Proceedings of Machine Learning Research vol 125:1–40, 2020 33rd Annual Conference on Learning Theory

Estimating Principal Components under Adversarial Perturbations

Pranjal Awasthi* PRANJALAWASTHI@GOOGLE.COM
Google and Rutgers University

Xue Chen XUE.CHEN1@NORTHWESTERN.EDU
Northwestern University

Aravindan Vijayaraghavan† ARAVINDV@NORTHWESTERN.EDU
Northwestern University

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Robustness is a key requirement for widespread deployment of machine learning algorithms, and has received much attention in both statistics and computer science. We study a natural model of robustness for high-dimensional statistical estimation problems that we call the *adversarial perturbation model*. An adversary can perturb *every* sample arbitrarily up to a specified magnitude δ measured in some ℓ_q norm, say ℓ_∞. Our model is motivated by emerging paradigms such as *low precision machine learning* and *adversarial training*.

We study the classical problem of estimating the top-r principal subspace of the Gaussian covariance matrix in high dimensions, under the adversarial perturbation model. We design a computationally efficient algorithm that given corrupted data, recovers an estimate of the top-r principal subspace with error that depends on a robustness parameter κ that we identify. This parameter corresponds to the $q \to 2$ operator norm of the projector onto the principal subspace, and generalizes well-studied analytic notions of sparsity. Additionally, in the absence of corruptions, our algorithmic guarantees recover existing bounds for problems such as sparse PCA and its higher rank analogs. We also prove that the above dependence on the parameter κ is almost optimal asymptotically, not just in a minimax sense, but remarkably for *every* instance of the problem. This *instance-optimal* guarantee shows that the $q \to 2$ operator norm of the subspace essentially *characterizes* the estimation error under adversarial perturbations.

1. Introduction

An important and active area of research in machine learning is the design of algorithms that are robust to modeling errors, noise and adversarial corruptions of different kinds. There is a rich body of work in the field of statistics, machine learning and theoretical computer science studying different models of robustness and the associated tradeoffs (e.g. Huber, 2011; Tukey, 1975; Hampel et al., 1986; Diakonikolas et al., 2019; Lai et al., 2016). In the context of statistical estimation problems the most widely studied model is Huber's ε-contamination

* PA acknowledges support from the National Science Foundation for the TRIPODS DATA-INSPIRE Institute through the award CCF-1934924.
† The last author is supported by the National Science Foundation (NSF) under Grant No. CCF-1652491, CCF-1637585 and CCF 1934931.

© 2020 P. Awasthi, X. Chen & A. Vijayaraghavan.

model (Huber, 2011). In Huber's model it is assumed that a small ε fraction of the data set is corrupted arbitrarily. The remaining portion of the dataset that is left uncorrupted is assumed to be generated from a structured distribution such as a Gaussian. Other notions of robustness that have been explored in unsupervised learning include distribution closeness of different kinds (Gao et al., 2019) and different semi-random models (Blum and Spencer, 1995; Feige and Kilian, 2001; Makarychev et al., 2012). Please see Appendix A for more detailed comparisons.

However there are several existing and emerging scenarios, where the data corruptions are not captured by these existing models of robustness. In many practical settings every data point is likely to perturbed with some small amount of noise, arising from various complex sources of errors. The reliability and security of learning algorithms could also be compromised by small imperceptible perturbations to the samples that are adversarial in nature (data poisoning). Moreover, *adversarial training* has emerged as a popular training paradigm where at each stage, the given training set is corrupted by adding (imperceptible) adversarial perturbations (typically measured in ℓ_∞ or ℓ_2 norm) (Madry et al., 2017), before performing stochastic gradient descent updates. This is empirically known to lead to more robust algorithms and also has implications for fair classification (Madras et al., 2018).

Data corruptions also arise naturally in popular emerging paradigms like *low-precision machine learning* (De Sa et al., 2017, 2018). Low precision computation gives substantial savings in time and energy costs by storing and processing only a few most significant bits e.g., 8-bit arithmetic is a popular choice. The lower memory utilization from low precision allows for processing of more training examples at the cost of quantization noise. This quantization noise is naturally captured as a small adversarial perturbation to every co-ordinate of the data point to an amount that depends on the number of bits used in the arithmetic (an ℓ_∞ norm bound). These adversarial perturbations lead to new tradeoffs in the estimation accuracy that are not well understood for many basic statistical tasks. In this work we take a step in this direction by studying a model of adversarial perturbations aimed at capturing the above scenarios.

Adversarial Perturbation model. We consider a natural model of robustness where *every* sample can be perturbed adversarially up to a bounded amount δ, say in ℓ_∞ norm (more generally, in ℓ_q norm where $q \in (2, \infty]$). In our model the input data $\tilde{A} \in \mathbb{R}^{m \times n}$ consisting of m samples in $\mathbb{R}^n$ is generated as follows:

1. The uncorrupted samples $A_1, \ldots, A_m \in \mathbb{R}^n$ are drawn i.i.d. from a Gaussian $\mathcal{N}(\mu, \Sigma)$, with unknown mean $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$.

2. An adversary can observe the samples $A_1, \ldots, A_m$, and perturb them arbitrarily to form $\tilde{A}_1, \ldots, \tilde{A}_m \in \mathbb{R}^n$ such that for each $j \in [m]$, $\|\tilde{A}_j - A_j\|_q \leq \delta$. These adversarial perturbations can be correlated.

We study the classical unsupervised learning problem of estimating the top-r principal subspace of the covariance matrix Σ, and the best rank-r approximation to Σ, for a specified $r \in [n]$. For $r = 1$, this corresponds to recovering the principal component of Σ.

In the above model, the adversarial perturbations are measured in ℓ_q norm where $q \in (2, \infty]$. As q goes to ∞, the perturbations become larger in magnitude and less

constrained. When $q = \infty$, *every co-ordinate* of *every point* can get perturbed adversarially up to δ in magnitude. For the sake of exposition, we will focus on the case of $q = \infty$ and present results for general $q \in (2, \infty]$ in the respective sections.

Our algorithms and guarantees will depend on certain quantity that we will call the *robustness parameter* κ, which captures the $q \to 2$ operator norm of the projector on to the target rank-r subspace, and generalizes analytic notions of sparsity. Surprisingly, we will see that this robustness parameter will be crucial in characterizing the estimation error under our model. To understand why sparsity (and the $\infty \to 2$ operator norm) is related to robustness under adversarial perturbations, let us first consider the simpler setting of mean estimation.

Warm up: Mean Estimation. Consider the problem of mean estimation where the uncorrupted data in $\mathbb{R}^n$ is generated from $\mathcal{N}(\mu, I)$. A valid ℓ_∞ adversarial perturbation is moving each of the samples by the same vector $z = \delta(1, 1, \ldots, 1)$, thereby moving the mean to μ' with $\|\mu' - \mu\|_2^2 = \delta^2 n$. In this case no estimator can tell apart μ, μ' from the data, hence this error of $\delta^2 n$ is unavoidable in the worst-case. Suppose however that mean μ was k-sparse i.e., it is supported on the set S of size at most $k \ll n$. If the support S is known beforehand, then by taking the empirical mean after projecting all the samples onto the support S, we can find an estimate $\widehat{\mu}$ with $\|\widehat{\mu} - \mu\|_2^2 \le \delta^2 k \ll \delta^2 n$ asymptotically (as the number of samples goes to infinity). While we do not know the the sparse support of μ beforehand[1]. the following proposition shows that one can indeed achieve the above improved rate when the mean is sparse in an analytic sense (the ratio of norms ℓ_1/ℓ_2).

Proposition 1 (Mean Estimation under Adversarial Perturbations) *Suppose we have m samples drawn according to the Adversarial Perturbation model with mean μ, covariance $\Sigma \preceq \sigma^2 I$ and $q = \infty$. There is a polynomial time algorithm (Algorithm 3) that outputs an estimate $\hat{\mu}$ for the (unknown) mean μ such that with probability at least $(1 - 1/n)$,*

$$\|\hat{\mu} - \mu\|_2^2 \le 4 \min\left\{ \|\mu\|_1(\delta + \eta), n(\delta + \eta)^2 \right\}, \text{ where } \eta := 2\sigma\sqrt{(\log n)/m}. \tag{1}$$

See Proposition 39 for general statement for all ℓ_q norms. If we use $\kappa = \frac{\|\mu\|_1}{\|\mu\|_2}$ to denote the analytic sparsity of μ, the first error term becomes $\kappa \cdot (\delta + \eta) \cdot \|\mu\|_2$. In fact, the above error of $\Omega(\kappa\delta\|\mu\|_2)$ is unavoidable for *every* instance for a broad range of parameters i.e., for every instance of the problem, there exists an adversarial perturbation that makes it statistically impossible to recover the mean with error $o(\delta\|\mu\|_1)$ (see Proposition 41).

Robustness Parameter κ. Similarly the estimation rates for finding the top-r principal subspace (or best rank-r approximation) of Σ will be characterized by the robustness parameter κ that is given by the $\infty \to 2$ operator norm:

$$\|\Pi\|_{\infty \to 2} = \max_{y : \|y\|_\infty \le 1} \|\Pi y\|_2,$$

where Π is the (orthogonal) projection matrix onto the subspace spanned by the top-r eigenvectors of Σ (for general q, the robustness parameter will correspond to $\|\Pi\|_{q \to 2}$ operator

1. This estimation problem is interesting even in the absence of adversarial perturbations, and corresponds to the *sparse mean estimation* problem that has been studied extensively in high-dimensional statistics Johnstone et al. (1994); Donoho et al. (1992); Donoho and Johnstone (1994).

norm). This robustness parameter generalizes analytic notions of sparsity (the ratio of ℓ_1/ℓ_2 norms) to projection matrices of subspaces[2]. Note that κ takes values in $[1, \sqrt{n}]$. The $\infty \to 2$ operator norm is also related to the famous Grothendieck inequality from functional analysis (Grothendieck, 1952; Alon and Naor, 2004). These parameters have also been used recently to characterize robustness to adversarial perturbations at test-time (Awasthi et al., 2019a) (see Section A for more discussion). Similar to mean estimation, the case of $r = 1$ for covariance estimation corresponds to the well studied sparse PCA problem (Johnstone et al., 2001; Amini and Wainwright, 2009; Ma et al., 2013; Vu and Lei, 2012, 2013; Berthet and Rigollet, 2013). Extensions of sparse PCA to estimating top r "sparse" subspaces have also been widely studied in the statistics community (Vu and Lei, 2013; Wang et al., 2014).

As we will see soon, our guarantees are not only minimax optimal in terms of these parameters, but they are essentially *instance-optimal*! Our upper bound and lower bound guarantees will work for *every* instance and will be tight up to logarithmic factors asymptotically (as number of samples becomes large). Hence our results give a surprising characterization of the estimation error under adversarial perturbations in terms of these robustness parameters (measured in $\infty \to 2$ norm), and highlight new robustness benefits of sparsity in high dimensional estimation.

1.1. Our Results

We now state our main results on recovering the principal subspace (and the best rank-r approximation) of the covariance Σ^* in terms of the $\infty \to 2$ operator norm of the corresponding rank-r projection matrix. The samples are drawn from the Adversarial Perturbation model where the covariance of the uncorrupted samples Σ^* has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. The unknown covariance matrix is split into $\Sigma = \Sigma_{\text{TOP}} + \Sigma_{\text{BOT}}$, where Σ_{TOP} corresponds to the best rank-r approximation of Σ i.e., the truncation of the SVD to the top-r eigenvalues $\lambda_1, \ldots, \lambda_r$. Let Π^* be the orthogonal projection matrix onto the span of Σ_{TOP}. We will assume that $\|\Pi^*\|_{\infty\to 2} \leq \kappa$. We will measure the estimation error in squared Frobenius norm. For the case of projection matrices, this is equivalent (up to a factor of 2) to the standard notion of subspace $\sin\Theta$ distance (see Appendix B).

Theorem 2 *[Algorithm] Suppose we have m samples drawn according to the the above Adversarial Perturbation model with (unknown) covariance Σ^* satisfying $\|\Pi^*\|_{\infty\to 2} \leq \kappa$. Assuming that $\kappa\delta \leq \frac{O(\lambda_r - \lambda_{r+1})}{\sqrt{r\lambda_1}}$, there exists an algorithm (Algorithm 2) that for any $\varepsilon > 0$ uses $m \geq Cr^2\kappa^4\big(\frac{\lambda_1^2}{(\lambda_r-\lambda_{r+1})^2}\big)\log n/\varepsilon^2$ samples and outputs a rank-r projection $\widehat{\Pi}$ with $\|\widehat{\Pi}\|_{\infty\to 2} = O(\kappa)$, and an estimate $\widehat{\Sigma}_{\text{TOP}}$ (restricted to the subspace $\widehat{\Pi}$) such that*

$$\|\widehat{\Pi} - \Pi^*\|_F^2 \leq \varepsilon_1 := \tfrac{\sqrt{\lambda_1}}{(\lambda_r-\lambda_{r+1})} \cdot O\big(\sqrt{r}\cdot\kappa\delta\big) + \varepsilon \text{ and } \|\widehat{\Sigma}_{\text{TOP}} - \Sigma_{\text{TOP}}\|_F^2 \leq O(\lambda_1^2\varepsilon_1 + \lambda_1\kappa^2\delta^2).$$

See Theorem 6 for the general statement for $q > 2$ and the proof. To interpret the results let's consider the case when $\Sigma^* = \theta\Pi^* + I$ (hence $\Sigma_{\text{TOP}} = (1+\theta)\Pi^*$), and $\theta = \Theta(1)$.[3] The above theorem shows that there is an efficient algorithm that obtains a rank-r projection $\widehat{\Pi}$

2. For the special case of a 1-dimensional subspace along the vector v, the orthogonal projector $\Pi_1 = \frac{1}{\|v\|_2^2}vv^\top$ satisfies $\|\Pi\|_{\infty\to 2} = \|\Pi\|_{2\to 1} = \|v\|_1/\|v\|_2$. See Fact 14 for details.
3. When $r = 1$, this special case is the sparse PCA setting where the principal component has ℓ_1 sparsity κ.

that is $O(\sqrt{r}\kappa\delta)$ close to Π^* in squared Frobenius norm, for sufficiently large polynomial m ($\widehat{\Pi}$ also has robustness parameter $O(\kappa)$). On the other hand, a random subspace of rank r will incur an error of $\Omega(r)$. Our algorithm can achieve an error of $o(1)$ while tolerating an additive perturbation that is as large as $\delta = o(1/(\sqrt{r}\kappa))$ (which could be $n^{-0.21}/\sqrt{r}$ if $\kappa = n^{0.2}$). On the other hand, if the top-r subspace has no special structure (robustness parameter $\kappa \approx \sqrt{n}$), then one requires $\delta = o(n^{-1/2}/\sqrt{r})$ for achieving similar error rates.

Next, we give a computational inefficient algorithm that achieves a better statistical rate in terms of the sample complexity.

Theorem 3 *[Statistical upper bound] Given m samples drawn according to the Adversarial Perturbation model with covariance Σ^* satisfying $\|\Pi^*\|_{\infty\to 2} \le \kappa$, there exists an algorithm that for any $\varepsilon > 0$ uses $m \ge Cr^2\kappa^2\big(\frac{\lambda_1^2}{(\lambda_r-\lambda_{r+1})^2}\big)\log n/\varepsilon^2$ samples and outputs a rank-r projection $\widehat{\Pi}$ with $\|\widehat{\Pi}\|_{\infty\to 2} \le \kappa$, and an estimate $\widehat{\Sigma}_{\text{TOP}}$ (restricted to the subspace $\widehat{\Pi}$) s.t.*

$$\|\widehat{\Pi} - \Pi^*\|_F^2 \le \varepsilon_1 := \tfrac{\sqrt{\lambda_1}}{(\lambda_r-\lambda_{r+1})} \cdot O\big(\sqrt{r}\cdot\kappa\delta\big) + \varepsilon \text{ and } \|\widehat{\Sigma}_{\text{TOP}} - \Sigma_{\text{TOP}}\|_F^2 \le O(\lambda_1^2\varepsilon_1 + \lambda_1\kappa^2\delta^2).$$

See Theorem 31 for the guarantees for general $q > 2$. The dominant error of $O(\sqrt{r}\kappa\delta)$ is the same for both Theorems 2 and 3, and represents the asymptotic error (error as $m \to \infty$). The main difference however is the number of samples m needed as a function of κ to drive the error to within ε of this asymptotic error. This gap of κ^4 vs κ^2 represents a computational vs statistical tradeoff that is unavoidable even when $r = 1$ (and $q = \infty$), assuming the hardness of the Planted Clique problem. This follows directly from computational lower bounds for sparse PCA with a $k = \kappa^2$-sparse vector (combinatorial sparsity) assuming Planted Clique hardness (Berthet and Rigollet, 2013; Gao et al., 2017). For smaller $q \in (2, \infty)$, there is an extra polynomial factor gap of $n^{2/q}$ in the sample complexity between Theorem 6 and Theorem 31 that would be interesting to resolve. Finally the estimation error in the absence of any adversarial errors is comparable to the existing state of the art results that are known to be tight (minimax optimal) (Vu and Lei, 2013; Awasthi et al., 2019a).

The following lower bound shows that our asymptotic error guarantees are almost optimal for *every* instance.

Theorem 4 *[Lower Bound] Suppose we are given parameters $r \in \mathbb{N}, \kappa \ge 2r$ and $\delta > 0$. In the notation of Theorem 3, for any Σ^*, given m samples $A_1, \dots, A_m$ generated i.i.d. from $\mathcal{N}(0, \Sigma^*)$ with $\kappa = \|\Pi^*\|_{\infty\to 2}$ satisfying $\sqrt{r\lambda_1}(\kappa/n) \le \delta \le \sqrt{r\lambda_1}/\kappa$, there exists a covariance matrix Σ' with a projector Π' onto its top-r principal subspace, and an alternate dataset $A'_1, \dots, A'_m$ drawn i.i.d. from $\mathcal{N}(0, \Sigma')$ satisfying $\|\Pi'\|_{\infty\to 2} \le (1 + o(1))\kappa$, and $\|A'_j - A_j\|_\infty \le \delta\ \forall j \in [m]$,*

$$\text{but } \|\Pi^* - \Pi'\|_F^2 \ge \big(\tfrac{\Omega(1)}{\sqrt{\lambda_1}\log(rm)\log n}\big)\cdot\sqrt{r}\kappa\delta, \text{ and } \quad \|\Sigma'_{\text{TOP}} - \Sigma_{\text{TOP}}\|_F^2 \ge \tfrac{(\lambda_1^2+\cdots+\lambda_r^2)}{r}\cdot\|\Pi' - \Pi^*\|_F^2$$

In particular, when $\Sigma_{\text{TOP}} = (1+\theta)\Pi^*$ then $\Sigma'_{\text{TOP}} = (1+\theta')\Pi'$ with $\theta' = (1 + o(1))\theta$.

See Section 4 for more details and proof of the construction, and Theorem 29 for the extension to general ℓ_q norms. Consider the previous setting where $\lambda_r - \lambda_{r+1} = \Omega(\lambda_1)$ and think of m as being any large polynomial in n. The above lower bound on the error

$\|\Pi' - \Pi^*\|_F^2 = \tilde{\Omega}(\sqrt{r}\kappa\delta)$ nearly matches the error bound obtain by our algorithm in Theorem 2 (as m becomes a sufficiently large polynomial and hence $\varepsilon \approx 0$) up to logarithmic factors, for *every* instance (i.e., every Π^*, Σ^*) i.e., our bounds are nearly *instance-optimal*. Note that this is much stronger than minimax optimality, which only requires the lower bounds to be tight for a specific choice of Σ^*, Π^*. Hence, Theorem 2 and Theorem 4 together show that the $\infty \to 2$ norm of the projection matrix essentially *characterizes* the robustness to training errors bounded in ℓ_∞ norm.

Discussion of the characterization. Our characterization of the robustness to adversarial perturbations is in terms of the robustness parameter $\kappa = \|\Pi^*\|_{\infty\to 2}$ ($\|\Pi^*\|_{q\to 2}$ for general q), which generalizes analytic notions of sparsity. For a $r = 1$-dimensional subspace, this exactly corresponds to the ℓ_1 sparsity of the unit vector v in that subspace. For higher-dimensional subspaces, there are several other notions of sparsity that have been explored (Vu and Lei, 2013; Wang et al., 2014). For a fixed orthonormal basis $V \in \mathbb{R}^{n\times r}$ of the subspace (so $\Pi^* = VV^\top$), some of the notions that have been considered include the entry-wise norm $\|V\|_1$ (the sum of the ℓ_1 norms of the basis vectors), the maximum ℓ_1 norm among the columns of V, the sparsity of the diagonal of Π^* and the sum of the row ℓ_2 norms of V, among other quantities. Many of these quantities are the same for $r = 1$ but may vary by factors of $\sqrt{r}$ or more depending on the quantity. On the other hand, our robustness parameter κ is a property only of the subspace and is basis independent. The $\|\Pi^*\|_{\infty\to 2}$ of a projector is the largest ℓ_1 norm among unit vectors (in ℓ_2 norm) that belong to the subspace.

Consider three different subspaces (or projectors) given by the orthonormal basis $V_1, V_2, V_3 \in \mathbb{R}^{n\times r}$ of the following form (think of $\kappa = \sqrt{k}$, $r \ll \kappa$); assume that the signs of the entries are chosen randomly in a way that also satisfies the necessary orthogonality properties (e.g., random Fourier characters over $\{\pm 1\}^k$).

$$V_1 = \begin{pmatrix} \frac{\pm 1}{\sqrt{k}} & \frac{\pm 1}{\sqrt{k}} & \cdots & \frac{\pm 1}{\sqrt{k}} \\ \frac{\pm 1}{\sqrt{k}} & \frac{\pm 1}{\sqrt{k}} & \cdots & \frac{\pm 1}{\sqrt{k}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\pm 1}{\sqrt{k}} & \frac{\pm 1}{\sqrt{k}} & \cdots & \frac{\pm 1}{\sqrt{k}} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \quad V_2 = \begin{pmatrix} \frac{\pm\sqrt{r}}{\sqrt{k}} & 0 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \frac{\pm\sqrt{r}}{\sqrt{k}} & 0 & \cdots & 0 \\ 0 & \frac{\pm\sqrt{r}}{\sqrt{k}} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & \frac{\pm\sqrt{r}}{\sqrt{k}} & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}, \quad V_3 = \begin{pmatrix} \frac{\pm 1}{\sqrt{r}} & \frac{\pm 1}{\sqrt{r}} & \cdots & \frac{\pm 1}{\sqrt{r}} & \frac{\pm 1}{\sqrt{k}} \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \frac{\pm 1}{\sqrt{r}} & \frac{\pm 1}{\sqrt{r}} & \cdots & \frac{\pm 1}{\sqrt{r}} & \frac{\pm 1}{\sqrt{k}} \\ 0 & 0 & \cdots & 0 & \frac{\pm 1}{\sqrt{k}} \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & \cdots & 0 & \frac{\pm 1}{\sqrt{k}} \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \end{pmatrix}$$

The main difference between V_1, V_2 is that in V_2 the sparse basis vectors have disjoint support, whereas in V_1 they are commonly supported. However, there is an alternate basis for the subspace V_2 which looks like V_1, but basis dependent quantities like the maximum ℓ_1 norm among columns get very different values for V_1, V_2. In the third example, the first $r-1$ basis vectors are extremely sparse with ℓ_1 norm $O(\sqrt{r})$, whereas only one of the basis vectors has ℓ_1 sparsity $\sqrt{k}$. Many aggregate notions of sparsity like $\|V\|_1$ or sum of the row ℓ_2 norms have very different values for V_1 and V_3 that differ by a $\sqrt{r}$ factor. On the other hand, our robustness parameter $\kappa \approx \sqrt{k}$; this is because each of these subspaces are supported on at most k co-ordinates (and a spread out vector of this form exists), so the maximum ℓ_1

length among unit ℓ_2 norm vector is $\sqrt{k}$. Hence, while our robustness parameter $\|\Pi^*\|_{\infty\to 2}$ *characterizes* the asymptotic error that can be obtained in all of these different cases (using Theorem 2 and Theorem 4), many other natural notions of sparsity are off by factors of $\sqrt{r}$ or more in at least one of these cases.

Finally, our robustness parameter κ also satisfies other useful properties like monotonicity (see Lemma 13), that will be very useful in the algorithm and analysis (this is not satisfied by various other norms like $\|\cdot\|_1$ etc.). While the $\infty \to 2$ operator norm is NP-hard to compute for PSD matrices, there exists polynomial time algorithms that can compute it up to a small constant factor (that corresponds to the Gröthendieck constant for PSD matrices) (see Nesterov, 1998; Alon and Naor, 2004).

Comparison to Prior Work and Related Work. There are several other notions of robustness that have been explored in both unsupervised and supervised learning. We place our work in the context of these existing works in Section A. The work that is closest to this paper is the recent work of Awasthi et al. (2019a). Our work is inspired by Awasthi et al. (2019a) and builds on some of those techniques. However, our work differs significantly from Awasthi et al. (2019a) both in terms of the problem focus, and the nature of the results, as we explain below. The main problem considered in Awasthi et al. (2019a) is finding a low-rank projection of a given data matrix A that achieves low approximation error, and is also robust to adversarial perturbations at *testing* time. Robustness at test time naturally places an upper bound constraint on the $q \to 2$ operator norm of the projection matrix. The paper also consider this problem under adversarial perturbations at training-time, and use these results as a black-box to obtain some guarantees for mean estimation and clustering in the presence of adversarial perturbations. The paper mainly studies the worst-case setting which is computationally hard, and hence focus on multiplicative approximation guarantees for an objective (like low-rank approximation error), as opposed to estimation or recovery.

On the other hand, the main focus of this paper is adversarial perturbations at *training time*; there is no requirement of robustness at testing-time. Hence, it is not clear why $\kappa = \|\Pi\|_{q\to 2}$ is a relevant parameter at all. The main message of this paper is that this parameter κ indeed characterizes the robustness to adversarial perturbations at training time as well (this is even if test-time robustness is not a consideration)! Moreover we focus on high-dimensional statistical estimation tasks where there is an underlying distribution for the uncorrupted data, and allows us to obtain the strong statistically optimal recovery guarantees. Hence the guarantees in the two works are incomparable.

2. Preliminaries

Norms. For a vector $v \in \mathbb{R}^n$ and any $q \geq 1$, we use $\|v\|_q$ to denote the q-norm: $\left(\sum_{i=1}^n |v(i)|^q\right)^{1/q}$. For any fixed $q \geq 1$, we use ℓ_{q^*} to denote the dual of ℓ_q, where $1/q + 1/q^* = 1$. We also apply Hölder's inequality extensively: $\forall q \geq 1$ and $u, v \in \mathbb{R}^n$, $\left|\langle u, v\rangle\right| \leq \|u\|_{q^*}\|v\|_q$. A direct corollary is that $\|v\|_q \leq |\text{support}(v)|^{1/q-1/p} \cdot \|v\|_p$ for any vector v and any $q < p$. In particular, $\|v\|_1 \leq \sqrt{k}$ for a unit vector v of sparsity k.

For a matrix $A \in \mathbb{R}^{n\times m}$ and $q \geq 1$, we will use $\|A\|_q$ to denote the entry-wise ℓ_q norm of A: $\left(\sum_{i,j} |A(i,j)|^q\right)^{1/q}$. When $q = 2$, we will also use the Frobenius norm $\|A\|_F \stackrel{\text{def}}{=} \|A\|_2$ equipped with trace inner product $\langle A, B\rangle = \text{tr}(A^\top B)$.

$p \to q$ norms. For any p and q, we define the operator $p \to q$ norm for a matrix $A \in \mathbb{R}^{n \times m}$:

$$\|A\|_{p \to q} = \max_{v \in \mathbb{R}^m \setminus \{0\}} \|Av\|_q / \|v\|_p.$$

For convenience, let $\|A\|$ denote the operator norm $\|A\|_{2 \to 2}$. A variational definition of the operator norm is as follows (See Section 4 in Awasthi et al. (2019a) for proofs).

Fact 5 *For any p and q, $\|A\|_{p \to q} = \max_{u \in \mathbb{R}^n \setminus \{0\}, v \in \mathbb{R}^m \setminus \{0\}} u^\top A v / (\|u\|_{q^*} \|v\|_p)$. Also, $\|A\|_{p \to q} = \|A^\top\|_{q^* \to p^*}$ and $\|A^\top A\|_{q \to q^*} = \|A\|_{q \to 2}^2$. In particular, $\|\Pi\|_{\infty \to 2} = \|\Pi\|_{2 \to 1}$ and $\|\Pi\|_{q \to q^*} = \|\Pi\|_{q \to 2}^2$ for projection matrices.*

Due to the space constraint, we defer a few properties of the operator norm to Appendix B.

3. Computational Upper Bound

In this section we present our computationally efficient algorithm for estimating the top-r principal subspace. We state our main claim regarding the error guarantees associated with the algorithm and describe the key ideas used in the analysis. All the proofs are deferred to Appendices D and E. A key subroutine in our algorithm is the following convex program that was proposed in Awasthi et al. (2019a). We use the program will be run on the corrupted data $\tilde{A}$ and the bulk of our analysis will involve showing that the solution output by the program can be used for estimation in spite of adversarial perturbations. The program takes in as parameters the rank r and an upper bound for the robustness parameter κ, whose target solution is the projection Π^* of Σ^*.

$$\min \frac{1}{m}\|\tilde{A}\|_F^2 - \frac{1}{m}\langle \tilde{A}\tilde{A}^\top, X \rangle \tag{2}$$

$$\textit{subject to} \quad \operatorname{tr}(X) \le r \tag{3}$$

$$0 \preceq X \preceq I \tag{4}$$

$$\|X\|_q \le r\kappa^2 \tag{5}$$

$$\|X\|_{q \to q^*} \le \kappa^2 \tag{6}$$

One can use the Ellipsoid algorithm to efficiently solve the program above via an efficient separation oracle (See Lemma 15). We briefly discuss the last two constraints in the above program and refer to Awasthi et al. (2019a) for a more detailed discussion: The constraint (5) is based on the fact that the projection $\Pi^* = \sum_{i=1}^r v_i v_i^\top$ has each $\|v_i\|_{q^*} \le \kappa$. At the same time, the last constraint (6) is based on the monotonicity of $q \to q^*$ norms from Lemma 13.

Below is the algorithm that uses the SDP solution above to outputs a robust projection matrix $\widehat{\Pi}$ of rank at most r.

Algorithm 1 Finding Robust Low-Rank Projection

1: **function** RobustProjection(data matrix $\tilde{A} \in \mathbb{R}^{m \times n}$, rank r, robustness κ, norm q)
2: Solve (2) on $\tilde{A}$ with parameters κ, q, r to find a solution $\widehat{X} \succeq 0$ (see Lemma 15).
3: Use SVD on $\widehat{X}$ to find the subspace spanned by the top-r eigenvectors of $\widehat{X}$. Output $\widehat{\Pi}$, the orthogonal projection matrix onto this subspace.

Finally, our algorithm for estimating the principal components of the covariance matrix in the presence of adversarial perturbations, described below, just uses RobustProjection as an additional pre-processing step to find a suitable robust subspace for computing the empirical covariance.

Algorithm 2 Principal Subspace Estimation under Adversarial Perturbations

function AdvRobustPCA($4m$ samples $\tilde{A}_1, \ldots, \tilde{A}_{4m} \in \mathbb{R}^n$, rank r, robustness κ, q)
  Split samples into two equal parts. Let $A^{(1)}, A^{(2)}$ denote these two datasets.
  For each $j \in [m]$, let $A'_j = \frac{1}{\sqrt{2}}(\tilde{A}_j - \tilde{A}_{m+j})$ and let $A''_j = \frac{1}{\sqrt{2}}(\tilde{A}_{2m+j} - \tilde{A}_{3m+j})$.
  Run RobustProjection(A', r, κ, q) to find a r-dimensional projection matrix $\widehat{\Pi}$.
  Output $\widehat{\Sigma}_r$ to be empirical covariance of $\widehat{\Pi}A''$.

Next, we state our main theorem regarding the estimation error associated with the algorithm above. We state the guarantee for a general $q \geq 2$. Substituting $q = \infty$ recovers the guarantee stated in Theorem 2.

Theorem 6 *Given $q \geq 2$, r, and κ, let $\widetilde{A} \in \mathbb{R}^{n \times m}$ be a δ-perturbation (in ℓ_q norm) of data points generated from $\mathcal{N}(0, \Sigma^*)$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of the covariance matrix Σ^* and Π^* be the projection matrix on to the top r eigenspace of Σ^*. There exists a universal constant C such that for any $\varepsilon > 0$, and $\kappa\delta \leq \frac{\lambda_r - \lambda_{r+1}}{C\sqrt{r\lambda_1}}$, Algorithm 2 when provided with $m \geq Cr^2\kappa^4 \cdot \frac{\lambda_1^2}{(\lambda_r - \lambda_{r+1})^2} \log n \cdot \frac{n^{4/q}}{\varepsilon^2}$ samples, outputs with probability at least 0.99 $\widetilde{\Sigma}_{\text{TOP}}$ of rank r and the projector onto its subspace $\widetilde{\Pi}$ that satisfies $\|\widetilde{\Pi}\|_{q \to 2} = O(\kappa)$,*

$$\|\widetilde{\Pi} - \Pi^*\|_F^2 \leq O\Big(\frac{\sqrt{\lambda_1 r} \cdot \kappa\delta}{\lambda_r - \lambda_{r+1}}\Big) + \varepsilon \text{ and } \|\widetilde{\Sigma}_{\text{TOP}} - \Sigma_{\text{TOP}}\|_F^2 \leq O\Big(\lambda_1^2 \|\widetilde{\Pi} - \Pi^*\|_F^2 + \lambda_1 \kappa^2 \delta^2\Big).$$

We describe the key ideas and supporting claims that are used in our analysis. Due to the space constraint, we will defer the formal proof of Theorem 6 to Appendix D.1. The proof consists of three main steps. We first argue about the error of the estimated projection matrix $\widetilde{\Pi}$ with respect to Π^*. One can show that the optimal solution to the convex program (2) (that we will refer to as the SDP) on the ideal instance $\mathbb{E}[AA^\top]$ in fact recovers the projection Π^*. However the SDP is solved on the given instance $\mathbb{E}[AA^\top] + E$ where E is the error matrix defined as $E := \frac{1}{m}\tilde{A}\tilde{A}^\top - \mathbb{E}[AA^\top]$ involving both the adversarial perturbations and sampling errors. The first part of the argument for the robustness of the SDP to adversarial perturbations is by providing an upper bound on $|\langle E, X\rangle|$ over all feasible SDP solutions X. Lemma 7 that is stated below crucially uses the constraints on $\|X\|_{q \to q^*}$ and $\|X\|_{q^*}$ to provide the required bound.

Lemma 7 *Let $\tilde{A}$ be a δ-perturbation (in ℓ_q norm) of the original data matrix A where $\mathbb{E}[AA^\top] = \Sigma^*$. Let $E := \frac{1}{m}\tilde{A}\tilde{A}^\top - \mathbb{E}[AA^\top]$ denote the error matrix and define*

$$\mathcal{P}_{c(q)} = \{X \in \mathbb{R}^{n \times n} : tr(X) = r, 0 \preceq X \preceq I, \|X\|_{q^*} \leq r\kappa^2, \|X\|_{q \to q^*} \leq c(q) \cdot \kappa^2\}$$

as the set of all solutions that can be obtained by solving the SDP in (2) via the Ellipsoid Algorithm (see Lemma 15). With high probability, $\Delta := \sup_{X \in \mathcal{P}_{c(q)}} |\langle E, X\rangle|$ satisfies

$$\Delta \leq O\Big(\sqrt{r \cdot \lambda_{\max}(\Sigma^*)}\kappa\delta + \kappa^2\delta^2 + \frac{r\kappa^2 \cdot \lambda_{\max}(\Sigma^*)\sqrt{\log n} \cdot n^{2/q}}{\sqrt{m}}\Big).$$

A key technical lemma that helps to establish the above bound is stated below.

Lemma 8 *Let $A_1, A_2, \ldots, A_m \in \mathbb{R}^n$ be generated i.i.d. from $\mathcal{N}(\mu, \Sigma^*)$. Let A be the $n \times m$ matrix with the columns being the points A_i. Let X be a solution to the SDP in program (2) and let B be any matrix, potentially chosen based on A, with $\|B_j\|_q \leq \delta \ \forall j \in [m]$. Then with probability at least $1 - \frac{1}{\text{poly}(n)}$ we have that*

$$\frac{1}{m}\Big|\langle (A - \mathbb{E}[A])B^T, X\rangle\Big| \leq O(\sqrt{r\|\Sigma^*\|}\kappa\delta) + O(\kappa^2\delta^2) + O\Big(\frac{r\kappa^2\|\Sigma^*\|\sqrt{\log n}\cdot n^{2/q}}{\sqrt{m}}\Big). \tag{7}$$

We defer the proof of Lemma 7 to Section D.2. The second step of the proof lower bounds the correlation of the SDP solution to Π^* in terms of the value obtained by the SDP solution on the ideal instance $\Sigma^* = \mathbb{E}[AA^\top]$. This is established in following claim whose proof is deferred to Section D.3.

Claim 9 *Given a PSD matrix Σ^*, let Π^* be the projection matrix on to the top r eigenspace of Σ^*. For any matrix X with $tr(X) = r$ and $0 \preceq X \preceq I$, it holds that*

$$\langle X, \Pi^*\rangle \geq r - \frac{\langle \Pi^*, \mathbb{E}[AA^\top]\rangle - \langle X, \mathbb{E}[AA^\top]\rangle}{\lambda_r - \lambda_{r+1}} = r - \frac{\langle \Pi^*, \Sigma^*\rangle - \langle X, \Sigma^*\rangle}{\lambda_r - \lambda_{r+1}}.$$

where λ_r and λ_{r+1} denote the rth and the $(r+1)$th largest eigenvalues of Σ^* respectively.

The above claim helps us argue that by truncating X to its top-r subspace we get a good approximation to Π^*. Finally, in the theorem below we show how to recover the top-r principal component Σ^* given $\widetilde{\Pi}$ that is a good estimate of Π^*.

Theorem 10 *Let $A_1, \ldots, A_m$ be data points drawn independently from $\mathcal{N}(0, \Sigma^*)$ where the covariance matrix $\Sigma^* = \sum_{i=1}^n \lambda_i v_i v_i^\top$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Let $\Sigma_{\text{TOP}} = \sum_{i=1}^r \lambda_i v_i v_i^\top$ and Π^* denote the projection matrix on to the eigenspace of Σ_{TOP}. Furthermore, let Π be a rank r projection matrix with $\|\Pi - \Pi^*\|_F^2 \leq \varepsilon$. Then given a delta perturbation $\widetilde{A}_1, \ldots, \widetilde{A}_m$, with probability at least 0.99 (over $A_1, \ldots, A_m$), the matrix $\widetilde{\Sigma}_{\text{TOP}} = \Pi\frac{1}{m}(\sum_{i=1}^m \widetilde{A}_i\widetilde{A}_i^\top)\Pi$ satisfies*

$$\|\widetilde{\Sigma}_{\text{TOP}} - \Sigma_{\text{TOP}}\|_F^2 = O(\lambda_1^2\varepsilon + \frac{\lambda_1^2 r^2}{m} + \kappa^4\delta^4 + \lambda_1 \cdot \kappa^2\delta^2) \text{ when } m = \Omega(\lambda_1^2 r^2).$$

Due to the space constraint, we defer the proof of Theorem 10 to Appendix D.4 and the proof of our main result, Theorem 6, to Appendix D.1.

4. Statistical Lower Bound and Instance-Optimality

We now describe the construction that establishes Theorem 4, the instance-optimal lower bound for recovering the principal subspace of a covariance matrix under adversarial perturbations. Recall that we have an arbitrary covariance matrix Σ^* with eigendecomposition $\Sigma^* = \sum_{i=1}^n \lambda_i v_i v_i^\top$ and $\Pi^* = \sum_{i=1}^r v_i v_i^\top$ being the projection matrix onto its top-r subspace. We construct based on Π another rank-r projection matrix Π' (and a corresponding Σ') s.t.

$$\|\Pi' - \Pi^*\|_F^2 \geq \frac{c\sqrt{r}\kappa\delta}{\sqrt{\lambda_1}\log(rm)\log n} \text{ and } \|\Sigma_{\text{TOP}} - \Sigma'_{\text{TOP}}\|_F^2 = \Omega\Big(\frac{\lambda_1^2 + \cdots + \lambda_r^2}{r} \cdot \|\Pi' - \Pi^*\|_F^2\Big),$$

and $\|\Pi'\|_{\infty\to 2} \le (1+o(1))\kappa$. Moreover, for any data matrix A composed of m samples generated from $\mathcal{N}(0,\Sigma)$, we prove that with high probability, $\exists$ a coupled data matrix $A' \in \mathbb{R}^{n\times m}$ generated from $\mathcal{N}(0,\Sigma')$ satisfying $\|A_j - A'_j\|_\infty \le \delta$.

We remark that our construction also extends in a straightforward fashion to general ℓ_q norms to also give the same asymptotic lower bound of $\tilde{\Omega}(\sqrt{r/\lambda_1}\cdot\kappa\delta)$, where the $\tilde{\Omega}$ hides polylogarithmic factors. We sketch the differences in the intermediate claims between the ℓ_∞ and general ℓ_q norm in the appendix (see Section F.3.1). To interpret the results, let $\lambda_1 = O(1)$, and let $\kappa \gg r$ (say $\kappa = n^{0.2}$ and $r = n^{0.1}$). The theorem gives a lower bound of $\tilde{\Omega}(\sqrt{r}\kappa\delta)$, which is meaningful when $\kappa\delta \le \sqrt{r}$; also δ can not be too small. The range of δ is quite natural (for e.g., it is $[n^{-0.85}, n^{-0.15}]$ for the above setting). Theorem 4 shows that the upper bounds are optimal up to poly-logarithmic factors for *every* principal subspace Π^* with $\|\Pi^*\|_{\infty\to 2} = \kappa$. The lower bound does not have the optimal dependence in terms of the gap between the eigenvalues $(\lambda_r - \lambda_{r+1})/\lambda_1$. Please also see Theorem 25 in the appendix for a simpler minimax lower bounds that achieves the correct dependence on the eigengap as well.

Construction. To construct Π' we take the basis vectors $v_1,\dots,v_r$ and add carefully chosen small perturbations $u_1,\dots,u_r$ to them to get a new basis $v'_1,\dots,v'_r$. Set $k' := \sqrt{\frac{\lambda_1}{r}}\cdot\left(\frac{\kappa}{\delta}\right)$ and $\varepsilon := \frac{c}{\log(rm)\log n}(\delta\kappa/\sqrt{r\lambda_1})$ for a small constant $c>0$. Note that $\varepsilon \in [0,\frac{1}{4})$ and $2r \le k' \le n/r$ from our choice of parameters. Let $S_1, S_2, \dots, S_r \subset \{1,\dots,n\}$ be arbitrary disjoint subsets of size k' each. Let for each $\ell\in[r]$, T_ℓ denote the subspace of dimension $d_\ell \ge k' - r \ge k'/2$ that corresponds to the subspace of $\mathbb{R}^{S_\ell}$ that is orthogonal to Π^* and let $\Pi_\ell^\perp \in \mathbb{R}^{n\times n}$ be its projector. Then we define the eigenvectors $v'_1,\dots,v'_r$ of Σ', while $v'_{r+1} = v_{r+1},\dots,v'_n = v_n$.

$$\forall \ell\in[r],\ \ u_\ell = \Big(\frac{1}{\sqrt{d_\ell}}\Big)\Pi_\ell^\perp g_\ell,\ \ \text{where } g_\ell \sim N(0, I_{n\times n}) \ \text{ independently.} \tag{8}$$

$$\text{Define, } \forall \ell\in[r],\ \ v'_\ell = (1-\varepsilon)v_\ell + \Big(\frac{\sqrt{2\varepsilon-\varepsilon^2}}{\|u_\ell\|_2}\Big)\,u_\ell. \tag{9}$$

Let Π' be the orthogonal projector on the subspace spanned by $v'_1,\dots,v'_\ell$. Recall $\forall j\in[m], A_j = \sum_{\ell=1}^n \zeta_\ell^{(j)}\sqrt{\lambda_\ell}\cdot v_\ell$ where $\zeta_\ell^{(j)} \sim N(0,1)$. We construct the alternate dataset A':

$$A'_j = \sum_{\ell=1}^r \zeta_\ell^{(j)}\sqrt{\lambda_\ell}\cdot\left(v_\ell + \Big(\frac{\sqrt{2\varepsilon-\varepsilon^2}}{(1-\varepsilon)\|u_\ell\|_2}\Big)u_\ell\right) + \sum_{\ell=r+1}^n \zeta_\ell^{(j)}\sqrt{\lambda_\ell}\cdot v_\ell. \tag{10}$$

(Note that the randomness in A_j and A'_j are coupled using the random variables $\{\zeta_\ell^{(j)} : \ell\in[r]\}, j\in[m]$.) Observe that each sample A'_j is also drawn independently from $\mathcal{N}(0,\Sigma')$ with

$$\Sigma' = \sum_{\ell=1}^r \lambda_\ell\Big(v_\ell + \big(\frac{\sqrt{2\varepsilon-\varepsilon^2}}{(1-\varepsilon)\|u_\ell\|_2}\big)u_\ell\Big)\Big(v_\ell + \big(\frac{\sqrt{2\varepsilon-\varepsilon^2}}{(1-\varepsilon)\|u_\ell\|_2}\big)u_\ell\Big)^\top + \sum_{\ell=r+1}^n \lambda_\ell v_\ell v_\ell^\top.$$

Its best rank-r approximation is $\Sigma'_{\text{TOP}} := \frac{1}{(1-\varepsilon)^2}\sum_{\ell=1}^r \lambda_\ell v'_\ell (v'_\ell)^\top$, where v'_ℓ is defined in (9). Moreover $v'_1,\dots,v'_r$ are orthonormal (since $u_1,\dots,u_r$ are mutually orthonormal and orthogonal to Π^*). Hence $\Pi' = \sum_{\ell=1}^r v'_\ell(v'_\ell)^\top$, and the top r eigenvalues of Σ' are $\{\lambda_\ell/(1-\varepsilon)^2 : \ell\in[r]\}$.

For our construction to work the u_i vectors must simultaneously satisfy a few properties. They must be (i) orthogonal to the given Π^*, (ii) have disjoint support, (iii) be sufficiently sparse, and (iv) and have sufficiently small ℓ_∞ norm. Ensuring these properties requires a careful balancing act, and the following lemma gives an appropriate random distribution that satisfies these properties.

Lemma 11 *The vectors $u_1, u_2, \ldots, u_r \in \mathbb{R}^n$ have disjoint supports $S_1, S_2, \ldots, S_r \subset [n]$, and $\Pi^* u_1 = \Pi^* u_2 = \cdots = \Pi^* u_r = 0$. Moreover given $k' \geq 2r$, for any $\eta < 1$, with probability at least $(1-\eta)$ we have*

$$\forall \ell \in [r], \qquad \left| \|u_\ell\|_2^2 - 1 \right| \leq 3\sqrt{\log(r/\eta)/k'} + 4\log(r/\eta)/k' \tag{11}$$

$$\|u_\ell\|_\infty \leq 3\sqrt{\log(rk'/\eta)/k'}. \quad \text{and} \quad \|u_\ell\|_1 \leq 2\sqrt{k'}. \tag{12}$$

The final hurdle in the construction comes from arguing that $\|\Pi'\|_{\infty\to 2}$ is comparable to $\|\Pi\|_{\infty\to 2}$. We argue this by analyzing the related $\|\Pi'\|_{\infty\to 1}$ norm instead which is known to have good monotonicity properties (see Lemma 13), and by using properties of $v_1, \ldots, v_r$ that follow from $\|\Pi^*\|_{\infty\to 2} = \kappa$. Please see Section F.3 for the proof of the theorem, and Section F.1 for proofs of the related lemmas.

Acknowledgments

The authors would like to think Sivaraman Balakrishnan for several helpful discussions, and for suggesting the thresholding algorithm for mean estimation.

References

Noga Alon and Assaf Naor. Approximating the cut-norm via grothendieck's inequality. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 72–80. ACM, 2004.

Arash A. Amini and Martin J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, 37:2877–2921, 2009.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4): 343–370, 1988.

Pranjal Awasthi and Aravindan Vijayaraghavan. Towards learning sparsely used dictionaries with arbitrary supports. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 283–296. IEEE, 2018.

Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 449–458. ACM, 2014.

Pranjal Awasthi, Vaggos Chatziafratis, Xue Chen, and Aravindan Vijayaraghavan. Adversarially robust low dimensional representations. *arXiv preprint arXiv:1911.13268*, 2019a.

Pranjal Awasthi, Abhratanu Dutta, and Aravindan Vijayaraghavan. On robustness to adversarial examples and polynomial optimization. In *Advances in Neural Information Processing Systems*, pages 13737–13747, 2019b.

Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212, 2017.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT*, pages 1046–1066, 2013.

Avrim Blum and Joel Spencer. Coloring random and semi-random k-colorable graphs. *J. Algorithms*, 19:204–234, September 1995. ISSN 0196-6774. doi: http://dx.doi.org/10.1006/jagm.1995.1034. URL http://dx.doi.org/10.1006/jagm.1995.1034.

Avrim Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1-2):35–52, 1998.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. ACM, 2017.

Mengjie Chen, Chao Gao, Zhao Ren, et al. A general decision theory for huber's ε-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.

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. URL http://proceedings.mlr.press/v75/cheng18b.html.

Alexandre d'Aspremont, Laurent E Ghaoui, Michael I Jordan, and Gert R Lanckriet. A direct formulation for sparse pca using semidefinite programming. In *Advances in neural information processing systems*, pages 41–48, 2005.

Fernando De La Torre and Michael J Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1-3):117–142, 2003.

Christopher De Sa, Matthew Feldman, Christopher Ré, and Kunle Olukotun. Understanding and optimizing asynchronous low-precision stochastic gradient descent. In *ACM SIGARCH Computer Architecture News*, volume 45, pages 561–574. ACM, 2017.

Christopher De Sa, Megan Leszczynski, Jian Zhang, Alana Marzoev, Christopher R Aberger, Kunle Olukotun, and Christopher Ré. High-accuracy low-precision training. *arXiv preprint arXiv:1803.03383*, 2018.

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. Society for Industrial and Applied Mathematics, 2018a.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018b.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1061–1073. ACM, 2018c.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

David L Donoho and Iain M Johnstone. Minimax risk overl p-balls forl p-error. *Probability Theory and Related Fields*, 99(2):277–303, 1994.

David L Donoho, Iain M Johnstone, Jeffrey C Hoch, and Alan S Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):41–67, 1992.

John Dunagan and Santosh Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. *Mathematical Programming*, 114(1):101–114, 2008.

Abhratanu Dutta, Aravindan Vijayaraghavan, and Alex Wang. Clustering stable instances of euclidean k-means. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6503–6512, Red Hook, NY, USA, 2017. ISBN 9781510860964.

Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. *J. Comput. Syst. Sci.*, 63:639–673, December 2001. ISSN 0022-0000. doi: 10.1006/jcss.2001.1773. URL http://dl.acm.org/citation.cfm?id=569473.569481.

Chao Gao, Zongming Ma, Harrison H Zhou, et al. Sparse cca: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

Chao Gao, Jiyi Liu, Yuan Yao, and Weizhi Zhu. Robust estimation via generative adversarial networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJgRDjR9tQ.

Alexander Grothendieck. Résumé des résultats essentiels dans la théorie des produits tensoriels topologiques et des espaces nucléaires. In *Annales de l'institut Fourier*, volume 4, pages 73–112, 1952.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, Inc, 1986.

Peter J Huber. *Robust statistics*. Springer, 2011.

Iain M Johnstone et al. On minimax estimation of a sparse normal mean vector. *The Annals of Statistics*, 22(1):271–289, 1994.

Iain M Johnstone et al. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008a.

Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 629–638. ACM, 2008b.

Adam Tauman Kalai, Varun Kanade, and Yishay Mansour. Reliable agnostic learning. *Journal of Computer and System Sciences*, 78(5):1481–1495, 2012.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Justin Khim and Po-Ling Loh. Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. *arXiv preprint arXiv:1803.03241*, 2018.

Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(Dec):2715–2740, 2009.

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.

Jerry Li. Robust sparse estimation tasks in high dimensions. *arXiv preprint arXiv:1702.05860*, 2017.

Zongming Ma et al. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Approximation algorithms for semi-random partitioning problems. In *Proceedings of the 44th Symposium on Theory of Computing (STOC)*, pages 367–384. ACM, 2012.

Shahar Mendelson. Empirical processes with a bounded $\psi 1$ diameter. *Geometric and Functional Analysis*, 20(4):988–1027, Oct 2010. doi: 10.1007/s00039-010-0084-5. URL https://doi.org/10.1007/s00039-010-0084-5.

Ankur Moitra, William Perry, and Alexander S. Wein. How robust are reconstruction thresholds for community detection. *CoRR*, abs/1511.01473, 2015.

Preetum Nakkiran. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

Yu Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization methods and software*, 9(1-3):141–160, 1998.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

Daureen Steinberg. Computation of matrix norms with applications to robust optimization. *Research thesis, Technion-Israel University of Technology*, 2005.

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

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

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.

Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.

Aravindan Vijayaraghavan and Pranjal Awasthi. Clustering semi-random mixtures of gaussians. In *International Conference on Machine Learning*, pages 5055–5064, 2018.

Vincent Vu and Jing Lei. Squared-norm empirical process in banach space. https://arxiv.org/abs/1312.1005, 2012.

Vincent Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *In: Ann. Statist.*, pages 2905–2947, 2013.

Zhaoran Wang, Huanran Lu, and Han Liu. Tighten after relax: Minimax-optimal sparse PCA in polynomial time. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3383–3391, 2014.

Yannis G Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, pages 768–774, 1985.

Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. *arXiv preprint arXiv:1810.11914*, 2018.

Appendix A. Related Work

Robustness in Supervised Learning. In the context of supervised learning problems such as classification and regression various models of robustness have been studied in the literature. These include the classical random classification noise model (Angluin and Laird, 1988), the statistical query model (Kearns, 1998), and the agnostic learning (Kearns et al., 1994) framework for modeling corruptions to the training labels. Model such as malicious noise (Kearns and Li, 1993) and nasty noise (Diakonikolas et al., 2018c) study settings where both the training data and the training labels could be corrupted. Typically these models assume that only a small ε fraction of the training data can be corrupted by an adversary. The study of these models has been very fruitful leading to a variety of algorithmic insights (Blum et al., 1998; Dunagan and Vempala, 2008; Kalai et al., 2008a; Klivans et al., 2009; Kalai et al., 2008b, 2012; Awasthi et al., 2014; Diakonikolas et al., 2018c).

Recently, motivated from properties of deep neural networks, there has also been a lot in interest in modeling robustness to adversarial perturbations of the test input (Madry et al., 2017; Schmidt et al., 2018; Nakkiran, 2019; Khim and Loh, 2018; Yin et al., 2018; Tsipras et al., 2018; Awasthi et al., 2019b). While these works also model the noise as ℓ_p perturbations to the input, the theory of test time robustness is poorly understood and we lack provably robust algorithms for many fundamental tasks.

Robustness in Unsupervised Learning. There is a large body of literature in the machine learning and statistics community on the design and study of robust algorithms for unsupervised learning tasks. Perhaps the most popular and widely studied model in this context is Huber's ε-contamination model (Huber, 2011). Here is it assumed that a given data set is generated from a mixture: $(1-\varepsilon)P + \varepsilon Q$ where P is the true distribution about which we want to reason and Q is an arbitrary distribution. Various works have studied the computational and statistical tradeoffs under Huber's model for fundamental tasks such as mean/covariance estimation (Yatracos, 1985; Chen et al., 2016; Diakonikolas et al., 2019, 2018a; Charikar et al., 2017; Steinhardt et al., 2017; Balakrishnan et al., 2017; Li, 2017), regression (Prasad et al., 2018; Klivans et al., 2018) and more general stochastic convex optimization (Prasad et al., 2018; Diakonikolas et al., 2018b). Dutta et al. (2017) consider a notion of additive perturbation stability for Euclidean k-means clustering, where the optimal clustering is stable even when each point is perturbed by a small amount in ℓ_2 norm. Our results together indicate that the $\infty \to 2$ norm of the principal may analogously capture a notion of stability for the subspace estimation problem when the perturbations are measured in ℓ_∞ norm (or ℓ_q for $q > 2$).

Principal Subspace Estimation in High Dimensions. The results of our paper characterize the robustness to adversarial perturbations for estimating the top r-principal subspace of the covariance matrix in terms of the sparsity of the subspace. In the area of

high dimensional statistics questions of estimating mean and covariance with rates depending on various notions of sparsity have been widely studied. These works however assume that the dataset is indeed generated from the idealized model. There is a long line work on the classical problem of sparse mean estimation in high dimensions (Donoho et al., 1992; Donoho and Johnstone, 1994). For the case of covariance estimation the sparse PCA formulation has been well studied and essentially corresponds to estimating the top principal component assuming that it is ℓ_0 or ℓ_1 sparse (Johnstone et al., 2001; Berthet and Rigollet, 2013; Amini and Wainwright, 2009). The works of Vu and Lei (2012, 2013); Ma et al. (2013); Wang et al. (2014) extend this to estimating the top-r principal subspace with rates depending on certain notions of sparsity of the subspace. Similar to our work, semidefinite programming (SDP) based approaches have been proposed for such sparse estimation problems (d'Aspremont et al., 2005).

Another related setting is the robust PCA formulation that has received significant interest in recent years (De La Torre and Black, 2003; Candès et al., 2011; Chandrasekaran et al., 2011). Here one assumes that a given data matrix is the sum of a low rank matrix and a sparse matrix, i.e., the one with very few non-zero entries. In this case it can be shown that if true signal (the low rank component) is well spread out then estimation is possible. In contrast, in our setting every data point could be corrupted and hence the data matrix $\tilde{A}$ cannot be written as the sum of low rank plus a sparse component. In fact, our characterization implies that under our model of perturbations, estimation is possible if and only if the signal is localized, i.e., is sparse.

Robustness in Combinatorial Settings. There is also a large body of work in the theoretical computer science community studying robust algorithm design for various combinatorial problems such as graph partitioning, independent set etc. A popular framework that is used in such contexts is *semi-random* models (Blum and Spencer, 1995). Semi-random models assume that the input is generated from an ideal distribution and then perturbed by an adversary in a non-worst case manner. The study of such models has led to the design of robust algorithms for many problems such as coloring (Blum and Spencer, 1995), independent set (Feige and Kilian, 2001), graph partitioning (Makarychev et al., 2012) and lately for machine learning problems as well (Moitra et al., 2015; Vijayaraghavan and Awasthi, 2018; Cheng and Ge, 2018; Awasthi and Vijayaraghavan, 2018).

Appendix B. Preliminaries

We discuss a few properties about the operator $p \to q$ norm, robust projections, and $\sin\Theta$ distance between subspaces and projections in this section.

A useful fact of the operator norms is the efficient approximation algorithms.

Lemma 12 (Nesterov (1998); Steinberg (2005)) *For any $q \le 2 \le p$, there exists an efficient randomized algorithm with an input matrix A that approximates $\|A\|_{p\to q}$ within a constant factor $C_{p,q} \le 3$. Moreover for any $q \ge 2$, and for PSD matrices M, there exists polynomial time algorithms that approximates $\|M\|_{q\to q^*}$ within a $1/\gamma_{q^*}^2$ factor where γ_{q^*} is the expected ℓ_{q^*} norm of a standard normal r.v. In particular for $q = \infty$, this gives a $\pi/2$ approximation.*

One crucial property in the rounding algorithm of the convex program (2) is the monotonicity of $q \to q^*$ norm stated below (See Section 5 in Awasthi et al. (2019a) for a proof, and counter examples for other norms).

Lemma 13 *For any $q > 2$, $q \to q^*$ norm is monotone for PSD matrices: for any $A, B \succeq 0$, $\|A + B\|_{q\to q^*} \geq \|A\|_{q\to q^*}$.*

Robust projections. We show basic properties of a projection matrix Π in terms of its $q \to 2$ norm.

Fact 14 *Given any projection matrix Π with $\|\Pi\|_{q\to 2} \leq \kappa$ for $q > 2$, we have the following properties.*

1. *For any δ and vectors u and v with $\|u - v\|_q \leq \delta$, $\|\Pi u - \Pi v\|_2 \leq \kappa\delta$.*
2. *Any vector v in this subspace has $\|v\|_{q^*}/\|v\|_2 \leq \kappa$. Moreover $\|\Pi\|_{q^*} \leq rank(\Pi) \cdot \kappa^2$.*

Proof The first property follows from the definition of $q \to 2$ norm.

For the second property, $\|v\|_q = \|\Pi v\|_q \leq \kappa\|v\|_2$ by definition. Morever, we could choose a orthonormal basis $v_1, \ldots, v_r$ for Π such that $\|\Pi\|_{q^*} = \|\sum_{i=1}^r v_i v_i^\top\|_{q^*} \leq \sum_{i=1}^r \|v_i v_i^\top\|_{q^*} = r\kappa^2$. ■

The constraint (5) in the convex program essentially comes from the 2nd property in the above fact.

$\sin\Theta$ distance of subspaces. Given two subspaces S and S^* of the same dimension, we always measure their distance in terms of the Frobenius norm of the $\sin\Theta(S, S^*)$ matrix, where Θ corresponds to the principal angles between the subspaces. This has a simple expression in terms of the projection matrices Π, Π^* when both have the same rank:

$\sin\Theta(S, S^*) = \Pi^\perp\Pi^*$. Hence $\|\sin\Theta(S, S^*)\|_F^2 = \|\Pi^\perp\Pi^*\|_F^2 = \|\Pi^*\|_F^2 - \langle\Pi, \Pi^*\rangle = \frac{1}{2}\|\Pi - \Pi^*\|_F^2$.

In particular, when we measure the distance between two projection matrices Π and Π^* of rank r, we will also use the following form

$$\|\sin\Theta(\Pi, \Pi^*)\|_F^2 = \|\Pi^\perp\Pi^*\|_F^2 = r - \langle\Pi, \Pi^*\rangle. \tag{13}$$

Appendix C. Solving the convex program (2)

Lemma 15 *For any $q \geq 2$, there exists a constant $c = c(q) \geq 1$ such that the following holds. There is a randomized polynomial time algorithm that given an instance $A \in \mathbb{R}^{n\times m}$ with an optimal solution X^* to the relaxation (2)-(6), with high probability finds a solution $\widehat{X}$ that is arbitrarily close in objective value compared to X^* such that $\|\widehat{X}\|_{q\to q^*} \leq c\kappa^2$.*

Proof We first observe that the feasible set of the program is convex. We now show how to use the Ellipsoid algorithm to approximately it. We will design an approximate hyperplane separation oracle for (6) and (5). The constraint (6) can be rewritten as $\langle yz^\top, X\rangle \leq \kappa^2$ for all $y, z \in \mathbb{R}^n$ such that $\|y\|_q, \|z\|_q \leq 1$. As described in Lemma 12, there exists SDP-based polynomial time algorithms that give constant factor $c = c(q)$ approximations for computing

the $q \to q^*$ matrix operator norm. Such an approximation algorithm immediately gives a $c(q)$-factor approximate separation oracle; when $\|X\|_{q\to q^*} > c\kappa^2$, the solution y', z' output by the algorithm gives a separating hyperplane of the form $\langle y'(z')^\top, X\rangle \le \kappa^2$. Finally, the constraint (5) is also convex and can be efficiently separated using the gradient at the given point X. ■

Appendix D. Computational Upper Bounds

In this section we provide proofs of the supporting clams that were used in establishing our main theorem (Theorem 6). We start with proving our main result — Theorem 6. Then we prove Lemma 7 in Appendix D.2, Claim 9 in Appendix D.3, and Theorem 10 in Appendix D.4.

D.1. Proof of Theorem 6

We finish the proof of our main theorem (Theorem 6) using the supporting claims.

Proof of Theorem 6. Recall that we define $E = \frac{1}{m}\widetilde{A}\widetilde{A}^\top - \mathbb{E}[AA^\top]$. Let X be the solution to the SDP in (2). From the optimality of X we have that

$$\langle X, \Sigma^* + E\rangle \ge \langle \Pi^*, \Sigma^* + E\rangle.$$

We bound $\langle X, E\rangle$ and $\langle \Pi^*, E\rangle$ by $\Delta := O\Big(\sqrt{r}\kappa\delta\sqrt{\lambda_1} + \kappa^2\delta^2 + \frac{r\kappa^2\cdot\lambda_1\sqrt{\log n}\cdot n^{2/q}}{\sqrt{m}}\Big)$ using Lemma 7. Hence we get that $\langle X, \Sigma^*\rangle \ge \langle \Pi^*, \Sigma^*\rangle - 2\Delta$. Then we apply Claim 9 to obtain

$$\langle X, \Pi^*\rangle \ge r - 2\Delta/(\lambda_r - \lambda_{r+1}) = r - 2\Delta/\theta, \tag{14}$$

where $\theta := \lambda_r - \lambda_{r+1}$. Let $X = \sum_{i=1}^n \lambda_i(X) u_i u_i^\top$ be the eigendecomposition of X with $\lambda_1(X) \ge \lambda_2(X) \ge \cdots \ge \lambda_n(X)$ and let $\widetilde{\Pi} = \sum_{i=1}^r u_i u_i^\top$. Since Π^* is a projection matrix, equation (14) implies that

$$\langle \Pi^*, X\rangle = \sum_{i=1}^n \lambda_i(X)\cdot\|\Pi^* u_i\|_2^2 \ge r - 2\Delta/\theta \quad\text{and}\quad \langle \Pi^*, \widetilde{\Pi}\rangle = \sum_{i=1}^r \|\Pi^* u_i\|_2^2.$$

Similarly since $\langle \widetilde{\Pi}, X\rangle \ge \langle \Pi^*, X\rangle \ge r - 2\Delta/\theta$, we have that

$$\sum_{i=1}^r \lambda_i(X) = \langle \widetilde{\Pi}, X\rangle \ge r - 2\Delta/\theta.$$

At the same time from the constraints of the SDP $\sum_{i=1}^n \lambda_i(X) = \mathrm{tr}(X) = r$. Hence

$$\sum_{i=r+1}^n \lambda_i(X)\cdot\|\Pi^* u_i\|_2^2 \le \sum_{i=r+1}^n \lambda_i(X) \le 2\Delta/\theta.$$

Using the above we get

$$\langle \Pi^*, \widetilde{\Pi} \rangle = \sum_{i=1}^{r} \|\Pi^* u_i\|_2^2 \geq \sum_{i=1}^{r} \lambda_i(X) \|\Pi^* u_i\|_2^2$$
$$= \sum_{i} \lambda_i(X) \|\Pi^* u_i\|_2^2 - \sum_{i=r+1}^{n} \lambda_i(X) \|\Pi^* u_i\|_2^2 \geq r - \frac{4\Delta}{\theta}.$$

This establishes $\|\widetilde{\Pi}^{\perp}\Pi^*\|_F^2 = \frac{1}{2}\|\widetilde{\Pi} - \Pi^*\|_F^2$ is at most $4\Delta/\theta$.

Finally we note $\lambda_r(X) \geq 1 - 2\Delta/\theta$ since $\sum_{i=1}^{r}(1 - \lambda_i(X)) \leq 2\Delta/\theta$, which implies $\|\widetilde{\Pi}\|_{q\to 2} \leq \|X\|_{q\to 2}/(1 - 2\Delta/\theta) = O(\kappa)$. The correctness of $\widetilde{\Sigma}_{\text{TOP}}$ then follows from Theorem 10. Note that $\lambda_1^2 r^2/m$ and $\kappa^4\delta^4$ are always less than $\lambda_1^2\varepsilon$ and $\lambda_1 \cdot \kappa^2\delta^2$ separately given our parameters. ■

D.2. Bounding Error over SDP Solutions

Here we provide the proof of Lemma 7. We first state and prove a few useful claims.

Claim 16 *For any X in $\mathcal{P}_{c(q)}$ and let $\widetilde{A}$ be an δ-perturbation of A. Then we always have that*

$$\|X^{1/2}(\widetilde{A} - A)\|_F \leq \sqrt{c(q)m} \cdot \kappa\delta$$

and

$$\langle (\widetilde{A} - A)(\widetilde{A} - A)^\top, X \rangle \leq c(q)m \cdot \kappa^2\delta^2.$$

Proof Define $B = (\widetilde{A} - A)$. The norm bound $\|X\|_{q\to q^*} \leq c(q)\kappa^2$ along with the fact that for any matrix M, $\|M^\top M\|_{q\to q^*} = \|M\|_{q\to 2}^2$, implies that $\|X^{\frac{1}{2}}\|_{q\to 2} \leq \sqrt{c(q)} \cdot \kappa$. Denoting B_i to be the ith column of B, we get that $\|B_i\|_q \leq \delta$ and that

$$\|X^{\frac{1}{2}}B\|_F^2 = \sum_{i=1}^{m} \|X^{\frac{1}{2}}B_i\|^2 \leq \sum_{i=1}^{m} c(q) \cdot \kappa^2\delta^2 = m \cdot c(q)\kappa^2\delta^2.$$

Next, note that $\langle (\widetilde{A} - A)(\widetilde{A} - A)^\top, X \rangle = \langle BB^\top, X \rangle = \|X^{1/2}B\|_F^2 \leq c(q) \cdot m\kappa^2\delta^2$. ■

We will also use the following standard fact extensively.

Fact 17 *For any two PSD matrices A and B, $\lambda_{\min}(A) \cdot tr(B) \leq \langle A, B \rangle \leq \lambda_{\max}(A) \cdot tr(B)$.*

Proof We rewrite $\langle A, B \rangle = \|A^{1/2}B^{1/2}\|_F^2$, which is sandwiched by $\lambda_{\min}(A^{1/2})^2 \cdot \|B^{1/2}\|_F^2 = \lambda_{\min}(A) \cdot \text{tr}(B)$ and $\lambda_{\max}(A^{1/2})^2 \cdot \|B^{1/2}\|_F^2 = \lambda_{\max}(A) \cdot \text{tr}(B)$. ■

We will use the following standard concentration bound on the moments of the covariance matrix of Gaussian random variables (see Lemma 8.12 in Awasthi et al. (2019a) for a proof).

Lemma 18 *Let $A_1, A_2, \ldots, A_m \in \mathbb{R}^n$ be generated i.i.d. from $\mathcal{N}(0, \Sigma^*)$. Let A be the $n \times m$ matrix with the columns being the points A_i. Then with probability at least $1 - \frac{1}{\text{poly}(n)}$*

$$\left\| \frac{1}{m}AA^\top - \mathbb{E}[AA^\top] \right\|_\infty \leq c\frac{\|\Sigma\|\sqrt{\log n}}{\sqrt{m}} \text{ and } \|\frac{1}{m}AA^\top - \Sigma^*\|_q \leq c\frac{\|\Sigma\| \cdot n^{2/q}\sqrt{\log n}}{\sqrt{m}}. \quad (15)$$

We now proceed to the proof of the main lemma that upper bounds $|\langle E, X\rangle|$.

Proof of Lemma 7. Using that fact that $\mathbb{E}[A] = 0$ and $B = \widetilde{A} - A$ we rewrite

$$\begin{aligned} E &= \frac{1}{m}(A+B)(A+B)^\top - \mathbb{E}[AA^\top] \\ &= \frac{1}{m}\Big(BB^\top + B(A-\mathbb{E}[A])^\top + (A-\mathbb{E}[A])B^\top + AA^\top\Big) - \mathbb{E}[AA^\top]. \end{aligned}$$

Hence we get that

$$|\langle E, X\rangle| \le \underbrace{\frac{1}{m}\langle BB^T, X\rangle}_{T_1} + \underbrace{\frac{2}{m}\Big|\Big\langle (A-\mathbb{E}[A])B^T, X\Big\rangle\Big|}_{T_2} + \underbrace{\Big|\Big\langle \frac{1}{m}AA^\top - \mathbb{E}[AA^\top], X\Big\rangle\Big|}_{T_3}. \tag{16}$$

Next we separately bound each of the terms above. Using Claim 16,

$$T_1 = \frac{1}{m}\langle BB^T, X\rangle = \frac{1}{m}\|X^{1/2}B\|_F^2 \le c(q)\kappa^2\delta^2.$$

Using the concentration bound from Lemma 18 on $\|\frac{1}{m}AA^\top - \mathbb{E}[AA^\top]\|_q$, t_3 can be bounded as

$$\begin{aligned} T_3 = \Big\langle \frac{1}{m}AA^\top - \mathbb{E}[AA^\top], X\Big\rangle &\le \|\frac{1}{m}AA^\top - \mathbb{E}[AA^\top]\|_q \cdot \|X\|_{q^*} \\ &= O\Big(\frac{\lambda_{\max}(\Sigma^*)\cdot n^{2/q}\sqrt{\log n}\cdot r\kappa^2}{\sqrt{m}}\Big). \end{aligned}$$

The second term T_2 in (16) is the crucial term to upper bound, and contributes the dominant term of $\sqrt{\lambda_1 r}\kappa\delta$ in the guarantees of Theorem 6. A naive upper bound on T_2 can be obtained by $|\langle M_1, M_2\rangle| \le \|M_1\|_{q^*}\|M_2\|_q$ as we did for T_3, but this leads to sub-optimal bounds that are off by factors involving r. The following technical claim which is a restatement of Lemma 8 from Section 3 crucially uses the constraint on $\|X\|_{q\to q^*}$. Its proof is deferred to Appendix E.

Lemma 19 *Let $A_1, A_2, \ldots, A_m \in \mathbb{R}^n$ be generated i.i.d. from $\mathcal{N}(\mu, \Sigma^*)$. Let A be the $n \times m$ matrix with the columns being the points A_i. Let X be a solution to the SDP in program (2) and let B be any matrix, potentially chosen based on A, with $\|B_j\|_q \le \delta\ \forall j \in [m]$. Then with probability at least $1 - \frac{1}{\mathrm{poly}(n)}$ we have that*

$$\frac{1}{m}\Big|\langle (A-\mathbb{E}[A])B^T, X\rangle\Big| \le O(\sqrt{r\|\Sigma^*\|}\kappa\delta) + O(\kappa^2\delta^2) + O\Big(\frac{r\kappa^2\|\Sigma^*\|\sqrt{\log n}\cdot n^{2/q}}{\sqrt{m}}\Big). \tag{17}$$

Combining the above bounds and using the fact that $\|X\|_{q\to q^*} \le c(q)\kappa^2$ we get that

$$\Delta \le c(q)\cdot O\Big(\kappa^2\delta^2 + \sqrt{r\lambda_{\max}(\Sigma^*)}\kappa\delta + \frac{r\kappa^2\lambda_{\max}(\Sigma^*)\sqrt{\log n}\cdot n^{2/q}}{\sqrt{m}}\Big).$$

■

D.3. Bounding Correlation with the Subspace [Proof of Claim 9]

In this section we provide the proof of Claim 9. For convenience, let ε denote the gap $\varepsilon := r - \langle X, \Pi^* \rangle$. Hence the goal is to show $\varepsilon \leq (\langle \Pi^*, \Sigma^* \rangle - \langle X, \Sigma^* \rangle)/(\lambda_r - \lambda_{r-1})$. To show this we will obtain an upper bound $\langle X, \Sigma^* \rangle$ in terms of $\varepsilon, (\lambda_r - \lambda_{r+1})$ and $\langle \Pi^*, \Sigma^* \rangle$.

Given the eigen-decomposition $\Sigma^* = \sum_{i=1}^n \lambda_i v_i v_i^\top$ with $\lambda_1 \geq \cdots \geq \lambda_n$, we define $\Sigma_{\text{TOP}} = \sum_{i=1}^r \lambda_i v_i v_i^\top$ and $\Sigma_{\text{BOT}} = \sum_{i=r+1}^n \lambda_i v_i v_i^\top$. Note $\langle \Pi^*, \Sigma^* \rangle = \langle \Pi^*, \Sigma_{\text{TOP}} \rangle = \text{tr}(\Sigma_{\text{TOP}})$. We will upper bound $\langle X, \Sigma^* \rangle$ as $\langle X, \Sigma_{\text{TOP}} + \Sigma_{\text{BOT}} \rangle$ given $\langle X, \Pi^* \rangle = r - \varepsilon$. Let $V = [v_1, \ldots, v_n]$ denote the matrix with columns being the eigenvectors of Σ^*. For convenience, we rewrite

$$\begin{aligned}
\Sigma^* &= \sum_{j=1}^n \lambda_j v_j v_j^\top = V \text{diag}[\lambda_1, \ldots, \lambda_n] V^\top, \\
\Sigma_{\text{TOP}} &= V \text{diag}[\lambda_1, \ldots, \lambda_r, 0, \ldots,] V^\top, \\
\Pi^* &= V \text{diag}[1, \ldots, 1, 0, \ldots,] V^\top.
\end{aligned}$$

The above implies that

$$r - \varepsilon = \langle X, \Pi^* \rangle = \langle X, V \text{diag}[1, \ldots, 1, 0, \ldots,] V^\top \rangle = \langle X', \text{diag}[1, \ldots, 1, 0, \ldots,] \rangle$$

where $X' = VXV^\top$. Similarly, $\langle X, \Sigma^* \rangle = \langle X', \text{diag}[\lambda_1, \ldots, \lambda_r, 0, \ldots] \rangle$. Since X' also satisfies $0 \preceq X' \preceq I$, we have that

$$\langle X, \Sigma_{\text{TOP}} \rangle = \langle X', \text{diag}[\lambda_1, \ldots, \lambda_r, 0, \ldots] \rangle \leq \text{tr}(\Sigma_{\text{TOP}}) - \varepsilon \cdot \lambda_r$$

as $\langle X', \text{diag}[1, \ldots, 1, 0, \ldots, 0] \rangle = r - \varepsilon$. Similarly, we have $\langle X', \text{diag}[0, \ldots, 0, 1, \ldots, 1] \rangle = \varepsilon$, so

$$\langle X, \Sigma_{\text{BOT}} \rangle = \langle X', \text{diag}[0, \ldots, 0, \lambda_{r+1}, \ldots, \lambda_n] \rangle \leq \varepsilon \cdot \lambda_{r+1}.$$

The above two bounds show that

$$\langle X, \Sigma^* \rangle \leq \text{tr}(\Sigma_{\text{TOP}}) - \varepsilon \lambda_r + \varepsilon \lambda_{r+1}.$$

Hence we get that

$$\begin{aligned}
\langle X, \Sigma^* \rangle &\leq \langle \Pi^*, \Sigma^* \rangle - \varepsilon(\lambda_r - \lambda_{r+1}), \\
\varepsilon &\leq \frac{\langle \Sigma, \Pi^* \rangle - \langle \Sigma^*, X \rangle}{\lambda_r - \lambda_{r+1}}.
\end{aligned}$$

D.4. Covariance Matrix Recovery

We end the section by showing how to recover a good approximation to the top-r subspace of Σ^* given a good approximation to Π^*. As stated before this is formalized in Theorem 10 which we prove below. We first state the following standard fact to bound the Frobenius error of the covariance estimation (see Theorem 4.7.1 in Vershynin (2018) for a proof).

Fact 20 *Let Σ^* be a covariance matrix of rank r and largest eigenvalue λ_1. For any m, and vectors $A_1, \ldots, A_m \sim N(0, \Sigma^*)$, it holds with probability at least $1 - 10^{-3}$, that $\|\widetilde{\Sigma} - \Sigma^*\| = \lambda_1 \cdot O(\sqrt{r/m} + r/m)$ for $\widetilde{\Sigma} = \frac{1}{m} \sum_i A_i A_i^\top$. Moreover, if $m = O(\lambda_1^2 r^2/\beta)$ then with prob. at least $1 - 10^{-3}$, $\|\widetilde{\Sigma} - \Sigma\|_F^2 \leq \beta$ for $\beta < r$.*

We also use the following lemma showing how to recover a good approximation to the top-r subspace of Σ^* in the absence of noise.

Lemma 21 *For any covariance matrix $\Sigma^* = \sum_{i=1}^n \lambda_i v_i v_i^\top$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, let $\Sigma_{\text{TOP}} = \sum_{i=1}^r \lambda_i v_i v_i^\top$ and Π^* be the projection matrix on to the top r eigenspace of Σ^*. Given any rank r projection matrix Π with $\|\Pi^* - \Pi\|_F^2 \leq \varepsilon$, and $m = \Omega(\lambda_1^2 \cdot r^2)$, we have that with probability at least $1 - 10^{-3}$, $\|\widetilde{\Sigma} - \Sigma_{\text{TOP}}\|_F^2 = O(\lambda_1^2 \cdot \varepsilon + \frac{\lambda_1^2 r^2}{m})$ for $\widetilde{\Sigma} = \Pi(\frac{1}{m}\sum_{i=1}^m A_i A_i^\top)\Pi$ and $A_1, \ldots, A_m$ are generated i.i.d. from $N(0, \Sigma^*)$.*

The above lemma is an extension of Lemma 8.2 in (Awasthi et al., 2019a). For completeness, we provide the proof in Appendix E. Next we establish Theorem 10 showing covariance recovery in the presence of adversarial perturbations.

Proof of Theorem 10. For the estimate $\Pi(\frac{1}{m}\sum_{i=1}^m \widetilde{A}_i \widetilde{A}_i^\top)\Pi$ output by the algorithm we have that

$$\begin{aligned}
&\|\Pi(\frac{1}{m}\sum_{i=1}^m \widetilde{A}_i \widetilde{A}_i^\top)\Pi - \Pi^*\Sigma^*\Pi^*\|_F^2 \\
\leq& 2\|\Pi(\frac{1}{m}\sum_{i=1}^m \widetilde{A}_i \widetilde{A}_i^\top)\Pi - \Pi(\frac{1}{m}\sum_{i=1}^m A_i A_i^\top)\Pi\|_F^2 + 2\|\Pi(\frac{1}{m}\sum_{i=1}^m A_i A_i^\top)\Pi - \Pi^*\Sigma^*\Pi^*\|_F^2 \\
\leq& 2\|\Pi(\frac{1}{m}\sum_{i=1}^m \widetilde{A}_i \widetilde{A}_i^\top)\Pi - \Pi(\frac{1}{m}\sum_{i=1}^m A_i A_i^\top)\Pi\|_F^2 + O(\lambda_1^2\varepsilon + \frac{\lambda_1^2 r^2}{m}),
\end{aligned}$$

where we use Lemma 21 in the last step. Let $\widetilde{A}_i = A_i + B_i$ such that B_i is the perturbation of the ith data point. We can rewrite the first term above as

$$\begin{aligned}
&\|\Pi(\frac{1}{m}\sum_{i=1}^m \widetilde{A}_i \widetilde{A}_i^\top)\Pi - \Pi(\frac{1}{m}\sum_{i=1}^m A_i A_i^\top)\Pi\|_F \\
=&\|\Pi(\frac{1}{m}\sum_{i=1}^m (A_i + B_i)(A_i + B_i)^\top)\Pi - \Pi(\frac{1}{m}\sum_{i=1}^m A_i A_i^\top)\Pi\|_F \\
\leq&\|\Pi(\frac{1}{m}\sum_{i=1}^m B_i B_i^\top)\Pi\|_F + \|\Pi(\frac{1}{m}\sum_{i=1}^m A_i B_i^\top)\Pi\|_F + \|\Pi(\frac{1}{m}\sum_{i=1}^m B_i A_i^\top)\Pi\|_F.
\end{aligned}$$

Now we bound each Frobenius norm separately. For the first term we have

$$\|\Pi(\frac{1}{m}\sum_{i=1}^m B_i B_i^\top)\Pi\|_F \leq \frac{1}{m}\sum_{i=1}^m \|\Pi B_i B_i^\top \Pi\|_F = O(\kappa^2\delta^2)$$

where we have used the fact that ΠB_i is a vector of norm at most $O(\kappa\delta)$. We bound the second term $\|\Pi(\frac{1}{m}\sum_{i=1}^m A_i B_i^\top)\Pi\|_F$ (and similary for the third one), by

$$\frac{1}{m}\|\Pi A\| \cdot \|B^\top \Pi\|_F \leq \frac{1}{m} \cdot \sqrt{\lambda_1 m} \cdot O(1 + \sqrt{r/m}) \cdot \sqrt{m}\kappa\delta = \sqrt{\lambda_1} \cdot O(\sqrt{r/m} + 1)\kappa\delta$$

where we bound $\|B^\top \Pi\|_F^2 \le \sqrt{m}\kappa\delta$ from the above bound on $\|\Pi(\frac{1}{m}\sum_{i=1}^m B_i B_i^\top)\Pi\|_F$ and $\|\Pi A\| \le \sqrt{\lambda_1 m}\cdot(1+\sqrt{r/m})$ as follows: $\mathrm{rank}(\Pi A) = r$ and Fact 20 implies that with probability at least $1-10^{-3}$, $\|\Pi\frac{1}{m}AA^\top\Pi - \mathbb{E}[(\Pi A)\cdot(\Pi A)^\top]\| \le O(\lambda_1\cdot\sqrt{r/m})$. Since $\|\mathbb{E}[(\Pi A)\cdot(\Pi A)^\top]\| \le \|\mathbb{E}[AA^\top]\| \le \lambda_1$, $\|\Pi\frac{1}{m}AA^\top\Pi\| \le \lambda_1 + \lambda_1\cdot O(\sqrt{r/m})$.

Combining the above bounds we get that $\|\Pi(\frac{1}{m}\sum_{i=1}^m \widetilde{A}_i\widetilde{A}_i^\top)\Pi - \Pi^*\Sigma^*\Pi^*\|_F^2$ can be bounded by

$$O(\lambda_1^2\varepsilon + \lambda_1^2\cdot r^2/m) + O(\kappa^4\delta^4) + O(\lambda_1\cdot(1+r/m)\cdot\kappa^2\delta^2).$$

■

Appendix E. Additional Proofs from Section 3

Proof of Lemma 19. We use the fact that for matrices M_1, M_2, Q_1, and Q_2, it holds that

$$\langle M_1M_2, Q_1Q_2\rangle \le \|M_1^\top Q_1\|_F\|M_2Q_2^\top\|_F$$

to rewrite

$$\begin{aligned}\frac{1}{m}\langle(A-\mathbb{E}[A])B^T, X\rangle &= \frac{1}{m}\langle(A-\mathbb{E}[A])B^T, X^{\frac{1}{2}}X^{\frac{1}{2}}\rangle\\ &\le \frac{1}{m}\|(A-\mathbb{E}[A])^\top X^{\frac{1}{2}}\|_F\|X^{\frac{1}{2}}B\|_F\end{aligned}$$

By Claim 16, $\|X^{\frac{1}{2}}B\|_F \le \sqrt{m}\kappa\delta$. Note that $\|(A-\mathbb{E}[A])^\top X^{\frac{1}{2}}\|_F^2 = \langle AA^\top, X\rangle$ given $\mathbb{E}[A]=0$. Then we split it into

$$\langle AA^\top, X\rangle = \langle AA^\top - m\cdot E[AA^\top], X\rangle + \langle m\cdot\mathbb{E}[AA^\top], X\rangle = O\left(r\kappa^2\cdot\|\Sigma^*\|\sqrt{\log n}\cdot n^{2/q}\cdot\sqrt{m} + \|\Sigma^*\|\cdot rm\right),$$

where the first bound comes from the above proof of Lemma 7 for the last term in (16) and the second bound comes from Fact 17.

We finish the proof by combining the above bounds:

$$\begin{aligned}\frac{2}{m}\langle(A-\mathbb{E}[A])B^T, X\rangle &\le \frac{1}{m}\cdot\sqrt{m}\kappa\delta\cdot O\left(\|\Sigma^*\|\cdot rm + r\kappa^2\cdot\|\Sigma^*\|\sqrt{\log n}\cdot n^{2/q}\sqrt{m}\right)^{1/2}\\ &= O(\sqrt{r\|\Sigma^*\|}\kappa\delta) + \kappa\delta\cdot O\left(\frac{\|\Sigma^*\|r\kappa^2\sqrt{\log n}\cdot n^{2/q}}{\sqrt{m}}\right)^{1/2}\\ &\le O(\sqrt{r\|\Sigma^*\|}\kappa\delta) + O(\kappa^2\delta^2) + O\left(\frac{\|\Sigma^*\|r\kappa^2\sqrt{\log n}\cdot n^{2/q}}{\sqrt{m}}\right).\end{aligned}$$

■

E.1. Proof of Lemma 21

We will use the following fact to apply triangle inequality.

Fact 22 *Given a rank r covariance matrix Σ^* with all eigenvalues upper bounded by $\lambda_{\max}$ and projection matrix Π^*, for any rank r projection Π with $\|\Pi^* - \Pi\|_F^2 \le \varepsilon$ and any $\widetilde{\Sigma}$ (not necessarily rank r), we have*

$$\|\Sigma^* - \Pi\widetilde{\Sigma}\Pi\|_F^2 \le 8\lambda_{\max}^2 \cdot \varepsilon + 2\|\Pi\Sigma^*\Pi \ - \Pi\widetilde{\Sigma}\Pi\|_F^2.$$

Proof At first, we have $\|\Sigma^* - \Pi\widetilde{\Sigma}\Pi\|_F^2 \le 2\|\Sigma^* - \Pi\Sigma^*\Pi\|_F^2 + 2\|\Pi\Sigma^*\Pi \ - \Pi\widetilde{\Sigma}\Pi\|_F^2$.

Since Π^* is the projection matrix of Σ^*, we have

$$\|\Sigma^* - \Pi\Sigma^*\Pi\|_F^2 = \|\Pi^*\Sigma^*\Pi^* - \Pi\Sigma^*\Pi\|_F^2 \le 2(\|\Pi^*\Sigma^*\Pi^* - \Pi\Sigma^*\Pi^*\|_F^2 + \|\Pi\Sigma^*\Pi^* - \Pi\Sigma^*\Pi\|_F^2).$$

Since $\|AB\|_F^2 \le \|A\|_{op}^2 \cdot \|B\|_F^2$, we further simplify it as

$$2(\|\Pi^* - \Pi\|_F^2 \cdot \|\Sigma^*\|_{op}^2 + \|\Sigma^*\|_{op}^2 \cdot \|\Pi^* - \Pi\|_F^2) \le 4\lambda_{\max}^2 \cdot \varepsilon.$$

■

We finish the proof of Lemma 21.

Proof of Lemma 21. Given $A_i \sim N(0, \Sigma^*)$, we know ΠA_i is a random vector generated by $N\big(0, \Pi\Sigma^*\Pi\big)$. So we apply Fact 20 to bound $\|\Pi\Sigma\Pi - \Pi(\frac{1}{m}\sum_{i=1}^m A_i A_i^\top)\Pi\|_F^2 \le \delta$. Then we consider Σ_{TOP}:

$$\|\Pi\Sigma\Pi - \Pi\Sigma_{\text{TOP}}\Pi\ \|_F = \|\Pi\Sigma_{\text{BOT}}\Pi\|_F \le \|(\Pi - \Pi^*)\Sigma_{\text{BOT}}\Pi\|_F + \|\Pi^*\Sigma_{\text{BOT}}\Pi\|_F.$$

Since Π^* is the projection matrix of Σ_{TOP}, $\Pi^*\Sigma_{\text{BOT}} = 0$ such that the second term becomes 0. For the first term $\|(\Pi - \Pi^*)\Sigma_{\text{BOT}}\Pi\|_F$, we upper bound it by

$$\|(\Pi - \Pi^*)\Sigma_{\text{BOT}}\Pi\|_F \le \|\Pi - \Pi^*\|_F \cdot \|\Sigma_{\text{BOT}}\|_{op} \cdot \|\Pi\|_{op} \le \lambda_1 \cdot \sqrt{\varepsilon}.$$

From the above discussion, we have

$$\|\Pi\Sigma_{\text{TOP}}\Pi - \Pi(\frac{1}{m}\sum_{i=1}^m A_i A_i^\top)\Pi\|_F^2 \le 2\|\Pi\Sigma\Pi - \Pi(\frac{1}{m}\sum_{i=1}^m A_i A_i^\top)\Pi\|_F^2 + 2\|(\Pi - \Pi^*)\Sigma_{\text{BOT}}\Pi\|_F^2 = O(\delta + \lambda_1^2\varepsilon).$$

The final bound follows from Fact 22 with $\Sigma^* = \Sigma_{\text{TOP}}$ there. ■

Appendix F. Statistical Lower Bound on the Error and Instance-Optimality

F.1. Auxiliary claims and Proofs.

Proof [Proof of Lemma 11] By construction $u_1, \ldots, u_r$ have disjoint supports, and for each $\ell \in [r]$, $\Pi^*\Pi_\ell^\perp = 0$; hence $\Pi^* u_\ell = 0$. We now show (11). Note that $\Pi_\ell^\perp g_\ell$ is distributed

according to the Gaussian $\mathcal{N}(0, \Pi_\ell^\perp)$. Hence $\mathbb{E}[\|\Pi_\ell^\perp g_\ell\|_2^2] = \mathrm{tr}(\Pi_\ell^\perp) = d_\ell$. For a fixed $\ell \in [r]$, using concentration bounds for χ^2 distributions we have for any $t > 0$

$$\mathbb{P}\left[\left|\|u_\ell\|_2^2 - 1\right| > 2\sqrt{\frac{t}{d_\ell}} + 2\frac{t}{d_\ell}\right] = \mathbb{P}\left[\left|\|\Pi_\ell^\perp g_\ell\|_2^2 - d_\ell\right| > 2\sqrt{d_\ell t} + 2t\right] \leq \exp(-t).$$

Substituting $t = \log(r/\eta)$, along with $d_\ell \geq k' - r \geq k'/2$ and a union bound over all $\ell \in [r]$ establishes (11). Then the last property of $\|u_\ell\|_1 \leq 2\sqrt{k'}$ follows from the Cauchy-Schwartz inequality with the fact that the support size of u_ℓ is at most k'.

Now we upper bound $\|u_\ell\|_\infty$. For each coordinate i and ℓ,

$$u_\ell(i) = \frac{1}{\sqrt{d_\ell}}\langle \Pi_\ell^\perp(i,:), g_\ell\rangle \text{ where } \Pi_\ell^\perp(i,:) \text{ represents the } i\text{th row of } \Pi_\ell^\perp.$$

This is a Gaussian random variable. Hence for a fixed $\ell \in [r]$, with probability at least $1 - \frac{\eta}{r}$,

$$\|u_\ell\|_\infty = \frac{1}{\sqrt{d_\ell}} \max_{i\in[n]} |\langle \Pi_\ell^\perp(i,:), g_\ell\rangle| \leq 2\sqrt{\log(rk'/\eta)} \cdot \frac{\max_{i\in[n]} \|\Pi_\ell^\perp(i,:)\|}{\sqrt{d_\ell}} \leq 2\sqrt{\log(rk'/\eta)} \cdot \frac{1}{\sqrt{k'/2}},$$

since $\Pi_\ell^\perp$ is an orthogonal projection matrix. After a union bound over $\ell \in [r]$, (12) follows. ■

Proof [Proof of Lemma 24] The proof just uses norm duality and relations between different norms.

$$\begin{aligned}
\Big\| \sum_\ell u_\ell v_\ell^\top \Big\|_{q\to q^*} &= \max_{\substack{x,y:\|x\|_q\leq 1\\ \|y\|_q \leq 1}} \sum_{\ell=1}^r \langle x, u_\ell\rangle\langle v_\ell, y\rangle \leq \sum_\ell \|u_\ell\|_{q^*} |\langle v_\ell, y\rangle| \\
&\leq \max_\ell \|u_\ell\|_{q^*} \cdot \sum_\ell |\langle v_\ell, y\rangle| = \max_{\ell\in[r]} \|u_\ell\|_{q^*} \cdot \|V^\top\|_{q\to q^*} \\
&= \max_{\ell\in[r]} \|u_\ell\|_{q^*} \cdot \|V\|_{q\to q^*} \leq r^{1/2-1/q} \|V\|_{q\to 2} \cdot \max_{\ell\in[r]} \|u_\ell\|_{q^*}.
\end{aligned}$$

The last inequality follows since $\|Vy\|_{q^*} \leq r^{1/2-1/q}\|Vy\|_2$ for any $y \in \mathbb{R}^n$ since V has r columns (see Section 2).

For the second statement, we have $\|UU^\top\|_{q\to q^*} = \|U^\top\|_{q\to 2}^2 = \|U\|_{2\to q^*}^2$ using the variational characterization of operator norms and norm duality (see Section 2). We now upper bound $\|U\|_{2\to q^*}$. Consider any $y \in \mathbb{R}^r$ with $\|y\|_2 = 1$. Then because of the disjoint supports of the columns of U

$$\begin{aligned}
\|Uy\|_{q^*}^{q^*} &= \Big(\sum_{\ell=1}^r |y_\ell|^{q^*} \|u_\ell\|_{q^*}^{q^*}\Big) \leq \max_{\ell\in[r]} \|u_\ell\|_{q^*}^{q^*} \cdot \|y\|_{q^*}^{q^*} \\
\|Uy\|_{q^*} &\leq \|y\|_{q^*} \cdot \max_{\ell\in[r]} \|u_\ell\|_{q^*} \leq r^{1/q^*-1/2}\|y\|_2 \cdot \max_{\ell\in[r]} \|u_\ell\|_{q^*} \leq r^{1/2-1/q} \max_{\ell\in[r]} \|u_\ell\|_{q^*}.
\end{aligned}$$

Hence the lemma holds. ■

The following simple lemma will be in upper bounding the magnitude of the perturbation for each sample point.

Lemma 23 *Given any $u_1, \dots, u_r \in \mathbb{R}^n$ with disjoint support, and any $\alpha_1, \dots, \alpha_r \in \mathbb{R}$, we have*

$$\Big\| \sum_{\ell=1}^{r} \alpha_\ell u_\ell \Big\|_q \le r^{1/q} \max_{\ell \in [r]} |\alpha_\ell| \|u_\ell\|_q.$$

Proof Since $u_1, \dots, u_r$ have disjoint support,

$$\Big\| \sum_{\ell=1}^{r} \alpha_\ell u_\ell \Big\|_q^q = \sum_{\ell=1}^{r} |\alpha_\ell|^q \|u_\ell\|_q^q \le r \Big(\max_{\ell \in [r]} |\alpha_\ell| \|u_\ell\|_q \Big)^q, \text{ as required.}$$

■

The following lemma is also useful to upper bound the $\infty \to 2$ operator norm of the alternate subspace projector Π'.

Lemma 24 *Given any vectors $u_1, \dots, u_r$ and vectors $v_1, \dots, v_r$ that form the columns of $U, V \in \mathbb{R}^{n \times r}$ separately, then for any $q \ge 1$*

$$\Big\| UV^\top \Big\|_{q \to q^*} \le \|V\|_{q \to q^*} \Big(\max_{\ell \in [r]} \|u_\ell\|_{q^*} \Big) \le r^{1/2 - 1/q} \|V\|_{q \to 2} \Big(\max_{\ell \in [r]} \|u_\ell\|_{q^*} \Big). \tag{18}$$

Moreover if $u_1, \dots, u_r$ have disjoint support then

$$\|UU^\top\|_{q \to q^*} = \|U\|_{2 \to q^*}^2 \le r^{1 - 2/q} \Big(\max_{\ell \in [r]} \|u_\ell\|_{q^*}^2 \Big). \tag{19}$$

F.2. Warmup: Min-max lower bound

We now give a min-max optimal lower bound. While Theorem 4 is much more general, we include this argument since it is simpler and helps build intuition, and also gives the correct dependence on the eigengap. The lower bound will apply for $\Sigma^* = \theta \Pi^* + I$; hence $\Sigma_{\text{TOP}} = (1+\theta)\Pi^*$ and $\Sigma_{\text{BOT}} = (I - \Pi^*) = (\Pi^*)^\perp$.

Theorem 25 *Suppose we are given parameters n, m, $\theta > 0$, $r \in \mathbb{N}$, κ, and $\delta > 0$ satisfying $\sqrt{r\lambda_1}(\frac{\kappa}{n}) < \delta \le \sqrt{r\theta}/\kappa$. There exist orthogonal projection matrices Π^*, Π' both of rank r with $\|\Pi^*\|_{\infty \to 2} \le \kappa$ and $\|\Pi'\|_{\infty \to 2} \le \kappa$ such that:*

- *We have the coupling data matrices A and $A' \in \mathbb{R}^{n \times m}$ with their columns generated i.i.d. from $\mathcal{N}(0, I + \theta\Pi^*)$ and $\mathcal{N}(0, I + \theta\Pi')$ respectively, such that $\|A - A'\| \le \delta$ with high probability.*
- $\|\Pi' - \Pi^*\|_F^2 = \Omega\big(\frac{1}{\sqrt{\theta}} \cdot \sqrt{r}\delta\kappa / \log nm\big)$.

We now prove the above theorem. We first show the constructions of Π' and A'. Choose k to be a power of 2 in $[\kappa^2/3, 2\kappa^2/3]$. Let $S := \{1, 2, \cdots, k\} \subset [n]$ and $v_1, v_2, \cdots, v_r$ be any r orthonormal vectors of the form $v_\ell(i) = \pm 1/\sqrt{k}$ if $i \in S$ and 0 otherwise. For example, there are k Fourier characters v_ℓ in $\{0,1\}^{\log k}$ that are orthogonal to each other with $\|v_\ell\|_\infty \le 1/\sqrt{k}$: For each $i \in [k]$, let $\overrightarrow{i} \in \{0,1\}^{\log k}$ be its binary form. Then each Fourier character is $v_\ell(i) = (-1)^{\langle \overrightarrow{\ell}, \overrightarrow{i} \rangle}/\sqrt{k}$.

Let $k' \in [\frac{1}{4}, \frac{1}{2}] \cdot \sqrt{\theta}\kappa/(\delta\sqrt{r})$ be a power of 2 to denote the support size of the perturbation vector. Let $u_1, \ldots, u_r$ be unit vectors supported on a disjoint set of k' coordinates each from $[n] \setminus S$ with $\|u_\ell\|_\infty = 1/\sqrt{k'}$ for each $\ell \in [r]$ using the same construction of $v_1, \ldots, v_r$. Set $\varepsilon := c_4 \frac{\delta\kappa}{\sqrt{r\theta}\log(nm)}$ for some small constant $c_4 > 0$ such that $\varepsilon \le 1/10$ from the parameters given in the statement. Finally, let

$$\forall \ell \in [r], \quad v'_\ell := (1-\varepsilon)v_\ell + \sqrt{2\varepsilon - \varepsilon^2}u_\ell,$$

and let Π' be the orthogonal projection onto the subspace spanned by $v'_1, \ldots, v'_r$. Now the original data point A_j and its coupling data point A'_j (for $j \in [m]$) for matrices A, A' are drawn i.i.d. as follows:

$$A_j = \sum_{\ell=1}^{r} \zeta_\ell v_\ell + g, \quad \text{and } A'_j = \sum_{\ell=1}^{r} \zeta_\ell v'_\ell + g, \tag{20}$$

$$\text{where } \forall \ell \in [r],\ \zeta_\ell \sim \mathcal{N}(0,\theta) \quad \text{and } g \sim \mathcal{N}(0, I). \tag{21}$$

Then we bound the $\infty \to 2$ norm of Π^* and Π'.

Claim 26 *$\|\Pi^*\|_{\infty\to 2} \le \kappa$ and $\|\Pi'\|_{\infty\to 2} \le \kappa$.*

Proof of Claim 26. We have $\Pi^* = \sum_{\ell=1}^{r} v_\ell v_\ell^\top$, since $v_1, \ldots, v_r$ is an orthonormal basis for the subspace given by Π^*, and

$$\|\Pi^*\|_{\infty\to 2} = \|\Pi^*\|_{2\to 1} = \max_{y:\|y\|_2=1} \|\Pi^* y\|_1 \le \sqrt{k}\|\Pi^* y\|_2 \le \sqrt{k} \le \sqrt{\frac{2}{3}}\kappa,$$

where the first inequality follows from Cauchy-Schwartz inequality and the support size being bounded by k. Now we compute the $\infty \to 1$ norm of Π'.

$$\Pi' = \Pi^* + \sum_{\ell\in[r]} (-2\varepsilon + \varepsilon^2) v_\ell v_\ell^\top + (2\varepsilon - \varepsilon^2) \sum_{\ell} u_\ell u_\ell^\top + \sqrt{2\varepsilon - \varepsilon^2}(1-\varepsilon)(v_\ell u_\ell^\top + u_\ell v_\ell^\top)$$

$$\|\Pi'\|_{\infty\to 1} \le \|\Pi^*\|_{\infty\to 1} + 2\varepsilon \|\sum_{\ell} u_\ell u_\ell^\top\|_{\infty\to 1} + 2\sqrt{2\varepsilon}\|\sum_{\ell} u_\ell v_\ell^\top\|, \tag{22}$$

due to triangle inequality and using the monotonicity of the $\infty \to 1$ norm (Lemma 13).

For the second term, we note $\|\sum_\ell u_\ell u_\ell^\top\|_{2\to 1} \le \sqrt{r} \cdot \max_\ell \|u_\ell\|_1 \le \sqrt{rk'}$.

We now bound the third term using (18) of Lemma 24.

$$\Big\|\sum_{\ell} u_\ell v_\ell^\top\Big\|_{\infty\to 1} \le \sqrt{r} \cdot \|V\|_{\infty\to 2} \cdot \max_\ell \|u_\ell\|_1 \le \sqrt{rk'} \cdot \sqrt{\frac{2}{3}}\kappa \le \frac{1}{\sqrt{3}}(r\theta)^{1/4}\sqrt{\frac{\kappa}{\delta}} \cdot \kappa$$

given $k' \le \frac{1}{2} \cdot \frac{\sqrt{\theta}\kappa}{\delta\sqrt{r}}$. Hence substituting in (22), and using (19) we have

$$\|\Pi'\|_{\infty\to 1} \le \frac{2}{3}\kappa^2 + 2\varepsilon \cdot r \max_\ell \|u_\ell\|_1^2 + \max_\ell \|u_\ell\|_1 \cdot \|V\|_{\infty\to 1} \le \kappa^2 + 2\varepsilon \cdot r\kappa' + \sqrt{8\varepsilon/3} \cdot (\theta r)^{1/4}\sqrt{\frac{\kappa}{\delta}} \cdot \kappa$$

$$\le \frac{2\kappa^2}{3} + 2 \cdot O\Big(\frac{\delta\kappa}{\sqrt{r\theta}\log nm}\Big) \cdot r \cdot \frac{\sqrt{\theta}\kappa}{2\delta\sqrt{r}} + \sqrt{8/3} \cdot \sqrt{\frac{\delta\kappa}{\sqrt{r\theta}\log nm} \cdot \sqrt{r\theta} \cdot \kappa/\delta} \cdot \kappa \le \kappa^2,$$

given $\varepsilon = \Theta\Big(\frac{\delta\kappa}{\sqrt{r\theta}\cdot\log nm}\Big)$. Hence $\|\Pi'\|_{\infty\to 2} \le \kappa$. ∎

Claim 27 $\|\Pi^* - \Pi'\|_F^2 = \Omega(\frac{\sqrt{r}\cdot\delta\kappa}{\sqrt{\theta}\cdot\log nm})$.

Proof of Claim 27. We lower bound the distance between the projections using the orthogonality between $u_1, \ldots, u_r$ and $v_1, \ldots, v_r$:

$$\begin{aligned}\Pi' - \Pi^* &= \sum_{\ell=1}^{r} v'_\ell (v'_\ell)^\top - v_\ell v_\ell^\top \\ &= \sum_{\ell\in[r]} (-2\varepsilon + \varepsilon^2) v_\ell v_\ell^\top + (2\varepsilon - \varepsilon^2) \sum_{\ell} u_\ell u_\ell^\top + \sqrt{2\varepsilon - \varepsilon^2}(1-\varepsilon)(v_\ell u_\ell^\top + u_\ell v_\ell^\top)\end{aligned}$$

So, $\|\Pi' - \Pi^*\|_F^2 \geq r(4\varepsilon - 2\varepsilon^2) = \Omega\left(\frac{\sqrt{r}\delta\kappa}{\sqrt{\theta}\log nm}\right)$.

■

Claim 28 *With high probability, the coupling data matrix A' satisfies $\|A - A'\|_\infty \leq \delta$.*

Proof Note that $\sum_\ell \zeta_\ell v_j$ is a Gaussian with co-variance $\mathcal{N}(0, \theta\Pi^*)$, and each co-ordinate of this vector is a normal R.V. with mean 0 and variance at most $\|v_j\|_\infty^2 \sum_\ell \zeta_\ell^2$.

$$\|A_j - A'_j\|_\infty \leq \varepsilon \Big\| \sum_{\ell=1}^{r} \zeta_\ell v_\ell \Big\|_\infty + \sqrt{2\varepsilon - \varepsilon^2} \Big\| \sum_{\ell=1}^{r} \zeta_\ell u_\ell \Big\|_\infty$$

$$\begin{aligned}\text{First, } \varepsilon \Big\| \sum_{\ell=1}^{r} \zeta_\ell v_\ell \Big\|_\infty &\leq 2\varepsilon\sqrt{\theta \cdot r\log(nm)} \max_\ell \|v_\ell\|_\infty \\ &\leq 2 \cdot \Theta\Big(\frac{\delta\kappa}{\sqrt{r\theta}\cdot\log nm}\Big) \cdot \sqrt{\theta r \log(nm)}\frac{1}{\kappa} \leq \frac{\delta}{2},\end{aligned}$$

when c_4 in ε is a small constant, and since $\|v_\ell\|_\infty \leq 1/\kappa$. Bounding the second term uses the fact that the $u_1, \ldots, u_r$ have disjoint support, along with the upper bounds for $\|u_\ell\|_\infty$.

$$\begin{aligned}\sqrt{2\varepsilon - \varepsilon^2} \Big\| \sum_{\ell=1}^{r} \zeta_\ell u_\ell \Big\|_\infty &\leq 2\sqrt{\theta \cdot \varepsilon \log(nm)} \max_\ell \|u_\ell\|_\infty \\ &\leq O\left(\sqrt{\theta\log(nm) \cdot \frac{\delta\kappa}{\sqrt{r\theta}\cdot\log nm}}\right) \cdot \sqrt{\frac{\delta\sqrt{r}}{\sqrt{\theta}\kappa}} \leq \frac{\delta}{2}.\end{aligned}$$

Combining the two bounds, we see that $\|A - A'\|_\infty \leq \delta$ with high probability. ■

The correctness of Theorem 25 now follows from Claim 26, Claim 27 and Claim 28.

F.3. Asymptotic Instance-Optimal Lower Bound: Proof of Theorem 4

Proof of Theorem 4. We now establish the required properties of Π'. Firstly $u_1, \dots, u_r$ are orthogonal to each other and to Π^* (i.e., $v_1, \dots, v_r$). So, $v'_1, v'_2, \dots, v'_\ell$ are orthonormal. Hence

$$\begin{aligned}\Pi' - \Pi^* &= \sum_{\ell=1}^{r} v'_\ell (v'_\ell)^\top - v_\ell v_\ell^\top \\ &= \sum_{\ell=1}^{r} -(2\varepsilon - \varepsilon^2) v_\ell v_\ell^\top + \sum_{\ell} \frac{(2\varepsilon - \varepsilon^2)}{\|u_\ell\|_2^2} u_\ell u_\ell^\top + \sum_{\ell} \frac{(1-\varepsilon)\sqrt{2\varepsilon - \varepsilon^2}}{\|u_\ell\|_2} \Big(u_\ell v_\ell^\top + v_\ell u_\ell^\top \Big)\end{aligned} \tag{23}$$

Since each of the terms in (23) are orthogonal (w.r.t. the trace inner product) we have

$$\begin{aligned}\|\Pi' - \Pi^*\|_F^2 &= \sum_{\ell=1}^{r} (2\varepsilon - \varepsilon^2)^2 + \sum_{\ell=1}^{r} \frac{(2\varepsilon - \varepsilon^2)^2}{\|u_\ell\|_2^4} + \sum_{\ell=1}^{r} 2(2\varepsilon - \varepsilon^2) \cdot \frac{(1-\varepsilon)^2}{\|u_\ell\|_2^2} \\ &\geq r\varepsilon = \Omega\Big(\frac{\sqrt{r}\kappa\delta}{\sqrt{\lambda_1}}\Big), \text{ with probability at least } 1 - n^{-\omega(1)},\end{aligned} \tag{24}$$

for our choice of parameters (here we used (11)). Then we lower bound the distance between Σ and Σ':

$$\begin{aligned}\Sigma' - \Sigma^* &= \sum_{\ell=1}^{r} \lambda_\ell \Big(v_\ell + \big(\frac{\sqrt{2\varepsilon - \varepsilon^2}}{(1-\varepsilon)\|u_\ell\|_2}\big) u_\ell \Big) \Big(v_\ell + \big(\frac{\sqrt{2\varepsilon - \varepsilon^2}}{(1-\varepsilon)\|u_\ell\|_2}\big) u_\ell \Big)^\top - \lambda_\ell v_\ell v_\ell^\top \\ &= \sum_{\ell=1}^{r} \lambda_\ell \frac{\sqrt{2\varepsilon - \varepsilon^2}}{(1-\varepsilon)\|u_\ell\|_2} (v_\ell u_\ell^\top + u_\ell v_\ell^\top) + \lambda_\ell \frac{2\varepsilon - \varepsilon^2}{(1-\varepsilon)^2 \|u_\ell\|_2^2} u_\ell u_\ell^\top .\end{aligned}$$

Because v_ℓ and u_ℓ are orthogonal and using (11), $\|\Sigma^* - \Sigma'\|_F^2$ is with high probability at least

$$\big(\sum_{\ell=1}^{r} \lambda_\ell^2\big)\varepsilon = \Big(\frac{\lambda_1^2 + \cdots + \lambda_r^2}{r}\Big) \|\Pi' - \Pi^*\|_F^2).$$

We now show that A' is a valid δ-perturbation of A. Recall the definition of A_j, A'_j in (10) respectively. For each fixed $j \in [m]$, by Lemma 23, we have with probability at least $1 - m^{-2}$ (over the randomness in $\{ \zeta_\ell^{(j)} : \ell \in [r] \}$) that

$$\begin{aligned}\|A_j - A'_j\|_\infty &= \Big\| \sum_{\ell} \sqrt{\lambda_\ell} \zeta_\ell^{(j)} \cdot \frac{\sqrt{2\varepsilon - \varepsilon^2}}{(1-\varepsilon)\|u_\ell\|_2} u_\ell \Big\|_\infty \\ &\leq \frac{2\sqrt{\log(rm)}}{(1-\varepsilon)} \cdot \max_{\ell \in [r]} \sqrt{2\varepsilon\lambda_\ell} \frac{\|u_\ell\|_\infty}{\|u_\ell\|_2},\end{aligned}$$

where the second term uses the fact that $u_1, \dots, u_r$ are disjoint and the concentration of Gaussian random variables (over $\zeta_\ell^{(j)}$). See also Lemma 23 for general q. After a union

bound over all $j \in [m]$, and using our bounds on $\|u_\ell\|_2$ and $\|u_\ell\|_\infty$ from Lemma 11 along with our definition of ε, we get with probability at least $1 - \eta - \frac{1}{m}$,

$$\max_{j\in[m]} \|A_j - A'_j\|_\infty \le \frac{2\sqrt{\log(rm)}}{(1-\varepsilon)} \cdot \sqrt{2\varepsilon\lambda_1} \cdot 2\sqrt{\frac{\log(rk'n)}{(k'-r)}} \cdot \frac{1}{1/2}$$
$$= O\Big(\frac{\sqrt{\log(rm)\log n}}{\sqrt{k'}} \cdot \sqrt{\varepsilon\lambda_1}\Big) \le \delta,$$

since $\varepsilon = c\delta^2 k'/(\lambda_1 \log(rm)\log n)$ for a small constant c (and $\varepsilon < 1/2$).

Upper bound on $\|\Pi'\|_{\infty\to 2}$*:* The proof will follow the same outline as for Theorem 25. We compute the $\infty \to 1$ norm of Π'; recall that the $\infty \to 1$ norm satisfies the matrix norm monotone property (Lemma 13). From (23), triangle inequality and monotonicity,

$$\|\Pi'\|_{\infty\to 1} \le \underbrace{\|\Pi^*\|_{\infty\to 1}}_{\text{equal to } \kappa^2} + \underbrace{2\Big\|\sum_\ell \varepsilon u_\ell u_\ell^\top\Big\|_{\infty\to 1}}_{\text{bound using (19)}} + \underbrace{2\Big\|\sum_\ell \sqrt{2\varepsilon-\varepsilon^2}u_\ell v_\ell^\top\Big\|_{\infty\to 1}}_{\text{bound using (18)}}. \tag{25}$$

We first bound the third term using (18) of Lemma 24.

$$\Big\|\sum_\ell \sqrt{2\varepsilon-\varepsilon^2}u_\ell v_\ell^\top\Big\|_{\infty\to 1} \le \sqrt{2\varepsilon}\sqrt{r}\|V\|_{\infty\to 2} \cdot \max_\ell \|u_\ell\|_1 \le \kappa\sqrt{2rk'\varepsilon}$$
$$\le \frac{\kappa^2}{(\log n \log m)^{1/2}},$$

by substituting the values for k', ε and using $rk'\varepsilon = O(\kappa^2/(\log n \log m))$. Hence substituting in (25) and using (19),

$$\|\Pi'\|_{\infty\to 1} \le \kappa^2 + 2\varepsilon\|U\|_{2\to 1}^2 + \kappa^2 \cdot \frac{1}{(\log n \log m)^{1/2}}$$
$$\le \kappa^2 + 2r\varepsilon \max_\ell \|u_\ell\|_1^2 + \kappa^2 \cdot \Big(\frac{1}{(\log n \log m)^{1/2}}\Big)$$
$$\le \kappa^2 + 4r \cdot \varepsilon k' + o(\kappa^2) \le (1+o(1))\kappa^2.$$

■

F.3.1. Extension to general ℓ_q norm

Theorem 4 extends in a straightforward fashion to also hold for ℓ_q norms.

Theorem 29 *Suppose we are given parameters $r \in \mathbb{N}, q \ge 1, \kappa \ge 2r^{1-2/q}$ and $\delta > 0$. In the notation of Theorem 3, for any Σ^*, given m samples $A_1, \ldots, A_m$ generated i.i.d. from $\mathcal{N}(0, \Sigma^*)$ with $\kappa = \|\Pi^*\|_{q\to 2}$ satisfying $\sqrt{r\lambda_1}(\kappa/n^{1-2/q}) \le \delta \le \sqrt{r\lambda_1}/\kappa$, there exists a covariance matrix Σ' with a projector Π' onto its top-r principal subspace, and an alternate dataset $A'_1, \ldots, A'_m$ drawn i.i.d. from $\mathcal{N}(0, \Sigma')$ satisfying $\|\Pi'\|_{q\to 2} \le (1+o(1))\kappa$, and $\|A'_j - A_j\|_q \le \delta \ \forall j \in [m]$,*

but $\|\Pi^* - \Pi'\|_F^2 \ge \big(\frac{\Omega(1)}{\sqrt{\lambda_1}\log(rm)\log n}\big) \cdot \sqrt{r}\kappa\delta$, *and* $\|\Sigma'_{\text{TOP}} - \Sigma_{\text{TOP}}\|_F^2 \ge \frac{(\lambda_1^2+\cdots+\lambda_r^2)}{r} \cdot \|\Pi' - \Pi^*\|_F^2$

In particular, when $\Sigma_{\text{TOP}} = (1+\theta)\Pi^*$ *then* $\Sigma'_{\text{TOP}} = (1+\theta')\Pi'$ *with* $\theta' = (1+o(1))\theta$.

In fact the same construction holds using $u_1, \ldots, u_r$ that are picked randomly but with disjoint support. However, there is a minor change in the parameters of the construction. We will set ε as before (and hence this will give the same lower bound on $\|\Pi' - \Pi^*\|_F^2$ and $\|\Sigma' - \Sigma^*\|_F^2$). We will set

$$\varepsilon = \frac{c\kappa\delta}{\sqrt{r\lambda_1}\log(rm)\log n} \text{ and } (k')^{1-2/q} := \Big(\frac{r^{2/q}\varepsilon\lambda_\ell}{\delta^2\log(rm)\log n}\Big),$$

for some constant $c > 0$. The assumptions of the theorem ensure that $2r \le k' \ll n/r$ as required for the construction.

We will need an additional simple claim that just extends Lemma 11.

Lemma 30 *In the notation of Lemma 11 for any $\eta < 1$, with probability at least $(1-\eta)$ we have*

$$\forall \ell \in [r], \qquad \|u_\ell\|_q \le 3\sqrt{\log(rk'/\eta)}(k')^{-1/2+1/q}. \tag{26}$$

$$\|u_\ell\|_{q^*} \le 2(k')^{1/2-1/q}. \tag{27}$$

The proof follows directly from Lemma 11 and using the relation between the ℓ_q, ℓ_∞ norms, and ℓ_{q^*}, ℓ_1 norms.

Completing the proof of Theorem 29. The proof follows the same argument as the proof of Theorem 4. As mentioned before, since we choose the same ε, it suffices to argue about $\max_{j\in[m]}\|A_j - A'_j\|_q$ and $\|\Pi'\|_{q\to q^*}$.

To establish the upper bound on $\|\Pi'\|_{q\to q^*}$ we use the bounds in Lemma 24 and (26). We have from Lemma 24

$$\begin{aligned}
\|\Pi'\|_{q\to q^*} &\le \|\Pi^*\|_{\infty\to 1} + 2\varepsilon\|UU^\top\|_{q\to q^*} + 2\sqrt{2\varepsilon-\varepsilon^2}\|UV^\top\|_{q\to q^*} \\
&\le \kappa^2 + 2\varepsilon r^{1-2/q}\Big(\max_{\ell\in[r]}\|u_\ell\|_{q^*}\Big)^2 + 2\sqrt{\varepsilon}r^{1/2-1/q}\Big(\max_{\ell\in[r]}\|u_\ell\|_{q^*}\Big)\cdot\|V\|_{q\to 2} \\
&\le \kappa^2 + 2\varepsilon r^{1-2/q}(k')^{1-2/q} + 2\sqrt{\varepsilon}r^{1/2-1/q}(k')^{1/1-1/q}\cdot\kappa \\
&\le \kappa^2 + o(\kappa^2) + o(\kappa)\cdot\kappa = \kappa^2(1+o(1)),
\end{aligned}$$

since from our choice of parameter k', we have $\varepsilon r^{1-2/q}\max_\ell\|u_\ell\|_{q^*}^2 = (\varepsilon^2 r\lambda_1)/(\delta^2\log(rm)\log n) = o(\kappa^2)$.

Finally, for the upper bound on $\max_{j\in[m]}\|A_j - A'_j\|_q \le \delta$ we use Lemma 23 and (27). For each fixed $j \in [m]$, by Lemma 23, we have with probability at least $1 - m^{-2}$ (over the randomness in $\{\zeta_\ell^{(j)} : \ell \in [r]\}$) that

$$\begin{aligned}
\|A_j - A'_j\|_q &= \Big\|\sum_\ell \sqrt{\lambda_\ell}\zeta_\ell^{(j)}\cdot\frac{\sqrt{2\varepsilon-\varepsilon^2}}{(1-\varepsilon)\|u_\ell\|_2}u_\ell\Big\|_q \\
&\le r^{1/q}\frac{2\sqrt{\log(rm)}}{(1-\varepsilon)}\cdot\max_{\ell\in[r]}\sqrt{2\varepsilon\lambda_\ell}\frac{\|u_\ell\|_q}{\|u_\ell\|_2} \\
&\le r^{1/q}\cdot\sqrt{\log(rm)}(k')^{-1/2+1/q} \le \delta,
\end{aligned}$$

for our choice of parameters and k'. This establishes the statement of Theorem 29 for general q.

Appendix G. Statistical Upper bounds (computationally inefficient algorithm)

We show the statistical upper bounds on the recovery of principle components in this section. By symmetrization (shown in Algorithm 2), we assume all data points are generated from $\mathcal{N}(0, \Sigma^*)$ rather than $\mathcal{N}(\mu, \Sigma^*)$ in this section.

Theorem 31 *Given $q > 2$, n, r, and κ, let $\mathcal{P} = \{$projection matrix $\Pi \big| rank = r$ and $\|\Pi\|_{q\to 2} \le \kappa\}$. Let Σ be an unknown covariance matrix with eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ whose projection matrix Π^* of the top r eigenspace is in $\mathcal{P}$.*

Let $\widetilde{A} \in \mathbb{R}^{n\times m}$ be the δ-perturbed (in ℓ_q norm) data matrix where each original column comes from $\mathcal{N}(0, \Sigma^*)$ for any $\delta > 0$, $\varepsilon > 0$ and $m \ge C \cdot \lambda_1^2 \cdot r^2\kappa^2 \log n \cdot n^{2/q}/\varepsilon^2$. Then

$$\widetilde{\Pi} \stackrel{def}{=} \arg\min_{\Pi\in\mathcal{P}} \{\|\widetilde{A}\|_F^2 - \|\Pi\widetilde{A}\|_F^2\}$$

satisfies $\|\widetilde{\Pi}^\perp \Pi^*\|_F^2 \le \frac{1}{\lambda_r - \lambda_{r+1}} \cdot O\left(\delta^2\kappa^2 + \sqrt{\lambda_1 r}\cdot\delta\kappa + \varepsilon\right)$ with probability 0.99. Moreover, one can obtain $\widetilde{\Sigma}_{\text{TOP}}$ satisfying $\|\widetilde{\Sigma}_{\text{TOP}} - \Sigma_{\text{TOP}}\|_F^2 \le O(\lambda_1^2 \cdot \|\widetilde{\Pi}^\perp\Pi^*\|_F^2 + \lambda_1\kappa^2\delta^2 + \kappa^4\delta^4)$ where $\|\widetilde{\Pi}^\perp\Pi^*\|_F^2$ is upper bounded above.

Remark 32 *Comparing to the computational upper bound in Theorem 6, the main difference is the dependency of m on κ: it becomes κ^2 here.*

We state the direct corollary in the spiked covariance model with $q = \infty$.

Corollary 33 *Given n, r, and κ, let $\mathcal{P} = \{\Pi \big| rank = r$ and $\|\Pi\|_{\infty\to 2} \le \kappa\}$. For any θ and $\Pi^* \in \mathcal{P}$, let $\widetilde{A} \in \mathbb{R}^{n\times m}$ be the δ-perturbed data matrix where each original column comes from $\mathcal{N}(0, I + \theta\Pi^*)$. For any $\delta > 0$, $\varepsilon > 0$ and $m \ge C \cdot (1+\theta)^2 \cdot r^2\kappa^2 \log n/\varepsilon^2$,*

$$\widetilde{\Pi} \stackrel{def}{=} \arg\min_{\Pi\in\mathcal{P}} \{\|\widetilde{A}\|_F^2 - \|\Pi\widetilde{A}\|_F^2\}$$

satisfies $\|\widetilde{\Pi}^\perp\Pi^*\|_F^2 \le \frac{1}{\theta} \cdot O\left(\delta^2\kappa^2 + (1+\theta)^{1/2}\sqrt{r}\cdot\delta\kappa + \varepsilon\right)$ with probability 0.99.

We show two technical results to prove the main theorem. The first one bounds the deviation of the inner product between all projection matrices and the original data matrix (before perturbation), whose proof is defered to Section G.1.

Lemma 34 *For any covariance matrix Σ^* whose eigenvalues are at most $\lambda_{\max}$, let $A \in \mathbb{R}^{n\times m}$ be a data matrix where each column is generated from $\mathcal{N}(0, \Sigma^*)$.*

Given n, q, r and κ, let $\mathcal{P} = \{\Pi \big| rank = r$ and $\|\Pi\|_{q\to 2} \le \kappa\}$. Then for any $m \ge C\lambda_{\max}^2 \cdot \kappa^2 \log n \cdot n^{2/q}$ with a sufficiently large constant C, we have that with probability 0.99,

$$\left|\left\langle \frac{1}{m}AA^\top - \Sigma^*, \Pi \right\rangle\right| = r \cdot O\left(\frac{\lambda_{\max}\cdot\kappa\cdot\sqrt{\log n}\cdot n^{1/q}}{\sqrt{m}}\right) \text{ for all } \Pi \in \mathcal{P}.$$

Then we bound the deviation of the inner product between all projection matrices and the actual data matrix (after perturbation) from the expectation.

Claim 35 *Given n, r, and κ, let $\mathcal{P} = \{\Pi | rank = r \text{ and } \|\Pi\|_{q\to 2} \le \kappa\}$. For an unknown covariance matrix Σ^*, let λ_1 denote the largest eigenvalue of Σ^*.*

Let $A \in \mathbb{R}^{n\times m}$ be the original data matrix where each column generated from $\mathcal{N}(0, \Sigma^*)$ and $\widetilde{A}$ be its δ-perturbation (ℓ_q norm in every column) for $m \ge C\lambda_1^2 \cdot \kappa^2 \log n \cdot n^{2/q}$ with a sufficiently large constant C. With probability 0.98,

$$\left|\left\langle \frac{1}{m}\widetilde{A}\cdot\widetilde{A}^\top - \Sigma^*, \Pi\right\rangle\right| = O\left(\lambda_1 \cdot r\kappa \cdot \sqrt{\frac{\log n}{m}} \cdot n^{1/q} + \delta^2\kappa^2 + \sqrt{\lambda_1 r}\cdot\delta\kappa\right) \text{ for all } \Pi \in \mathcal{P}.$$

Proof of Claim 35. We rewrite the left hand side as

$$\begin{aligned}
&\left|\left\langle \frac{1}{m}\widetilde{A}\widetilde{A}^\top - \Sigma^*, \Pi\right\rangle\right| \\
\le&\left|\left\langle \frac{1}{m}AA^\top - \Sigma^* + \frac{1}{m}(\widetilde{A}-A)A^\top + \frac{1}{m}\widetilde{A}(\widetilde{A}-A)^\top, \Pi\right\rangle\right| \\
\le&\left|\left\langle \frac{1}{m}AA^\top - \Sigma^*, \Pi\right\rangle\right| + \left|\left\langle \frac{1}{m}(\widetilde{A}-A)A^\top, \Pi\right\rangle\right| + \left|\left\langle \frac{1}{m}\widetilde{A}(\widetilde{A}-A)^\top, \Pi\right\rangle\right| \\
\le&\left|\left\langle \frac{1}{m}AA^\top - \Sigma^*, \Pi\right\rangle\right| + 2\left|\left\langle \frac{1}{m}(\widetilde{A}-A)A^\top, \Pi\right\rangle\right| + \left|\left\langle \frac{1}{m}(\widetilde{A}-A)(\widetilde{A}-A)^\top, \Pi\right\rangle\right|
\end{aligned}$$

By Lemma 34, the first term $\left|\left\langle \frac{1}{m}AA^\top - \Sigma^*, \Pi\right\rangle\right|$ is upper bounded by $O\left(r\cdot\lambda_1\kappa\cdot\sqrt{\frac{\log n}{m}}\cdot n^{1/q}\right)$ with probability 0.99. Since $\|\widetilde{A}_i - A_i\|_q \le \delta$ and $\|\Pi\|_{q\to 2} \le \kappa$, the last term is upper bounded by

$$\frac{1}{m}\left|\left\langle (\widetilde{A}-A)(\widetilde{A}-A)^\top, \Pi^2\right\rangle\right| = \frac{1}{m}\|\Pi(\widetilde{A}-A)\|_F^2 \le \delta^2\kappa^2.$$

We bound the second term here.

$$\frac{1}{m}\left|\left\langle (\widetilde{A}-A)A^\top, \Pi\right\rangle\right| = \frac{1}{m}\left|\langle \Pi(\widetilde{A}-A), \Pi A\rangle\right| \le \frac{1}{m}\|\Pi(\widetilde{A}-A)\|_F \cdot \|\Pi A\|_F.$$

The first part $\|\Pi(\widetilde{A}-A)\|_F$ is always $\le \sqrt{m}\delta\kappa$ from the definition of Π. For the second part, notice that

$$\|\Pi A\|_F^2 = \langle AA^\top, \Pi\rangle \le \langle m\Sigma^*, \Pi\rangle + \left|\langle AA^\top - m\Sigma^*, \Pi\rangle\right| \le \lambda_1\cdot rm + O\left(r\lambda_1\cdot\kappa\cdot\sqrt{m\log n}\cdot n^{1/q}\right),$$

where the two bounds come from Fact 17 and Lemma 34 separately. So the second term is upper bounded by

$$\frac{1}{m}\cdot\sqrt{m}\delta\kappa\cdot\left(\lambda_1\cdot rm + C_0\cdot r\lambda_1\cdot\kappa\cdot\sqrt{m\log n}\cdot n^{1/q}\right)^{1/2} \le \sqrt{r\lambda_1}\cdot\delta\kappa + \lambda_1^{1/2}\cdot C_0^{1/2}\cdot\delta\kappa\cdot\left(\frac{r\kappa\sqrt{\log n}\cdot n^{1/q}}{\sqrt{m}}\right)^{1/2}.$$

So the total error is

$$O\left(r\cdot\lambda_1\kappa\cdot\sqrt{\frac{\log n}{m}}\cdot n^{1/q}\right) + \delta^2\kappa^2 + \sqrt{\lambda_1\cdot r}\delta\kappa + \lambda_1^{1/2}\cdot C_0^{1/2}\cdot\delta\kappa\cdot\left(\frac{r\kappa\sqrt{\log n}\cdot n^{1/q}}{\sqrt{m}}\right)^{1/2}. \quad (28)$$

Finally we simplify the error terms. The last term

$$\lambda_1^{1/2} C_0^{1/2} \cdot \delta\kappa \cdot (\frac{r\kappa\sqrt{\log n} \cdot n^{1/q}}{\sqrt{m}})^{1/2} = O\left(\delta^2\kappa^2 + \lambda_1 \cdot \frac{r\kappa\sqrt{\log n} \cdot n^{1/q}}{\sqrt{m}}\right),$$

which are the first two terms in the total error (28). ■

Finally, we finish the proof of Theorem 31.

Proof of Theorem 31. Notice that the output projection $\widetilde{\Pi}$ could also be defined as $\arg\max_{\Pi\in\mathcal{P}}\{\|\Pi\widetilde{A}\|_F^2\}$ and for any projection matrix Π,

$$\frac{1}{m}\|\Pi\widetilde{A}\|_F^2 = \frac{1}{m}\langle\widetilde{A}\widetilde{A}^\top, \Pi\rangle.$$

By Claim 35, every Π has $\frac{1}{m}\langle\widetilde{A}\widetilde{A}^\top, \Pi\rangle$ around $\langle\Sigma^*, \Pi\rangle \pm \Delta$ for

$$\Delta := O\left(r\lambda_1 \cdot \kappa \cdot \sqrt{\frac{\log n}{m}} \cdot n^{1/q} + \delta^2\kappa^2 + \sqrt{r\lambda_1} \cdot \delta\kappa\right) \text{ (the error in Claim 35).}$$

Since $\widetilde{\Pi}$ attains a better objective value than Π^*, we have

$$\begin{aligned}
\langle\Sigma^*, \widetilde{\Pi}\rangle &\geq \left\langle\frac{1}{m}\widetilde{A}\widetilde{A}^\top, \widetilde{\Pi}\right\rangle - \Delta \\
&\geq \left\langle\frac{1}{m}\widetilde{A}\widetilde{A}^\top, \Pi^*\right\rangle - \Delta && \text{(using the definition of } \widetilde{\Pi}) \\
&\geq \langle\Sigma^*, \Pi^*\rangle - 2\Delta.
\end{aligned}$$

Next, we apply Claim 9 to conclude $\langle\Pi^*, \widetilde{\Pi}\rangle \geq r - \frac{2\Delta}{\lambda_r - \lambda_{r+1}}$, which upper bounds $\|\widetilde{\Pi}^\perp\Pi^*\|_F^2 \leq \frac{2\Delta}{\lambda_r - \lambda_{r+1}}$. Finally we use Theorem 10 to get $\widetilde{\Sigma}_{\text{TOP}}$ satisfying $\|\widetilde{\Sigma}_{\text{TOP}} - \Sigma_{\text{TOP}}\|_F^2 \leq O(\lambda_1^2 \cdot \frac{2\Delta}{\lambda_r - \lambda_{r+1}} + \lambda_1\kappa^2\delta^2 + \kappa^4\delta^4)$. ■

G.1. Proof of Lemma 34

We use the following concentration result from Mendelson (2010) to bound the supremum.

Lemma 36 (See Corollary 4.1 in Vu and Lei (2012)) *Let $A_1, \ldots, A_m \in \mathbb{R}^n$ be i.i.d. mean 0 random vectors with*

$$\Sigma = \mathbb{E}\, A_1 A_1^\top \text{ and } \sigma = \sup_{\|u\|_2=1} \big\|\langle A_1, u\rangle\big\|_{\psi_2}.$$

For $S_n = \frac{1}{m}\sum_{i=1}^m A_i \cdot A_i^\top$ and a symmetric subset $\mathcal{V}$ in $\mathbb{R}^n$, we have

$$\mathop{\mathbb{E}}_{A_1,\ldots,A_m}\left[\sup_{v\in\mathcal{V}}\left|\left\langle S_n - \Sigma, vv^\top\right\rangle\right|\right] \leq c\left(\frac{\sigma^2}{\sqrt{m}} \cdot \sup_{v\in\mathcal{V}}\|v\|_2 \cdot \mathop{\mathbb{E}}_g\left[\sup_{v\in\mathcal{V}}\langle g, v\rangle\right] + \frac{\sigma^2}{m}\mathop{\mathbb{E}}_g\left[\sup_{v\in\mathcal{V}}\langle g, v\rangle\right]^2\right)$$

for a vector $g \in \mathbb{R}^n$ with i.i.d. Gaussian entries and a universal constant c.

To use the above lemma, we first upper bound σ^2 in our setting.

Claim 37 *Let $X \sim \mathcal{N}(0, \Sigma^*)$ for a matrix Σ^* with eigenvalues at most $\lambda_{\max}$. Then $\big\|\langle X, u\rangle\big\|_{\psi_2} \le \sqrt{\lambda_{\max}(\Sigma^*)}$ for any u with $\|u\|_2 = 1$.*

Proof Let $v_1, \dots, v_n$ be the eigenvectors of Σ^* with eigenvalues $\lambda_1, \dots, \lambda_n$. Then $\langle X, u\rangle = \sqrt{\lambda_1} \cdot \langle v_1, u\rangle g_1 + \cdots + \sqrt{\lambda_n} \cdot \langle v_n, u\rangle g_n$ for i.i.d. Gaussian random variable $g_1, \dots, g_n$. So the variance is $\lambda_1 \langle v_1, u\rangle^2 + \cdots + \lambda_n \langle v_n, u\rangle^2 \le \max\{\lambda_1, \dots, \lambda_n\}$ and

$$\big\|\langle X, u\rangle\big\|_{\psi_2} \le \sqrt{\lambda_{\max}}.$$

■

We apply Lemma 36 to all vectors that could be in the basis of possible Π.

Claim 38 *For any covariance matrix Σ^* with eigenvalues at most $\lambda_{\max}$, let $A_1, \dots, A_m \in \mathbb{R}^n$ be i.i.d. vectors generated from $\mathcal{N}(0, \Sigma^*)$. Given n and q, let $\mathcal{V}$ be the set of all vectors v with $\|v\|_2 = 1$ and $\|v\|_{q^*} \le \kappa$.*

Then for any $m \ge C\lambda_{\max}^2 \cdot \kappa^2 \log n \cdot n^{2/q}$ with a sufficiently large constant C, we have that with probability 0.99,

$$\left|\left\langle \frac{1}{m}\sum_{i=1}^m A_i A_i^\top - \Sigma^*, vv^\top \right\rangle\right| = O\left(\frac{\lambda_{\max}\kappa\sqrt{\log n}\cdot n^{1/q}}{\sqrt{m}}\right) \text{ for all } v \in \mathcal{V}.$$

Proof To apply Lemma 36, we notice that $\sup_{v\in\mathcal{V}} \|v\|_2 = 1$ and

$$\mathop{\mathbb{E}}_{g}\left[\sup_{v\in\mathcal{V}}\langle g, v\rangle\right] \le \mathbb{E}\left[\sup_{v} \|g\|_q \cdot \|v\|_{q^*}\right] = \mathbb{E}[\|g\|_q] \cdot \sup \|v\|_{q^*} = O(n^{1/q}\sqrt{\log n}\cdot\kappa).$$

Thus Lemma 36 shows that for some absolute constant $c' > 0$

$$\mathop{\mathbb{E}}_{A_1,\dots,A_m}\left[\sup_{v\in\mathcal{V}}\left|\left\langle \frac{1}{m}\sum_{i=1}^m A_i A_i^\top - \Sigma^*, vv^\top\right\rangle\right|\right] = \frac{c'\lambda_{\max}\cdot 1\cdot\kappa\sqrt{\log n}\cdot n^{1/q}}{\sqrt{m}} + \frac{c'\lambda_{\max}\kappa^2\log n\cdot n^{2/q}}{m}.$$

When $m > C\lambda_{\max}^2\cdot\kappa^2\log n\cdot n^{2/q}$, the right hand is at most twice the first term $O(\frac{\lambda_{\max}\cdot\kappa\sqrt{\log n}\cdot n^{1/q}}{\sqrt{m}})$. Next we apply the Markov inequality to replace the expectation by probability 0.99. ■

Lemma 34 follows as a corollary of the above claim: for any Π of rank r and $\|\Pi\|_{q\to 2} \le \kappa$, we have $\|\Pi\|_{2\to q^*} = \|\Pi\|_{q\to 2} = \kappa$ such that all its eigenvectors $v_1, \dots, v_r$ are in $\mathcal{V}$ with $\|v_i\|_{q^*} \le \kappa$ (by considering $\|\Pi v_i\|_{q^*} \le \kappa$). Thus

$$\begin{aligned}\left|\left\langle \frac{1}{m}\sum_{i=1}^m A_i A_i^\top - \Sigma^*, \Pi\right\rangle\right| &= \left|\left\langle \frac{1}{m}\sum_{i=1}^m A_i A_i^\top - \Sigma^*, \sum_{j=1}^r v_j v_j^\top\right\rangle\right| \\ &\le \sum_{j=1}^r \left|\left\langle \frac{1}{m}\sum_{i=1}^m A_i A_i^\top - \Sigma^*, v_j v_j^\top\right\rangle\right| = r\cdot O\left(\frac{\lambda_{\max}\kappa\cdot\sqrt{\log n}\cdot n^{1/q}}{\sqrt{m}}\right).\end{aligned}$$

Appendix H. Robust Mean Estimation

In this section we present an analysis of the robust mean estimation procedure sketched below, thereby establishing Proposition 1.

Algorithm 3 Mean Estimation under Adversarial Perturbations

1: **function** AdvRobustMean(m samples $\tilde{A}_1, \ldots, \tilde{A}_m \in \mathbb{R}^n$, norm q, perturbation δ, error η)
2: Compute the empirical mean μ' of all the given samples.
3: Output $\tilde{\mu}$, where $\tilde{\mu}$ is the point in the ℓ_q ball of size $\delta+\eta$ around μ' with the minimum ℓ_{q^*} norm i.e.,

$$\min_{u \in \mathbb{R}^n} \|u\|_{q^*}^{q^*}, \text{ s.t. } \|u - \mu'\|_q \leq \delta + \eta.$$

We remark that the above algorithm in the case of $q = \infty$ specializes to $\forall i \in [n], \quad \tilde{\mu}(i) = \text{sign}(\mu'(i)) \cdot \max\{\ |\mu'(i)| - (\delta+\eta), 0\}$. This is the same as the soft-thresholding algorithm that has been explored in the sparse mean estimation literature. More generally, we will prove the statement for any ℓ_q norm for $q \geq 2$. The main theorem of this section is the following

Proposition 39 *Fix $q \geq 2$. Suppose we have m samples drawn according to the Adversarial Perturbation model with ℓ_q perturbations. There is a polynomial time algorithm (Algorithm 3) that outputs an estimate $\hat{\mu}$ for the (unknown) mean μ such that with probability at least $(1 - 1/n)$,*

$$\|\hat{\mu} - \mu\|_2^2 \leq 4 \min\left\{\|\mu\|_{q^*}(\delta+\eta), n^{1-\frac{1}{q}}(\delta+\eta)^2\right\}, \text{ where } \eta := 2\sigma n^{\frac{1}{q}}\sqrt{\frac{\log n}{m}}. \tag{29}$$

Proof Let $\mu' = \text{mean}(\tilde{A})$. Since $\|\tilde{A}_j - A_j\|_q \leq \delta$ for each $j \in [m]$, we know that $\|\mu' - \text{mean}(A)\|_q \leq \delta$. Furthermore, from standard Gaussian concentration as stated in Fact 40 below we have that with probability at least $1 - \frac{1}{n}$ it holds that

$$\|\mu - \text{mean}(A)\|_q \leq \eta = 2\sigma n^{\frac{1}{q}}\sqrt{\frac{\log n}{m}}. \tag{30}$$

This implies that with probability at least $1 - \frac{1}{n}$,

$$\|\mu - \mu'\|_q \leq \delta + \eta \tag{31}$$

and hence is a valid solution to the convex program in Algorithm 3. Moreover the convex program can be solved in polynomial time using the Ellipsoid method. This is because the objective is separable over the data points, and for each constraint is of the form $\|z\|_p \leq \tau$, where τ is specified and $p \geq 1$. A simple hyperplane separation oracle for a constraint of the form $\|z\|_p \leq \tau$ is given by the duality since

$$\|z\|_p = \max_{y \in \mathbb{R}^n : \|y\|_{p^*} \leq 1} \langle y, z\rangle = \left\langle \frac{z^*}{\|z^*\|_{p^*}}, z \right\rangle, \text{ where } z_i^* = \text{sign}(z_i)|z(i)|^{p-1} \quad \forall i \in [n].$$

Hence a hyperplane of the form $\langle w, z\rangle \leq \tau$ with $w = z^*/\|z^*\|_{p^*}$ gives a valid separation oracle. A similar separation oracle can also be used for the objective. (Note that one can also use the projected sub-gradient method for a more effective algorithm).

This implies that the Algorithm outputs a vector $\hat{\mu}$ in polynomial time. It satisfies

$$\|\hat{\mu}\|_{q^*} \leq \|\mu\|_{q^*} \tag{32}$$

Hence, via Hölder's inequality we get that

$$\begin{aligned}
\|\hat{\mu} - \mu\|_2^2 &\leq \|\hat{\mu} - \mu\|_q \|\hat{\mu} - \mu\|_{q^*} \\
&\leq (\|\hat{\mu} - \mu'\|_q + \|\mu - \mu'\|_q)(\|\hat{\mu}\|_{q^*} + \|\mu\|_{q^*}) \\
&\leq 2(\|\hat{\mu} - \mu'\|_q + \|\mu - \mu'\|_q)\|\mu\|_{q^*} \text{ [from (32)]} \\
&\leq 4\|\mu - \mu'\|_q \|\mu\|_{q^*} \text{ [from the optimality of } \hat{\mu}.] \\
&\leq 4(\delta + \eta)\|\mu\|_{q^*} \text{ [from (31)]}
\end{aligned} \tag{33}$$

Alternately, using the fact that for any vector $x \in \mathbb{R}^n$, $\|x\|_p \leq n^{\frac{1}{p} - \frac{1}{q}}\|x\|_q$ we get that

$$\begin{aligned}
\|\hat{\mu} - \mu\|_2^2 &\leq n^{1-\frac{1}{q}}\|\hat{\mu} - \mu\|_q^2 \\
&\leq n^{1-\frac{1}{q}}\Big(\|\hat{\mu} - \mu'\|_q + \|\mu - \mu'\|_q\Big)^2 \\
&\leq 4n^{1-\frac{1}{q}}\|\mu - \mu'\|_q^2 \text{ [from the optimality of } \hat{\mu}.] \\
&\leq 4n^{1-\frac{1}{q}}(\delta + \eta)^2 \text{ [from (31)].}
\end{aligned} \tag{34}$$

Combining (33) and (34) we get the claim. Setting $q = \infty$ establishes Proposition 1 from the introduction. ■

To complete the argument we provide a self contained proof of the fact stated below.

Fact 40 *Fix $q \geq 2$. Let $A_1, \ldots, A_m$ be drawn i.i.d. from $N(0, \Sigma_{n\times n})$ with $\|\Sigma\| \leq \sigma^2$. Then with probability at least $1 - \frac{1}{n}$ it holds that,*

$$\|\frac{1}{m}\sum_{i=1}^{m} A_i\|_q \leq 2\sigma n^{\frac{1}{q}}\sqrt{\frac{\log n}{m}}.$$

Proof *Noticing that each coordinate of $\frac{1}{m}\sum_{i=1}^m A_i$ is a mean Gaussian with variance bounded by σ^2/m and using union bound we get that with probability at least $1 - \frac{1}{n}$,*

$$\|\frac{1}{m}\sum_{i=1}^{m} A_i\|_\infty \leq 2\sigma\sqrt{\frac{\log n}{m}}.$$

Then it easily follows that with probability at least $1 - \frac{1}{n}$,

$$\begin{aligned}
\|\frac{1}{m}\sum_{i=1}^{m} A_i\|_q &\leq n^{\frac{1}{q}}\|\frac{1}{m}\sum_{i=1}^{m} A_i\|_\infty \\
&\leq 2\sigma n^{\frac{1}{q}}\sqrt{\frac{\log n}{m}}.
\end{aligned}$$

■

Notice that the bound of $n^{1-\frac{1}{q}}(\delta+\eta)^2$ is the naive bound that is simply achieved by always outputting the mean of the points in $\tilde{A}$. Hence, for small values of the perturbation δ, the algorithm achieves a non-trivial guarantee of $\|\mu\|_{q^*}(\delta+\eta)$. In fact we next show that the guarantee of the algorithm is optimal. In particular, provide an instance wise lower bound, stated below, for robust mean estimation in our model of corruption.

Proposition 41 *Fix $q=\infty$. Let μ be any vector such that the analytical sparsity of μ, i.e., $\frac{\|\mu\|_1}{\|\mu\|}$ is bounded by $\sqrt{n}/4$. Then there exist $\delta,\sigma>0$ and another vector $\|\mu'\|$ such that $\frac{\|\mu'\|_1}{\|\mu'\|_2}=\frac{\|\mu\|_1}{\|\mu\|_2}(1+o(1))$, and $\|\mu-\mu'\|_2=\Omega(\sqrt{\delta\|\mu\|}_1)$ and with high probability, i.i.d. samples $A_1,A_2,\dots A_m$ generated from $\mathcal{N}(\mu,\sigma^2 I)$ and $\tilde{A}_1,\tilde{A}_2,\dots\tilde{A}_m$ generated from $\mathcal{N}(\mu',\sigma^2 I)$ satisfy $\|A_j-\tilde{A}_j\|_\infty\le\delta$, for all $j\in[m]$.*

Proof The construction builds upon the argument presented in Awasthi et al. (2019a) with most of the details unchanged. We provide a proof sketch here. Pick a subset S of $s=(\frac{\|\mu\|_1}{\|\mu\|_2})^2$ coordinates and define $\mu'=\mu+\delta sign(\mu_S)$, where μ_S is the vector that equals μ over S and 0 outside of S. Notice that since the analytical sparsity of μ is bounded by $\sqrt{n}/4$, S will be non-empty. We will pick δ such that $\delta=o(\|\mu\|^2)/\|\mu\|_1$. It is easy to see that $\|\mu'\|^2\ge\|\mu\|^2$ and we also have that $\|\mu\|_1=\|\mu\|_1+\delta s=\frac{\|\mu\|_1}{\|\mu\|_2}(1+o(1))$. Also if σ is small enough then samples generated from $\mathcal{N}(\mu,\sigma^2 I)$ and from $\mathcal{N}(\mu',\sigma^2 I)$ will be δ-close to each other. Finally, notice that

$$\begin{aligned}\|\mu-\mu'\|&=\delta\sqrt{s}\\&=\Omega(\sqrt{\delta\|\mu\|_1}).\end{aligned}$$

■