solves Problem 3. Each iteration is implementable in $O(\mathrm{nnz} \cdot \frac{p\log(nd/\epsilon)}{\epsilon^2})$, where* nnz *is the number of nonzero entries amongst all $\{\mathbf{A}_i\}_{i\in[n]}$, losing $O(\epsilon)$ in the quality of Problem 3 with probability* $1 - \mathrm{poly}((nd/\epsilon)^{-1})$.

Proof. The proof is analogous to that of Theorem 4; we sketch the main differences here. By applying Lemma 15 and monotonicity of Schatten norms in the Loewner order, we again have $\Phi_0 \leq 1$, implying correctness whenever the algorithm terminates on Line 4. Correctness of dual certification again follows from lack of termination and the choice of T, as well as setting u to indicate each coordinate. Finally, the returned matrix in Line 8 is correct by convexity of the Schatten-q norm, and the fact that all $\mathbf{V}_t^{p-1}$ have unit Schatten-q norm.

We now discuss issues regearding computing g_t in Line 5 of the algorithm, the bottleneck step; these techniques are standard in the approximate SDP literature, and we defer a more formal discussion to e.g. [JLL$^+$20]. First, note that each coordinate of g_t requires us to compute

$$\frac{1}{\left\|\sum_{i\in[n]}[w_t]_i\mathbf{A}_i\right\|_p^{p-1}} \cdot \left\langle \mathbf{A}_i, \left(\sum_{i\in[n]}[w_t]_i\mathbf{A}_i\right)^{p-1} \right\rangle. \tag{34}$$

We estimate the two quantities in the above expression each to $1+\epsilon$ multiplicative error with high probability. Union bounding over iterations, and modifying Lemma 4 to use the potential $\left\|\sum_{i\in[n]}[w_t]_i\mathbf{A}_i\right\|_p - (1 + O(\epsilon))\|w_t\|_1$, the analysis remains valid up to constants in ϵ with this multiplicative approximation quality. We now discuss our approximation strategies.

For shorthand, denote $\mathbf{M} = \sum_{i\in[n]}[w_t]_i\mathbf{A}_i$. To estimate the denominator of (34), it suffices to multiplicatively approximate $\|\mathbf{M}\|_p^p = \mathrm{Tr}[\mathbf{M}^p]$ within a $1+\epsilon$ factor, as raising to the $\frac{p-1}{p}$ power can only improve this. To do so, we use the well-known fact (e.g. [DG03]) that letting $\mathbf{Q}$ be a $k \times d$ matrix with independent entries $\sim \mathcal{N}(0, \frac{1}{k})$, for $k = O(\frac{\log(\frac{nd}{\epsilon})}{\epsilon^2})$, with probability $1 - \mathrm{poly}((\frac{nd}{\epsilon})^{-1})$,

$$\mathrm{Tr}[\mathbf{M}^p] \approx \sum_{\ell\in[k]} \mathbf{Q}_{\ell:}^\top \mathbf{M}^p \mathbf{Q}_{\ell:}$$

to a $1+\epsilon$ factor. To read this from the standard Johnson-Lindestrauss guarantee, it suffices to factorize $\mathbf{M}^p$ and use that each row of the square root's ℓ_2 norm is preserved with high probability under multiplication by $\mathbf{Q}$, and then apply the cyclic definition of trace. Similarly, for each $i \in [n]$, we can approximate the numerators via

$$\operatorname{Tr}\left(\mathbf{Q}\mathbf{M}^{\frac{p-1}{2}}\mathbf{A}_i\mathbf{M}^{\frac{p-1}{2}}\mathbf{Q}^\top\right).$$

We can simultaneously compute all such quantities by first applying $O(p)$ matrix-vector multiplications through $\mathbf{M}$ to each row of $\mathbf{Q}$, and then computing all quadratic forms. In total, the computational cost per iteration of all approximations is $O(\text{nnz} \cdot \frac{p\log(\frac{nd}{\epsilon})}{\epsilon^2})$ as desired. □

C.3 Proof of Proposition 2

In this section, following our prior developments, we prove the following claim.

Proposition 2. *Following Theorem 5's notation, let p be odd, $\{\mathbf{A}_i\}_{i\in[n]} \in \mathbb{S}^d_{\geq 0}$, $0 < \epsilon = O(\alpha)$, and*

$$\min_{\substack{x\in\Delta^n \\ \|x\|_\infty \leq \frac{1+\alpha}{n}}} \|\mathcal{A}(x)\|_p = \text{OPT}. \tag{4}$$

for $\mathcal{A}(x) := \sum_{i\in[n]} x_i\mathbf{A}_i$. Given estimate of OPT *exponentially bounded in $\frac{nd}{\epsilon}$, there is a procedure calling Algorithm 7 $O(\log\frac{nd}{\epsilon})$ times giving $x \in \Delta^n$ with $\|x\|_\infty \leq \frac{(1+\alpha)(1+\epsilon)}{n}$, $\|\mathcal{A}(x)\|_p \leq (1+\epsilon)$*OPT. *Algorithm 7 runs in $O(\frac{\log(nd/\epsilon)\log n}{\epsilon^2})$ iterations, each requiring time $O(\text{nnz}\cdot\frac{p\log(nd/\epsilon)}{\epsilon^2})$.*

C.3.1 Reduction to a decision problem

Given access to an oracle for the following approximate decision problem, we can implement an efficient binary search for estimating OPT. Specifically, letting the range of OPT be $(\mu_{\text{lower}}, \mu_{\text{upper}})$, we can subdivide the range into $O(\frac{1}{\epsilon}\log\frac{\mu_{\text{upper}}}{\mu_{\text{lower}}})$ multiplicative intervals of range $1+\epsilon$, and then compute a binary search using our decision oracle. This incurs a multiplicative $\log(\frac{nd}{\epsilon})$ overhead in the setting of Proposition 2 (see Appendix A, [JLL$^+$20], for a more formal treatment).

Problem 4. *Given $\{\mathbf{A}_i\}_{i\in[n]} \in \mathbb{S}^d_{\geq 0}$, either find primal solution $x \in \Delta^n$ with $\|\mathcal{A}(x)\|_p \leq 1+\epsilon$, $\|x\|_\infty \leq \frac{(1+\epsilon)(1+\alpha)}{n}$, or conclude no $x \in \Delta^n$ satisfies $\|\mathcal{A}(x)\|_p \leq 1-\epsilon$, $\|x\|_\infty \leq \frac{(1-\epsilon)(1+\alpha)}{n}$.*

The hard constraint $\|x\|_\infty \leq \frac{1+\alpha}{n}$ in the definition (4) can be adjusted by constant factors to admit the ℓ_∞ bound in Problem 4, since we assumed $\epsilon = O(\alpha)$ is sufficiently small.

C.3.2 Preliminaries

We use the shorthand $\mathbf{S} := \frac{n}{1+\alpha}\mathbf{I}$, and $p' := \frac{\log n}{\epsilon}$, so $\ell_{p'}$ and ℓ_∞ are interchangeable up to $1+O(\epsilon)$ factors. In other words, Problem 4 asks to certify whether there exists $x \in \Delta^n$ with

$$\max\left(\|\mathcal{A}(x)\|_p,\ \|\mathbf{S}x\|_{p'}\right) \leq 1, \tag{35}$$

up to multiplicative $1+\epsilon$ tolerance on either side. Consider the potential function

$$\Phi(w) := \log\left(\exp\left(\|\mathcal{A}(w)\|_p\right) + \exp\left(\|\mathbf{S}w\|_{p'}\right)\right) - \|w\|_1. \tag{36}$$

It is clear that the first term of $\Phi(w)$ approximates the left hand side of (35) up to a $\log 2$ additive factor, so if any of $\|\mathcal{A}(w)\|_p$, $\|\mathcal{A}(w)\|_{p'}$, or $\|w\|_1$ reaches the scale $3\epsilon^{-1}$ and $\Phi(w)$ is bounded by 1, we can safely terminate. and conclude primal feasibility for Problem 4. Next, we compute

$$\begin{aligned}\nabla_i\Phi(w) = 1 - \frac{\exp\left(\|\mathcal{A}(w)\|_p\right)\langle\mathbf{A}_i, \mathbf{Y}(w)\rangle + \exp\left(\|\mathbf{S}w\|_{p'}\right)[\mathbf{S}z(w)]_i}{\exp\left(\|\mathcal{A}(w)\|_p\right) + \exp\left(\|\mathbf{S}w\|_{p'}\right)} \text{ for all } i \in [n],\\ \text{where } \mathbf{Y}(w) := \left(\frac{\mathcal{A}(w)}{\|\mathcal{A}(w)\|_p}\right)^{p-1},\ z(w) := \left(\frac{\mathbf{S}w}{\|\mathbf{S}w\|_{p'}}\right)^{p'-1}\end{aligned} \tag{37}$$

The following helper lemma will be useful in concluding dual infeasibility of Problem 4.

Lemma 16. *In the setting of Problem 4, suppose there exists $x^* \in \Delta^n$ with*

$$\|\mathcal{A}(x^*)\|_p \le 1-\epsilon,\ \|\mathbf{S}x^*\|_{p'} \le 1-\epsilon.$$

Then, for any w,

$$\langle \nabla\Phi(w), x^*\rangle \ge \epsilon.$$

Proof. From the definitions in (37), it is clear that $\|\mathbf{Y}(w)\|_q = \|z(w)\|_{q'} = 1$, where q, q' are the dual norms of p, p' respectively. Moreover, by the definition of x^*, we have for all $\|\mathbf{Y}\|_q = \|z\|_{q'} = 1$,

$$\langle \mathbf{Y}, \mathcal{A}(x)\rangle \le 1-\epsilon,\ \langle z, \mathbf{S}x\rangle \le 1-\epsilon.$$

This follows from the dual definition of the ℓ_p norm (see Fact 2). Now, note that for some nonnegative $\alpha(w), \beta(w)$ summing to 1, using the above claim and (37),

$$\langle \nabla\Phi(w), x^*\rangle = 1 - (\alpha(w)\langle \mathbf{Y}(w), \mathcal{A}(x^*)\rangle + \beta(w)\langle z(w), \mathbf{S}x^*\rangle) \ge \epsilon,$$

as desired (here, we used positivity of all relevant quantities). □

C.3.3 Potential monotonicity

We prove a monotonicity property regarding the potential Φ in (36).

Lemma 17. *Let $w \in \mathbb{R}^n_{\ge 0}$ satisfy $\|\mathcal{A}(w)\|_p \le 3\epsilon^{-1}$, $\|\mathbf{S}w\|_{p'} \le 3\epsilon^{-1}$, let $g = \max(0, \nabla\Phi(w))$ entrywise, and let $w' = (1+\eta g)\circ w$, where $\eta = (4p')^{-1}$. Then, $\Phi(w') \le \Phi(w)$.*

Proof. Denote for simplicity the threshold $K = 3\epsilon^{-1}$ and the step vector $\delta = \eta g$. First, by prior calculations in Lemma 3 and Lemma 4, it follows that

$$\|\mathcal{A}(w')\|_p \le \|\mathcal{A}(w)\|_p + \Delta_{\mathcal{A}},\ \|\mathbf{S}w'\|_{p'} \le \|\mathbf{S}w\|_{p'} + \Delta_{\mathbf{S}},$$
$$\text{where } \Delta_{\mathcal{A}} := \sum_{i\in[n]} \langle \mathbf{A}_i, \mathbf{Y}(w)\rangle\, \delta_i w_i(1+p\delta_i),\ \Delta_{\mathbf{S}} := \sum_{i\in[n]} [\mathbf{S}z(w)]_i \delta_i w_i(1+p'\delta_i).$$

Next, note that by $\delta \le \eta$ entrywise and lack of termination (i.e. the threshold K),

$$\Delta_{\mathcal{A}} \le (1+p\eta)\eta\langle \mathbf{Y}(w), \mathcal{A}(w)\rangle \le 2\eta\|\mathcal{A}(w)\|_p \le 1.$$

Therefore, by $\exp(x) \le 1 + x + x^2$ for $x \le 1$,

$$\exp\left(\|\mathcal{A}(w')\|_p\right) \le \exp\left(\|\mathcal{A}(w)\|_p\right)\left(1 + \Delta_{\mathcal{A}} + \Delta_{\mathcal{A}}^2\right). \tag{38}$$

Moreover, by applying Cauchy-Schwarz and the threshold $\|\mathcal{A}(w)\|_p \le K$ once more,

$$\begin{aligned}\Delta_{\mathcal{A}}^2 &\le (1+p\eta)^2\left(\sum_{i\in[n]} \langle \mathbf{A}_i, \mathbf{Y}(w)\rangle\, \delta_i w_i\right)^2 \\ &\le 2\left(\sum_{i\in[n]} \langle \mathbf{A}_i, \mathbf{Y}(w)\rangle\, \delta_i^2 w_i\right)\langle \mathbf{Y}(w), \mathcal{A}(w)\rangle \le 2K\left(\sum_{i\in[n]} \langle \mathbf{A}_i, \mathbf{Y}(w)\rangle\, \delta_i^2 w_i\right).\end{aligned} \tag{39}$$

Combining (38) and (39) (and applying similar reasoning to the term $\Delta_{\mathbf{S}}$), we conclude

$$\exp\left(\|\mathcal{A}(w')\|_p\right) \le \exp\left(\|\mathcal{A}(w)\|_p\right)\left(1 + \sum_{i\in[n]} \langle \mathbf{A}_i, \mathbf{Y}(w)\rangle\, \delta_i w_i(1+(p+2K)\delta_i)\right),$$
$$\exp\left(\|\mathbf{S}w'\|_{p'}\right) \le \exp\left(\|\mathbf{S}w\|_{p'}\right)\left(1 + \sum_{i\in[n]} [\mathbf{S}z(w)]_i \delta_i w_i(1+(p'+2K)\delta_i)\right).$$

Recall the inequality $\log(1+x) \le x$ for nonnegative x. Expanding the definition of Φ and $\nabla\Phi$ (cf. (36)), and plugging in the above bounds, we conclude that

$$\begin{aligned}\Phi(w') - \Phi(w) &= \log\left(\frac{\exp\left(\|\mathcal{A}(w')\|_p\right) + \exp\left(\|\mathbf{S}w'\|_{p'}\right)}{\exp\left(\|\mathcal{A}(w)\|_p\right) + \exp\left(\|\mathbf{S}w\|_{p'}\right)}\right) - \langle \delta, w\rangle \\ &\le \sum_{i\in[n]} (1 - \nabla_i\Phi(w))\delta_i w_i(1 + (p'+2K)\delta_i) - \langle \delta, w\rangle \\ &= \sum_{i\in[n]} \left((1-\nabla_i\Phi(w))(1+(p'+2K)\delta_i) - 1\right)\delta_i w_i.\end{aligned}$$

As before, we show that this sum is entrywise nonpositive. For any $i \in [n]$ with $\delta_i \neq 0$, we have

$$\begin{aligned}(1-\nabla_i\Phi(w))(1+(p'+2K)\delta_i) - 1 &= (1-\nabla_i\Phi(w))(1+(p'+2K)\eta\nabla_i\Phi(w)) - 1 \\ &\le (1-\nabla_i\Phi(w))(1+\nabla_i\Phi(w)) - 1 \le 0,\end{aligned}$$

as desired, where we used that $\eta^{-1} \ge p' + 2K$. This yields the conclusion $\Phi(w') \le \Phi(w)$. □

C.3.4 Algorithm and analysis

Finally, we state Algorithm 7 and prove Proposition 2.

Algorithm 7 BoxedSchattenPacking$(\{\mathbf{A}_i\}_{i\in[n]}, \epsilon, p, \alpha)$

1: **Input:** $\{\mathbf{A}_i\}_{i\in[n]} \in \mathbb{S}^d_{\succeq 0}, \epsilon \in [0, \frac{1}{2}], p \ge 2, \alpha \in [0, n-1]$
2: $p' \leftarrow \frac{\log n}{\epsilon}$, $\mathbf{S} \leftarrow \frac{n}{1+\alpha}\mathbf{I}$
3: $\eta \leftarrow (4p')^{-1}$, $K \leftarrow 3\epsilon^{-1}$, $T \leftarrow \frac{6\log(\frac{nd}{\epsilon})}{\eta\epsilon}$
4: $[w_0]_i \leftarrow \frac{\epsilon}{n^2 d}$ for all $i \in [n]$, $t \leftarrow 0$
5: **while** $\|\mathcal{A}(w_t)\|_p, \|\mathbf{S}w_t\|_{p'}, \|w_t\|_1 \le K$ **do**
6: $\quad g_t \leftarrow \max(0, \nabla\Phi(w_t))$ entrywise, where we use the definition (36)
7: $\quad w_{t+1} \leftarrow w_t \circ (1+\eta g_t)$, $t \leftarrow t+1$
8: $\quad$ **if** $t \ge T$ **then**
9: $\quad\quad$ **return** Infeasible
10: $\quad$ **end if**
11: **end while**
12: **return** $x = \frac{w_t}{\|w_t\|_1}$

Proof of Proposition 2. Correctness of the reduction to deciding Problem 4 follows from the discussion in Section C.3.1. Moreover, by the given Algorithm 7, it is clear (following e.g. the preprocessing of Lemma 15) that $\Phi(w_t) \le 1$ throughout the algorithm, so whenever the algorithm terminates we have primal feasibility. It suffices to prove that whenever the problem admits x^* with

$$\|\mathcal{A}(x^*)\|_p \le 1-\epsilon, \ \|\mathbf{S}x^*\|_{p'} \le 1-\epsilon,$$

then the algorithm terminates on Line 5 in T iterations. Analogously to Theorem 4, we have

$$\eta(1-\eta)\sum_{0\le t<T}\langle g_t, x^*\rangle \le \log n - \log\|w_0\|_1 + \log\|w_T\|_1 \le 2\log\left(\frac{nd}{\epsilon}\right) + \log\|w_T\|_1.$$

Next, since g_t is an upwards truncation of $\nabla\Phi(w_t)$, applying Lemma 16 implies that

$$\|w_T\|_1 \ge \exp\left(\frac{\eta\epsilon T}{2} - 2\log\left(\frac{nd}{\epsilon}\right)\right).$$

The conclusion follows by the definition of T, as desired. Finally, the iteration complexity follows analogously to the discussion in Theorem 5's proof, where the only expensive cost is estimating coordinates of the $\mathcal{A}$ component of $\nabla\Phi(w_t)$ every iteration. □

Finally, we remark that by opening up the dual certificates $\mathbf{Y}(w), \mathbf{Z}(w)$ of our mirror descent analysis, we can in fact implement a stronger version of the decision Problem 4 which returns a feasible dual certificate whenever the primal problem is infeasible. We omit this extension for brevity, as it is unnecessary for our applications, but it is analogous to the analysis of Theorem 5.

D Deferred proofs from Section 5

D.1 Proof of Proposition 3

Proposition 3. *There is an algorithm* Power *(Algorithm 1, [MM15]), parameterized by $t \in [d]$, tolerance $\tilde{\epsilon} > 0$, $p \geq 1$, and $\mathbf{A} \in \mathbb{S}^d_{\geq 0}$, which outputs orthonormal $\{z_j\}_{j\in[t]}$ with the guarantee*

$$\left.\begin{aligned} \left|z_j^\top \mathbf{A}^p z_j - \lambda_j^p(\mathbf{A})\right| &\leq \tilde{\epsilon}\lambda_j^p(\mathbf{A}) \\ \left|z_j^\top \mathbf{A}^{p-1} z_j - \lambda_j^{p-1}(\mathbf{A})\right| &\leq \tilde{\epsilon}\lambda_j^{p-1}(\mathbf{A}) \end{aligned}\right\} \text{ for all } j \in [t]. \tag{6}$$

Here, $\lambda_j(\mathbf{A})$ is the j^{th} largest eigenvalue of $\mathbf{A}$. The total time required by the method is $O(\mathrm{nnz}(\mathbf{A})\frac{tp\log d}{\varepsilon})$.

Proof. We claim that Algorithm 1 in [MM15] applied to the matrix $\mathbf{A}^p$ with a careful choice of exponent q in their Algorithm 1 yields this guarantee. Specifically, we choose q_1, q_2, both of which satisfy the criteria in their main theorem, such that the iterates produced by simultaneous power iteration $\mathbf{M}^p$ with exponent q_1 and $\mathbf{M}^{p-1}$ with exponent q_2 are identical; it suffices to choose q a multiple of $p(p-1)$. Thus, we can also apply their guarantees to $\mathbf{A}^{p-1}$ and apply a union bound. Notice that their Algorithm 1 also contains some postprocessing to ensure that they obtain singular values in the right space, which is unnecessary for us, as our matrices are Hermitian. □

D.2 Proof of Lemma 5

Lemma 5. *Let $n = \Omega\left(\frac{d+\log\delta^{-1}}{(\epsilon\log\epsilon^{-1})^2}\right)$. With probability $1-\frac{\delta}{2}$, the uniform distribution over G attains value $(1+\frac{\tilde{\epsilon}}{2})\|\mathbf{\Sigma}\|_p$ for objective* (5)*, where $\tilde{\epsilon} = C'\epsilon\log\epsilon^{-1}$ for a universal constant $C' > 0$.*

Proof. Lemma 8 implies that letting w^* be the uniform distribution over the uncorrupted samples amongst $X_1, \ldots, X_n$, we have with probability at least $1-\frac{\delta}{2}$, and denoting $\tilde{\epsilon} := 2C_3 \cdot \epsilon\log\epsilon^{-1}$,

$$\left\|\sum_{i\in[n]} w_i^* X_i X_i^\top\right\|_p \leq \left(1+\frac{\tilde{\epsilon}}{2}\right)\|\mathbf{\Sigma}\|_p.$$

Therefore, the mixed ℓ_∞-ℓ_p packing semidefinite program

$$\exists w^* \in \Delta^n \text{ with } \|w^*\|_\infty \leq \frac{1}{(1-\epsilon)n},\ \left\|\sum_{i\in[n]} w_i^* X_i X_i^\top\right\|_p \leq \left(1+\frac{\tilde{\epsilon}}{2}\right)\|\mathbf{\Sigma}\|_p$$

is feasible. This completes the proof. □

D.3 Proof of Lemma 6

Lemma 6. *Let $n = \Omega\left(\frac{d+\log\delta^{-1}}{(\epsilon\log\epsilon^{-1})^2}\right)$. With probability at least $1-\frac{\delta}{2}$, $(1+\tilde{\epsilon})\mathbf{\Sigma} \succeq \mathbf{M}_G \succeq (1-\tilde{\epsilon})\mathbf{\Sigma}$.*

Proof. We follow the notation of (7). First, by the guarantees of Corollary 1,

$$w_G = 1 - w_B \geq 1 - \frac{\epsilon n}{(1-2\epsilon)n} = \frac{1-3\epsilon}{1-2\epsilon} \geq 1-2\epsilon.$$

Therefore, again applying Corollary 1, for all $i \in G$,

$$\frac{w_i}{w_G} \leq \frac{1}{(1-2\epsilon)n}\cdot\frac{1-2\epsilon}{1-3\epsilon} = \frac{1}{(1-3\epsilon)n}.$$

We conclude that the set of weights $\{\frac{w_i}{w_G}\}_{i\in G}$ belong to $\mathfrak{S}_{3\epsilon}^{(1-\epsilon)n}$. By applying Corollary 3 to these weights and adjusting the definition of C_3 by a constant, we conclude with probability at least $1-\frac{\delta}{2}$

$$\left(1+C_3\cdot\epsilon\log\epsilon^{-1}\right)\mathbf{\Sigma} \succeq \sum_{i\in G}\frac{w_i}{w_G}X_iX_i^\top \succeq \left(1-C_3\cdot\epsilon\log\epsilon^{-1}\right)\mathbf{\Sigma}.$$

The conclusion follows by multiplying through by w_G, and using the definition $\tilde{\epsilon} = 2C_3\cdot\epsilon\log\epsilon^{-1}$. □

D.4 Proof of Proposition 4

Proposition 4. *Let $\mathbf{M} = \mathbf{M}_G + \mathbf{M}_B$ be so that $\|\mathbf{M}\|_p \le (1+\tilde{\epsilon})\|\mathbf{\Sigma}\|_p$, $\mathbf{M}_G \succeq 0$ and $\mathbf{M}_B \succeq 0$, and so that $(1+\tilde{\epsilon})\mathbf{\Sigma} \succeq \mathbf{M}_G \succeq (1-\tilde{\epsilon})\mathbf{\Sigma}$. Following notation of Algorithm 4, let*

$$\mathbf{M} = \sum_{j\in[d]} \lambda_j v_j v_j^\top, \ \mathbf{\Sigma} = \sum_{j\in[d]} \sigma_j u_j u_j^\top \tag{8}$$

be sorted eigendecompositions of $\mathbf{M}$ and $\mathbf{\Sigma}$, so $\lambda_1 \ge \ldots \ge \lambda_d$, and $\sigma_1 \ge \ldots \ge \sigma_d$. Let γ be as in Theorem 2, and assume $\sigma_{t+1} < (1-\gamma)\sigma_1$. Then,

$$\max_{j\in[t]} v_j^\top \mathbf{\Sigma} v_j \ge (1-\gamma)\|\mathbf{\Sigma}\|_\infty.$$

Proof. For concreteness, we will define the parameters

$$p = \frac{2}{7}\sqrt{\frac{\log(3d)}{\tilde{\epsilon}}}, \ \gamma = 14\sqrt{\tilde{\epsilon}\log(3d)} = 49p\tilde{\epsilon}.$$

For these choices of p, γ, we will use the following (loose) approximations for sufficiently small $\tilde{\epsilon}$:

$$\left(1-\frac{\gamma}{4}\right)^p = \left(1-\frac{\gamma}{4}\right)^{\frac{4}{\gamma}\log(3d)} \le \frac{1}{3d}, \ (1+\tilde{\epsilon})^p - (1-\tilde{\epsilon})^p \le \exp(p\tilde{\epsilon}) - (1-p\tilde{\epsilon}) \le 3p\tilde{\epsilon}. \tag{40}$$

Suppose for contradiction that all $v_j^\top \mathbf{\Sigma} v_j < (1-\gamma)\sigma_1$ for $j \in [t]$. By applying the guarantee of Corollary 1 and Fact 2, it follows that

$$\left\langle \mathbf{M}, \mathbf{M}^{p-1}\right\rangle = \|\mathbf{M}\|_p^p \le (1+\tilde{\epsilon})^p \|\mathbf{\Sigma}\|_p^p. \tag{41}$$

Let $s \in [d]$ be the largest index such that $\sigma_s > \left(1-\frac{\gamma}{4}\right)\sigma_1$, and note that $s \le t$. We define

$$\mathbf{N} := \sum_{j\in[s]} \lambda_j^{p-1} v_j v_j^\top \preceq \mathbf{M}^{p-1}.$$

That is, $\mathbf{N}$ is the restriction of $\mathbf{M}^{p-1}$ to its top s eigendirections. Then,

$$\begin{aligned}\left\langle \mathbf{M}, \mathbf{M}^{p-1}\right\rangle &= \left\langle \mathbf{M}_B, \mathbf{M}^{p-1}\right\rangle + \left\langle \mathbf{M}_G, \mathbf{M}^{p-1}\right\rangle \\ &\ge \left\langle \mathbf{M}_B, \mathbf{M}^{p-1}\right\rangle + \left\langle (1-\tilde{\epsilon})\mathbf{\Sigma}, \mathbf{M}^{p-1}\right\rangle \ge \langle \mathbf{M}_B, \mathbf{N}\rangle + (1-\tilde{\epsilon})^p \|\mathbf{\Sigma}\|_p^p.\end{aligned} \tag{42}$$

In the second line, we used Lemma 6 twice, as well as the trace inequality Lemma 18 with $\mathbf{A} = \mathbf{M}$ and $\mathbf{B} = (1-\tilde{\epsilon})\mathbf{\Sigma}$. Combining (41) with (42), and expanding the definition of $\mathbf{M}_B$, yields

$$\begin{aligned}((1+\tilde{\epsilon})^p - (1-\tilde{\epsilon})^p)\|\mathbf{\Sigma}\|_p^p \ge \langle \mathbf{M}_B, \mathbf{N}\rangle &= \left\langle \mathbf{M}_B, \sum_{j\in[s]} \lambda_j^{p-1} v_j v_j^\top \right\rangle \\ &= \left\langle \mathbf{M}, \sum_{j\in[s]} \lambda_j^{p-1} v_j v_j^\top \right\rangle - \left\langle \mathbf{M}_G, \sum_{j\in[s]} \lambda_j^{p-1} v_j v_j^\top \right\rangle \\ &\ge \left\langle \mathbf{M}, \sum_{j\in[s]} \lambda_j^{p-1} v_j v_j^\top \right\rangle - (1+\tilde{\epsilon}) \left\langle \mathbf{\Sigma}, \sum_{j\in[s]} \lambda_j^{p-1} v_j v_j^\top \right\rangle \\ &= \sum_{j\in[s]} \left(\lambda_j^p - (1+\tilde{\epsilon})\lambda_j^{p-1} v_j^\top \mathbf{\Sigma} v_j\right) \ge \sum_{j\in[s]} \left(\lambda_j^p - \lambda_j^{p-1}(1+\tilde{\epsilon})(1-\gamma)\sigma_1\right).\end{aligned} \tag{43}$$

The third line followed from from the spectral bound $\mathbf{M}_G \preceq (1+\tilde{\epsilon})\mathbf{\Sigma}$ of Lemma 6, and the fourth followed from the fact that $\{\lambda_j\}_{j\in[d]}$, $\{v_j\}_{j\in[d]}$ eigendecompose $\mathbf{M}$, as well as the assumption $v_j^\top \mathbf{\Sigma} v_j \le (1-\gamma)\sigma_1$ for all $j \in [t]$. Letting $S := \sum_{j\in[s]} \sigma_j^p$, and using both approximations in (40),

$$\|\mathbf{\Sigma}\|_p^p \le \sum_{j\in[s]} \sigma_j^p + \left(1-\frac{\gamma}{4}\right)^p (d-s)\sigma_1^p \le \frac{4}{3}S \implies ((1+\tilde{\epsilon})^p - (1-\tilde{\epsilon})^p)\|\mathbf{\Sigma}\|_p^p \le 4p\tilde{\epsilon}S. \tag{44}$$

Next, we bound the last term of (43). By using $(1+\tilde{\epsilon})(1-\gamma) \le 1 - \frac{\gamma}{2}$,

$$\begin{aligned}\sum_{j\in[s]} \left(\lambda_j^p - \lambda_j^{p-1}(1+\tilde{\epsilon})(1-\gamma)\sigma_1\right) &\ge \sum_{j\in[s]} \lambda_j^{p-1}\left(\lambda_j - \left(1-\frac{\gamma}{2}\right)\sigma_1\right) \\ &\ge \frac{\gamma}{6}\sum_{j\in[s]} \lambda_j^{p-1}\sigma_1 \ge \frac{\gamma}{6}(1-\tilde{\epsilon})^{p-1}\sum_{j\in[s]}\sigma_j^p \ge \frac{\gamma}{12}S.\end{aligned} \tag{45}$$

The second line used $\lambda_j - (1-\frac{\gamma}{2})\sigma_1 \ge (1-\tilde{\epsilon})\sigma_j - (1-\frac{\gamma}{2})\sigma_1 \ge \frac{\gamma}{6}\sigma_1$ by definition of s, Lemma 19 (twice), and $(1-\tilde{\epsilon})^{p-1} \ge \frac{1}{2}$. Combining (45) and (44) and plugging into (43),

$$4p\tilde{\epsilon}S \ge \frac{\gamma}{12}S \implies 48p\tilde{\epsilon} \ge \gamma.$$

By the choice of γ and p (i.e. $\gamma = 49p\tilde{\epsilon}$), we attain a contradiction. □

In the proof of Proposition 4, we used the following facts.

Lemma 18. *Let $\mathbf{A} \succeq \mathbf{B} \succeq 0$ be symmetric matrices and p a positive integer. Then we have*

$$\mathrm{Tr}\left(\mathbf{A}^{p-1}\mathbf{B}\right) \ge \mathrm{Tr}\left(\mathbf{B}^p\right).$$

Proof. For any $1 \le k \le p-1$,

$$\mathrm{Tr}\left(\mathbf{A}^k\mathbf{B}^{p-k}\right) \ge \mathrm{Tr}\left(\mathbf{A}^{k-1}\mathbf{B}^{\frac{p-k}{2}}\mathbf{A}\mathbf{B}^{\frac{p-k}{2}}\right) \ge \mathrm{Tr}\left(\mathbf{A}^{k-1}\mathbf{B}^{\frac{p-k}{2}}\mathbf{B}\mathbf{B}^{\frac{p-k}{2}}\right) = \mathrm{Tr}\left(\mathbf{A}^{k-1}\mathbf{B}^{p-k+1}\right).$$

The first step used the Extended Lieb-Thirring trace inequality $\mathrm{Tr}(\mathbf{M}\mathbf{N}^2) \ge \mathrm{Tr}(\mathbf{M}^\alpha\mathbf{N}\mathbf{M}^{1-\alpha}\mathbf{N})$ for $\alpha \in [0,1]$, $\mathbf{M}, \mathbf{N} \in \mathbb{S}^d_{\succeq 0}$ (see e.g. Lemma 2.1, [ALO16]), and the second $\mathbf{A} \succeq \mathbf{B}$. Finally, induction on k yields the claim. □

Lemma 19. *For all $j \in [d]$, $\lambda_j \ge (1-\tilde{\epsilon})\sigma_j$.*

Proof. By the Courant-Fischer minimax characterization of eigenvalues,

$$\lambda_j \ge \min_{k\in[j]} u_k^\top \mathbf{M} u_k.$$

However, we also have $\mathbf{M} \succeq \mathbf{M}_G \succeq (1-\tilde{\epsilon})\mathbf{\Sigma}$ (Lemma 6), yielding the conclusion. □

D.5 Proof of Theorem 2

The guarantees of Proposition 4 were geared towards exact eigenvectors of the matrix $\mathbf{M}$. We now modify the analysis to tolerate inexactness in the eigenvector computation, in line with the processing of Line 5 of our Algorithm 4. This yields our final claim in Theorem 2.

Corollary 5. *In the setting of Proposition 4, and letting $\{z_j\}_{j\in[t]}$ satisfy* (6)*, set for all $j \in [t]$*

$$y_j := \frac{\mathbf{M}^{\frac{p-1}{2}} z_j}{\left\|\mathbf{M}^{\frac{p-1}{2}} z_j\right\|_2}.$$

Then with probability at least $1-\delta$,

$$\max_{j\in[t]} y_j^\top \mathbf{\Sigma} y_j \ge (1-\gamma)\left\|\mathbf{\Sigma}\right\|_\infty.$$

Proof. Assume all y_j have $y_j^\top \mathbf{\Sigma} y_j \le (1-\gamma)\sigma_1$ for contradiction. We outline modifications to the proof of Proposition 4. Specifically, we redefine the matrix $\mathbf{N}$ by

$$\mathbf{N} := \mathbf{M}^{\frac{p-1}{2}}\left(\sum_{j\in[s]} z_j z_j^\top\right)\mathbf{M}^{\frac{p-1}{2}}.$$

Because $\sum_{j\in[s]} z_j z_j^\top$ is a projection matrix, it is clear $\mathbf{N} \preceq \mathbf{M}^{p-1}$. Therefore, by combining the derivations (41) and (42), it remains true that

$$\left((1+\tilde{\epsilon})^p - (1-\tilde{\epsilon})^p\right) \|\mathbf{\Sigma}\|_p^p \geq \langle \mathbf{M}_B, \mathbf{N}\rangle = \langle \mathbf{M}, \mathbf{N}\rangle - \langle \mathbf{M}_G, \mathbf{N}\rangle .$$

We now bound these two terms in an analogous way from Proposition 4, with negligible loss; combining these bounds will again yield a contradiction. First, we have the lower bound

$$\left\langle \mathbf{M}, \sum_{j\in[s]} \mathbf{M}^{\frac{p-1}{2}} z_j z_j^\top \mathbf{M}^{\frac{p-1}{2}} \right\rangle = \sum_{j\in[s]} z_j^\top \mathbf{M}^p z_j \geq (1-\tilde{\epsilon}) \sum_{j\in[s]} \lambda_j^p.$$

Here, the last inequality applied the assumption (6) with respect to $\mathbf{M}^p$. Next, we upper bound

$$\begin{aligned}
\left\langle \mathbf{M}_G, \sum_{j\in[s]} \mathbf{M}^{\frac{p-1}{2}} z_j z_j^\top \mathbf{M}^{\frac{p-1}{2}} \right\rangle &\leq (1+\tilde{\epsilon}) \left\langle \mathbf{\Sigma}, \sum_{j\in[s]} \mathbf{M}^{\frac{p-1}{2}} z_j z_j^\top \mathbf{M}^{\frac{p-1}{2}} \right\rangle \\
&= (1+\tilde{\epsilon}) \sum_{j\in[s]} \left\| \mathbf{M}^{\frac{p-1}{2}} z_j \right\|_2^2 y_j^\top \mathbf{\Sigma} y_j \\
&\leq (1+\tilde{\epsilon})(1-\gamma)\sigma_1 \sum_{j\in[s]} z_j^\top \mathbf{M}^{p-1} z_j \\
&\leq (1-\gamma)(1+\tilde{\epsilon})^2 \sigma_1 \sum_{j\in[s]} \lambda_j^{p-1},
\end{aligned}$$

The first line used $\mathbf{M}_G \preceq (1+\tilde{\epsilon})\mathbf{\Sigma}$, the second used the definition of y_j, the third used our assumption $y_j^\top \mathbf{\Sigma} y_j \leq (1-\gamma)\sigma_1$, and the last used (6) with respect to $\mathbf{M}^{p-1}$. Finally, the remaining derivation (45) is tolerant to additional factors of $1+\tilde{\epsilon}$, yielding the same conclusion up to constants. □

Finally, we prove Theorem 2 by combining the tools developed thus far.

Proof of Theorem 2. Correctness of the algorithm is immediate from Corollary 5 and the guarantees of 1DRobustVariance. Concretely, Corollary 5 guarantees that one of the vectors we produce will be a $(1-\gamma)$-approximate top eigenvector (say some index $j \in [t]$), and 1DRobustVariance will only lose a negligible fraction $O(\epsilon \log \epsilon^{-1})$ of this quality (see Lemma 1); the best returned eigenvector as measured by 1DRobustVariance can only improve the guarantee. Finally, the failure probability follows by combining the guarantees of Lemmas 1, 5, and 6.

We now discuss runtime. The complexity of lines 2, 4, and 5, as guaranteed by Proposition 2, Proposition 3, and Lemma 1 are respectively (recalling $p = \widetilde{O}(\epsilon^{-0.5})$)

$$\widetilde{O}\left(\frac{nd}{\epsilon^{4.5}}\right),\ \widetilde{O}\left(\frac{ndt}{\epsilon^{1.5}}\right),\ \widetilde{O}\left(ndt\right).$$

Throughout we use that we can compute matrix-vector products in an arbitrary linear combination of the $X_i X_i^\top$ in time $O(nd)$; it is easy to check that in all runtime guarantees, nnz can be replaced by this computational cost. Combining these bounds yields the final conclusion. □