Adversarially Robust Low Dimensional Representations

Pranjal Awasthi PRANJAL.AWASTHI@RUTGETS.EDU
Google Research and Rutgers University, New York

Vaggos Chatziafratis VAGGOS@CS.STANFORD.EDU
Google Research, New York

Xue Chen XUECHEN@GMU.EDU
George Mason University

Aravindan Vijayaraghavan ARAVINDV@NORTHWESTERN.EDU
Northwestern University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Many machine learning systems are vulnerable to small perturbations made to inputs either at test time or at training time. This has received much recent interest on the empirical front due to applications where reliability and security are critical. However, theoretical understanding of algorithms that are robust to adversarial perturbations is limited.

In this work we focus on Principal Component Analysis (PCA), a ubiquitous algorithmic primitive in machine learning. We formulate a natural robust variant of PCA where the goal is to find a low dimensional subspace to represent the given data with minimum projection error, that is in addition robust to small perturbations measured in ℓ_q norm (say $q = \infty$). Unlike PCA which is solvable in polynomial time, our formulation is computationally intractable to optimize as it captures a variant of the well-studied sparse PCA objective as a special case. We show the following results:

• Polynomial time algorithm that is constant factor competitive in the worst-case with respect to the best subspace, in terms of the projection error and the robustness criterion.

• We show that our algorithmic techniques can also be made robust to adversarial training-time perturbations, in addition to yielding representations that are robust to adversarial perturbations at test time. Specifically, we design algorithms for a strong notion of training-time perturbations, where every point is adversarially perturbed up to a specified amount.

• We illustrate the broad applicability of our algorithmic techniques in addressing robustness to adversarial perturbations, both at training time and test time. In particular, our adversarially robust PCA primitive leads to computationally efficient and robust algorithms for both unsupervised and supervised learning problems such as clustering and learning adversarially robust classifiers.

Keywords: adversarial robustness, PCA, training robustness, sparse PCA

1. Introduction

Reliability and trustworthiness of machine learning systems are key requirements for their secure adoption in day to day life. Many algorithms in machine learning are brittle to small perturbations made to the data points either at test time or at training time. While the design of robust machine learning algorithms has seen exciting recent developments in both statistics and computer science (Huber, 2011; Diakonikolas and Kane, 2019), our theoretical

understanding of robustness to adversarial perturbations is limited. This lack of robustness to adversarial perturbations poses significant practical hurdles (Szegedy et al., 2013; De Sa et al., 2017, 2018), and raises foundational questions of whether and how we can design basic machine learning primitives that are robust to adversarial perturbations.

In this work we study the above question in the context of *principal component analysis (PCA)*, that is the predominant tool for obtaining succinct data representations, and used as a preprocessing primitive in many machine learning pipelines. Given data in a high-dimensional space $\mathbb{R}^n$ represented by the columns of a matrix A, the goal in PCA is to find a subspace of dimension at most $r \leq n$ to represent the points, that minimizes the projection error (or reconstruction error) onto the subspace. This is formalized as follows where the matrix norm $\|\cdot\|$ is either the Frobenius norm or the spectral norm:

$$\min_{\Pi \in \mathcal{P}} \|\Pi^{\perp} A\|^2 = \min_{\Pi \in \mathcal{P}} \|A - \Pi A\|^2, \text{ where } P = \{\text{orthogonal projections of rank} \leq r\}. \quad (1)$$

The representation of each point $x \in \mathbb{R}^n$ corresponds to the projection Πx onto the r-dimensional subspace given by Π (one can also represent the point as an r-dimensional vector in terms of a basis for Π).

We propose a robust variant of PCA that corresponds to learning representations that are robust to adversarial perturbations to the data. We model an adversarial perturbation x' of a point x as one for which the ℓ_q norm of the difference is small, i.e., $\|x - x'\|_q \leq \delta$, for some fixed $\delta > 0$ and $q > 2$. It is instructive to keep in mind the case of $q = \infty$, that is of particular interest in emerging paradigms such as adversarial machine learning (Szegedy et al., 2013; Madry et al., 2017). A low dimensional subspace with an associated projection matrix Π is robust if $\|\Pi x - \Pi x'\|_2$ is small for any adversarial perturbation x' of x.

What data representations are adversarially robust? Given an r-dimensional subspace of $\mathbb{R}^n$ with projection matrix Π, the adversarial robustness of Π to δ-perturbations in the ℓ_q norm is precisely captured by

$$\sup_{x, x' : \|x - x'\|_q \leq \delta} \|\Pi(x - x')\|_2 = \delta \|\Pi\|_{q \to 2}. \quad (2)$$

The quantity $\kappa = \|\Pi\|_{q \to 2}$ characterizes the robustness of the projection Π to perturbations in ℓ_q norm around *every* point $x \in \mathbb{R}^n$ in the following sense. The distance between the projections of x and a δ-perturbation x' of x (in ℓ_q norm) is upper bounded by $\kappa\delta$. On the other hand, around each point x one can also realize a perturbation $x' = x + z$ with $\|z\|_q \leq \delta$ such that $\|\Pi x - \Pi x'\|_2 = \kappa\delta$. We will call Π a (κ, q)-robust rank-r projection when Π is an orthogonal projection matrix of rank at most r with $\|\Pi\|_{q \to 2} \leq \kappa$; when the robustness parameter κ and norm q are understood, we will just call it a robust rank-r projection.

This leads to the following natural formulation. Given a data matrix $A \in \mathbb{R}^{n \times m}$ composed of m points in $\mathbb{R}^n$, a robustness parameter $\kappa \geq 1$ and the norm $q \in [2, \infty]$, find a robust rank-r projection with low error:

$$\min_{\Pi} \quad \|\Pi^{\perp} A\|^2 = \min_{\Pi} \|A - \Pi A\|^2 \quad (3)$$

$$\text{s.t. } \Pi \text{ is an (orthogonal) projection matrix of rank at most } r, \text{ and } \|\Pi\|_{q \to 2} \leq \kappa. \quad (4)$$

One can also switch the objective and the constraint to consider the alternate formulation where we want to find a projection matrix with the minimum $\|\Pi\|_{q \to 2}$ (i.e., the most robust

projection), that achieves a prescribed projection error. We will be interested in two versions of the problem, depending on whether we measure the projection error in *Frobenius norm* or *spectral norm*. Recall that the top-r terms of the Singular Value Decomposition of A simultaneously solve both of these problems in polynomial time, when there is no additional robustness constraint, or when $q = 2$ (since $\|\Pi\|_{2\to 2} = 1$ for any non-trivial projection Π). We also remark that just as for the PCA objective (1), the above objective (3) can be equivalently rephrased as finding the best approximation among low-rank matrices, but among those with a "robust column space" (see Claim 9).

Training-time robustness. Our formulation in (3) finds robust representations assuming access to the uncorrupted training dataset denoted by the matrix A. However in practice, large scale datasets often contain various kinds of measurement errors (Sloutsky et al., 2013), or even data that is poisoned by adversarial perturbations. Hence, it is important to design algorithms that are robust to such training-time perturbations as well.

To capture training-time perturbations, we extend our formulation in (3) by assuming that we only have access to a corrupted dataset $\tilde{A}$, whose ith column $\tilde{A}_i$ is an adversarial perturbation of the corresponding column A_i of the uncorrupted dataset A, i.e., $\|\tilde{A}_i - A_i\|_q \leq \delta$. Given as input $\tilde{A}$, our goal is to output a robust projection $\widehat{\Pi}$ that achieves *near optimal* error for the true dataset A, i.e., $\|\widehat{\Pi}^{\perp}A\|^2 \approx \min_{\Pi} \|\Pi^{\perp}A\|^2$.

We will show how to design algorithms for finding robust representations that are robust to adversarial perturbations at *training-time*. In other words, we achieve robustness to both test-time and training-time perturbations simultaneously. As we will see in Section 2.1.2, the resilience to adversarial perturbations in the training set will crucially depend on the the $q \to 2$ operator norm of the projection matrix associated with the minimizer of (3).

Problem motivation. Studying robust variants of PCA can lead to new robust primitives for problems in data analysis and machine learning. (See Section 2.2 for specific examples.) Our work is also motivated by emerging paradigms such as *adversarial machine learning* and *low precision machine learning*. The recent phenomenon of *adversarial robustness* identified by Szegedy et al. (2013) shows that learning algorithms even when trained on high quality datasets are susceptible to small adversarial perturbations at test time. Even though empirical approaches have been proposed (Madry et al., 2017; Zhang et al., 2019) for designing algorithms that are robust to such perturbations, the current theoretical understanding is limited. Moreover in low-precision machine learning, one can achieve substantial performance improvements by quantizing the data to a few most significant bits (e.g., 8-bit arithmetic); this quantization noise is naturally captured as a small perturbation (in ℓ_∞ norm) to each training data point (De Sa et al., 2017, 2018).

Practical implications. Surprisingly, our techniques for learning robust linear representations also lead to algorithms for making deep neural networks that are highly non-linear in nature, more robust to test-time perturbations. In a very recent work, Awasthi et al. (2020b) directly build on the theoretical insights developed in this work to design a practical algorithm for making deep neural networks more robust to adversarial perturbations as compared to the state-of-the-art.

Connection to Sparse PCA and generalizations. While our formulation in (3) that is motivated by robustness is new to the best of our knowledge, it has rich connections to

(and implications for) well studied problems like the sparse PCA problem (Zou et al., 2006; Johnstone and Lu, 2009). Consider the setting when the perturbations are measured in ℓ_∞ norm and rank $r = 1$. The robustness constraint on the projection $\Pi = vv^\top$ imposes an upper bound of κ on the "analytic sparsity" of $v \in \mathbb{R}^n$ (measured as the ratio of ℓ_1 and ℓ_2 norms). In the special case of $r = 1$ the formulation is

$$\min \|A - vv^\top A\|_F^2 = \mathrm{tr}(AA^\top) - \max v^\top AA^\top v \;\text{ subject to } \|v\|_1 \le \kappa, \;\text{ and } \|v\|_2 = 1. \quad (5)$$

The complementary objective (i.e., $\max v^\top AA^\top v$) is the ℓ_1 version of the maximization SPARSE PCA objective;[1] both the ℓ_0 and the ℓ_1 versions are notoriously hard in the worst-case (Chan et al., 2016) (see also Theorem 54 in Appendix H.3). For general $q \ge 2$, requiring robustness places a constraint on the dual ℓ_{q^*} norm of the direction v. Moreover for projection matrices of higher rank $r \ge 1$, $\|\Pi\|_{q\to 2}$ is a basis-independent quantity that captures the maximum ℓ_{q^*} norm over all directions (unit vectors in ℓ_2) in the subspace given by Π (see Lemma 7). Hence robust projection matrices correspond to subspaces comprised of analytically sparse vectors measured in an appropriate norm e.g., ℓ_1 norm, when $q = \infty$ (see also Claim 8 for an approximate converse in terms of the sparsity of a basis for Π). Appendix H.2 gives some examples of what robust projection matrices i.e., matrices with small $q \to 2$ operator norm, look like.

The range of values of the robustness parameter κ is $1 \le \|\Pi\|_{q\to 2} \le n^{1/2-1/q}$. For several real world datasets we expect κ to be significantly smaller than the upper bound. As an example Figure 1 shows that most of the signal in images from the CIFAR-10 dataset can be captured by a robust subspace with $\infty \to 2$ norm that is significantly smaller than $\sqrt{n}$. The smaller the value of κ, the more robust the subspace is (it will be instructive to think of $\kappa \approx n^\varepsilon$, for some small constant $\varepsilon = 0.01$).

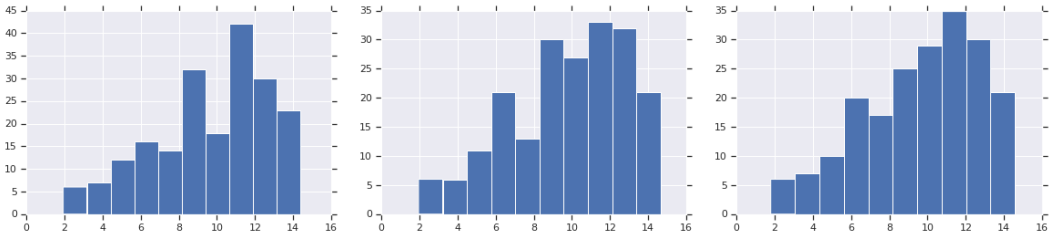

Figure 1: We study the CIFAR-10 image dataset (Krizhevsky et al., 2009) transformed in the *discrete cosine* (DCT) basis. We project each channel, of dimensionality $n = 1024$, onto a robust 200-dimensional subspace. The figure shows the histogram of the ℓ_1/ℓ_2 sparsity of the corresponding basis vectors. Each of the three projection matrices capture more than 99% of the signal in the respective channel. Furthermore, for each projection matrix, the $\|\cdot\|_{\infty\to 2}$ is in $[19, 21]$, significantly smaller than $\sqrt{n} = 32$.

1. It is within a factor 2 of the ℓ_0 version where the constraint $\|v\|_1 \le \kappa$ is replaced by $\|v\|_0 \le \kappa^2$ (see Section 10.3.3 of Vershynin (2018)).

2. Our Results

While all the results that follow apply to $q \geq 2$, it will be useful to think of $q = \infty$ (perturbations measured in ℓ_∞ norm), and the robustness parameter $\kappa \ll n^{1/2}$, say $\kappa < n^{0.1}$. Intuitively, the larger the choice of q, the more unrestricted the adversarial perturbations can be (since $\|x\|_p \leq \|x\|_q$ when $p \geq q$).

2.1. Algorithmic Guarantees for Robust Low-Rank Projections

2.1.1. Approximation Algorithms for Adversarially Robust PCA

We first consider the two variants of problem (3), where the matrix norm $|||\cdot|||$ represents either the Frobenius norm or spectral norm, in the worst-case setting.

(Informal) Theorem 1 *There exist polynomial time algorithms that given $q \geq 2$, any $\gamma \in (0, 1]$ and a data matrix A with a (κ, q)-robust projection matrix Π^* of rank at most r satisfying $|||A - \Pi^* A|||^2 \leq \varepsilon |||A|||^2$ for some $\varepsilon \in [0, 1]$, find a projection matrix $\widehat{\Pi}$ of rank at most r s.t.*

$$\|\widehat{\Pi}\|_{q\to 2} \leq O(1/\sqrt{\gamma}) \cdot \kappa, \text{ and } |||A - \widehat{\Pi} A|||^2 \leq (\alpha + \gamma) \cdot \varepsilon |||A|||^2, \tag{6}$$

where $\alpha = 2$ for the Frobenius norm error objective and $\alpha = 3$ for the spectral norm error objective. Moreover, for any $\gamma \in (0, 1]$, there exist polynomial time algorithms that find an $r' \leq r(1 + O(\gamma^{-1}))$-dimensional projection $\widehat{\Pi}$ that gets a projection error of $(1 + \gamma)|||A - \Pi^* A|||^2$, and relaxes the robustness parameter by $O(1/\sqrt{\gamma})$ factor.

In other words, our algorithms attain small constant factor bicriteria approximation to the adversarially robust PCA problem. The algorithms for both objectives – Frobenius norm and spectral norm, use convex relaxations and similar ideas, yet the algorithms (and relaxations) are different, unlike the case for standard PCA. Please see Theorem 11 (Frobenius norm objective) and Theorem 18 (spectral norm objective) for the formal statements. Our algorithms take in as input a guess for the robustness parameter κ. Recall that $\kappa \in [1, n^{1/2-1/q}]$.[2] Alternately, one can also input a guess for the optimal projection error (or the desired projection error), and minimize the robustness parameter κ approximately.

Observe that the approximation guarantee in Theorem 1 is a constant independent of the desired rank r. Even if we do not restrict the rank r (set $r = n$) our algorithm finds among all subspaces that are $O(\kappa)$-robust, the one with approximately optimal error. The constant factor loss in the robustness parameter depends on the value of $q \in [2, \infty)$. It is the largest for $q = \infty$ (where it is $\sqrt{\pi/2}$), and this is related to a variant of the Grothendieck problem (Alon and Naor, 2004; Nesterov, 1998). This loss in the robustness parameter is unavoidable when $q > 2$, due to the inapproximability for certifying the $q \to 2$ norm, even for projection matrices (Bhattiprolu et al., 2018b) (see Section A.2).

Our result also has new implications for approximating the minimization objective for sparse PCA specified in (5). Most existing theoretical guarantees for Sparse PCA have been established for average case models (Berthet and Rigollet, 2013). There has also been

2. When $q = 2$, the robustness constraint becomes trivial, and problem reduces to the standard PCA problem as discussed earlier.

work on studying the maximization version of the sparse PCA objective in worst case models (Chan et al., 2016). To the best of our knowledge, we are not aware of any existing worst case guarantees for the minimization version as defined in (5). Our results (applied with $r = 1$) provides a small constant factor bicriteria approximation to problem (5). This is in stark contrast to the approximability of the maximization version of the problem. Even when $r = 1$ the best known polynomial time algorithm gives a $O(n^{1/3})$ factor approximation in the worst-case (for both the ℓ_1 and ℓ_0 versions); moreover no constant factor approximation is possible assuming the SSE conjecture (Chan et al., 2016). (see its implication to computational hardness of our minimization version (3) in Appendix H.3). Furthermore, the minimization variant of the problem that we study (and our small approximation factors) will be crucial in various downstream applications such as clustering.

2.1.2. Robustness to Adversarial Errors during Training

We now discuss how to handle *data poisoning*, where points in the *training* data set $\widetilde{A}$ are adversarially perturbed. Recall that in the corruption model, *every* sample $A_i \in \mathbb{R}^n$ can potentially be adversarially perturbed up to a δ amount, as measured in ℓ_q norm for $q \geq 2$. So every column of $\widetilde{A}$ satisfies $\|\tilde{A}_i - A_i\|_q \leq \delta$; and we will refer to such an $\tilde{A}$ as a δ-*corrupted instance* of A. While the input instance is $\widetilde{A}$, our goal now is to recover a robust low-rank projection for the uncorrupted matrix A. We will show that we can in fact output a robust low-rank projection $\widehat{\Pi}$ that is competitive with the best robust low-rank projection of A, even though A is not known to us! We first state our result when the error is measured in Frobenius norm, and later describe the guarantees for the spectral norm variant.

(Informal) Theorem 2 *Suppose $q \geq 2$ and $A \in \mathbb{R}^{n\times m}$ is the unknown uncorrupted data matrix, with a (κ, q)-robust projection matrix Π^* of rank at most r satisfying $\|A - \Pi^* A\|_F^2 \leq \varepsilon \|A\|_F^2$ for some $\varepsilon \in [0, 1]$. There exists a polynomial time algorithm that given as input a δ-corrupted instance $\tilde{A}$ of A outputs a projection $\widehat{\Pi}$ of rank at most r that is approximately optimal:*

$$\forall \eta > 0,\ \|\widehat{\Pi}\|_{q\to 2} \leq O(\kappa),\ \text{and}\ \|A - \widehat{\Pi} A\|_F^2 \leq O(\varepsilon + \eta) \cdot \|A\|_F^2 + O(\tfrac{1}{\eta}) \cdot \delta^2 \kappa^2 m. \tag{7}$$

In particular this gives an $O(1)$ approximation when $\delta^2 < (\varepsilon^2/\kappa^2) \cdot \frac{1}{m}\|A\|_F^2$.

To interpret the results let $q = \infty$ and consider an uncorrupted dataset A where every column (sample) is a unit vector in $\mathbb{R}^n$, and let $\kappa = n^{0.1}$. The total corruption to each point is at most $o(1)$ in Euclidean norm when $\delta = o(n^{-1/2})$; in this case one would expect that standard PCA applied to $\tilde{A}$ may recover a good solution. The above Theorem 2 on the other hand guarantees to find a good (robust) low-rank approximation for the unknown matrix A even when $\delta = o(1/\kappa) = o(n^{-0.1})$. Note that in this setting *every* point can be completely overwhelmed by the adversarial noise (in Euclidean norm). The algorithm first denoises the input by solving a convex minimization problem before applying the algorithm from Theorem 1. Furthermore, Proposition 28 shows that the additive factor of $O(m\delta^2\kappa^2)$ is unavoidable for every $\kappa, \delta = O(1/\kappa)$. These results suggest that the robust projection structure (measured in $q \to 2$ operator norm) is key in understanding the resilience to small adversarial perturbations of every point during training, even without any test-time robustness considerations. Subsequent work by Awasthi et al. (2020a) also characterize principal

subspace recovery in an average-case setting in the presence of adversarial perturbations at training time using the $q \to 2$ norm robustness criterion (in an instance-optimal sense).

Our guarantees for spectral norm error in the presence of training-time adversarial perturbations are somewhat similar to Theorem B.1. However, there is a qualitative difference: given as input an adversarial δ-perturbation $\tilde{A}$ of an uncorrupted matrix A that has a good solution (i.e., $\|A - \Pi^* A\| \leq \varepsilon \|A\|$ for some small $\varepsilon \in (0, 1)$), we will either find a robust low-dimensional projection of the unknown dataset A, or we will certify that the dataset has been poisoned substantially (i.e., $\|\tilde{A} - A\| > \varepsilon \|A\|$). In particular, the algorithm will never output a low-dimensional representation that is bad for the unknown data matrix A. Please see Theorem 25 for a formal statement. We remark that information-theoretically we can design an estimator (that is computationally inefficient) that achieves the stronger qualitative guarantees as in Theorem 2. Designing a computationally efficient algorithm to do the same is an open question that we describe in more detail in Section 2.3.2.

2.2. Applications to Learning Problems.

The algorithmic results that we have described so far, may be used as a robust primitive in lieu of standard PCA. These lead to efficient, adversarially robust algorithms for learning problems of different flavors, further validating our formulation. In particular, we demonstrate the versatility of our robust primitive via the following three applications across both unsupervised and supervised learning:

1. Clustering with training-time perturbations. We study the classical unsupervised learning problem of k-means clustering. Let $A \in \mathbb{R}^{n \times m}$ denote m data points with an unknown ground truth clustering into k clusters. It is well known that if the ground truth clusters are well separated then the popular Lloyd's algorithm (Lloyd, 1982), when properly initialized, recovers the ground truth clustering (Kumar and Kannan, 2010; Awasthi and Sheffet, 2012). We extend this setting to consider a scenario where the input to the algorithm is the data matrix $\tilde{A}$ with each data point being adversarially corrupted up to a perturbation of a certain amount. Existing algorithms based on variants of the Lloyd's heuristic fail to handle large amounts of noise in this setting. This is due to the fact that these algorithms use PCA to initialize the cluster centers, and as we saw in previous sections, PCA is not robust to adversarial perturbations.

We instead design a robust variant of the Lloyd's heuristic that can handle a large amount of perturbation while successfully clustering the data according to the ground truth. In our algorithm, the adversarially robust PCA primitive plays a crucial role. We use the adversarially robust PCA primitive to obtain a good set of initial cluster centers. Additionally, during the iterative Lloyd's updates, we compute new cluster means via a new robust mean estimation procedure that we design in this work (this is the special case of clustering with $k = 1$). As a result we obtain a clustering algorithm that can handle adversarial perturbations to the training set, of magnitude up to $o(1/\kappa)$ where κ is the robustness of the k-dimensional subspaces spanned by the cluster means. Hence our algorithm can handle significantly more noise when this subspace is robust, compared to standard approaches that break down unless the perturbation amount (in ℓ_∞ norm, say) is of the order of $o(1/\sqrt{n})$. On the other hand, such a dependence on ℓ_1 sparsity of the means is needed even in the case where $k = 1$ (i.e., mean estimation in the presence of adversarial perturbations).

See Section D for details, including Theorem 31 for the general case, and Theorem 33 for the specialization to clustering mixtures of Gaussians.

2. Learning intersection of halfspaces under training-time and test-time perturbations. We consider the problem of learning an intersection of k halfspaces over the Gaussian distribution on $\mathbb{R}^n$ in the presence of adversarial perturbations to the samples, both at testing-time and training-time. We will represent an intersection of halfspaces by a function $h : \mathbb{R}^n \to \{0, 1\}$ denoted by $h(x) = \prod_{i=1}^{k} \mathbf{1}(w_i^\top x \geq \theta_i)$, where $\forall i \in [k],\ \|w_i\|_2 = 1$ and $\theta_i \in \mathbb{R}$ and where $\mathbf{1}(\cdot)$ denotes the indicator function. Let $\mathcal{H}_k$ represent the hypothesis class of all intersections of at most k halfspaces. In the uncorrupted setting, the training points $x_1, \ldots, x_m \in \mathbb{R}^n$ are drawn i.i.d. from a Gaussian distribution, and their corresponding labels $y_i = h^*(x_i)$ for some $h^* \in \mathcal{H}_k$. A series of well-known results (Vempala, 2010b,a; Klivans et al., 2008) shows that when we are given access to uncorrupted training samples in $\mathbb{R}^n$ drawn from a Gaussian distribution, one can learn an intersection of halfspaces in the PAC learning model, in time $f(k) \cdot \mathrm{poly}(n)$. Crucially, these algorithms use PCA as a first step to reduce the learning problem to a low dimensional space. Our adversarially robust PCA primitive can be used to learn an intersection of $k = O(1)$ halfspaces even when there are adversarial perturbations *both* at *training-time* and *test-time*.

What does a classifier h, say $h(x) = \mathbf{1}(w_1^\top x \geq 0) \cdot \mathbf{1}(w_2^\top x \geq 0)$, that is robust to adversarial δ-perturbations at test-time look like? First observe that $\max\{\|w_1\|_1, \|w_2\|_1\} \leq O(1/\delta)$ is necessary, otherwise there exists a δ-adversarial perturbation $\tilde{x}$ with $\|\tilde{x} - x\|_\infty \leq \delta$ that h misclassifies w.h.p! Moreover the subspace Π^* spanned by w_1, w_2 is robust as measured in $\kappa = \|\Pi^*\|_{\infty \to 2}$ (see Claim 45 for a formal claim). For general k, we will assume that there exists a robust classifier $h^*(x) = \prod_{i=1}^{k} \mathbf{1}(w_i^\top x \geq \theta_i)$ for the data such that the projection matrix Π^* onto the span of the normals $w_1, \ldots, w_k$ satisfies $\|\Pi^*\|_{q \to 2} \leq \kappa$.

We consider a natural model of training-time perturbations, where each training datapoint is δ-adversarially perturbed in ℓ_q norm ($q \geq 2$). Our robust algorithm follows the same general approach as in Vempala (2010a); however we use our primitive for adversarially robust PCA instead of standard PCA to bring down the dimension to k. This allows us to handle adversarial perturbations of magnitude $\delta = o(1/\kappa)$ (as opposed to existing approaches that need $\delta = o(1/\sqrt{n})$ for $q = \infty$), and output a robust classifier (intersection of k-halfspaces) that incurs an error of $o(1)$. Recall from the earlier discussion, that such a condition is necessary qualitatively: even a single half-space $\mathbf{1}(w_1^\top x \geq 0)$ is not robust when $\|w_1\|_1 = \kappa$ and $\delta \gg 1/\kappa$. See Section E for details.

3. Trading off natural accuracy in classification for robustness to test-time perturbations. Finally, in many scenarios it might be desirable to trade off natural accuracy for significant robustness to test-time perturbations. We demonstrate how our robust primitive can be used for this purpose. Specifically, we consider the *Gaussian data model* (Anderson, 2003) that has been studied in recent works to understand adversarial robustness (Tsipras et al., 2018; Schmidt et al., 2018). In this model a labeled example (x, y) is generated by first picking the label as $+1$ or -1 with equal probability. Then $x \in \mathbb{R}^n$ is drawn from either $\mathcal{N}(\mu_1, \Sigma)$ or $\mathcal{N}(\mu_2, \Sigma)$ depending on whether $y = -1$ or $y = +1$. We denote this model as $\mathcal{M}(\mu_1, \mu_2, \Sigma)$.

In the above model the (Bayes) optimal classifier is a linear classifier of the form $sgn(\langle w, x\rangle + b)$ with weight vector $w = \Sigma^{-1}(\mu_1 - \mu_2)$. If the means are well separated

then the above classifier has good accuracy but can be easily fooled during test-time via small perturbations. In other words, the robust accuracy of the above classifier is close to zero. In order to get a better trade-off of standard accuracy and robust accuracy, we could instead aim to look for robust subspaces where the variance is low and the means when projected are still separated by a non-trivial amount. We will show how our robust PCA primitive helps us achieve this and obtain a classifier with better robust accuracy. See Section F for details.

2.3. Proof Sketches and Technical Overview

We give a flavor of the technical ideas involved in obtaining our main algorithmic results.

2.3.1. Constant Factor Approximation Algorithms

Let us first consider the version of problem (3) of finding a robust rank-r projection that has small error measured in Frobenius norm. A natural mathematical programming relaxation is the following:

$$\min_X \|A\|_F^2 - \langle AA^\top, X\rangle \tag{8}$$

$$\text{subject to} \quad \operatorname{tr}(X) \le r, \quad 0 \preceq X \preceq I \quad \text{and } \|X\|_{q\to 2} \le \kappa \tag{9}$$

This is a valid convex relaxation for the problem since the constraints are all satisfied by any rank-r projection matrix that is robust i.e, $\|\Pi\|_{q\to 2} \le \kappa$.

The first challenge however is that the operator norm constraint (9) is NP-hard to verify efficiently, even for the case of projection matrices. However, these operator norm $\|\cdot\|_{q\to p}$ computation problems form a rich class of problems related to the Grothendieck problem (Alon and Naor, 2004; Nesterov, 1998), and polynomial time $O(1)$ factor approximations are known for general $q \to 2$ norms with $q \ge 2$ (see Section A.2).

The bigger challenge is in producing a projection matrix from X^* that simultaneously (a) achieves a good objective value, (b) has rank at most r, and (c) is $O(\kappa)$-robust i.e., has bounded $q \to 2$ norm. A natural approach for producing a good low-rank solution is to output a rank-r projection matrix Π_r that corresponds to the large singular values of X^*. However we have no control on the robustness of the subspace $\|\Pi_r\|_{q\to 2}$. In fact, the algorithmic problem (3) is challenging even when there is no rank constraint ($r = n$). The main issue is to relate the $q \to 2$ operator norm of the projection matrix we output to that of the relaxation solution $\|X^*\|_{q\to 2}$ which is upper bounded by κ.

Our crucial insight is that we can indeed design a rounding scheme that achieves all three goals if the norm in the constraint (9) is a *monotone norm*!

$$\text{A matrix norm } |||\cdot||| \text{ is monotone iff} \quad \forall A, B \succeq 0, \quad \text{we have} \quad |||A + B||| \ge |||A|||.$$

(See Definition 5 for details.) This monotonicity property allows us to truncate terms in the eigendecomposition of X^* without any loss in robustness κ, and get fine control on the robustness κ when we rescale different rank-1 terms appropriately. Unfortunately however, the $q \to 2$ operator norm is not monotone in general (see e.g., Claim 51, Claim 50).

Our next important observation is that we can replace the constraint (9) by a similar constraint in terms of the $q \to q^*$ norm. This is because for any matrix B, we have that

$\|B\|_{q\to 2}^2 = \|B^\top B\|_{q\to q^*}$ where ℓ_{q^*} is the dual norm for ℓ_q and satisfies $\frac{1}{q^*}+\frac{1}{q}=1$. The main advantage of this reformulation is that the $q \to q^*$ operator norms *are* indeed *monotone*.

Claim 3 (Same as Claim 15) *For any $q \geq 1$, the operator norm $\|\cdot\|_{q\to q^*}$ is monotone.*

Moreover polynomial time $O(1)$-approximate separation oracles based on semidefinite programs exist for these norms when $q \geq 2$. This motivates convex programming relaxation **CP1** and its equivalent but more elegant convex relaxation **CP2** shown in Figure 2.

CP1 :	**CP2** :
$\min_X \|A\|_F^2 - \langle AA^\top, X\rangle$	$\min_{X\in\mathbb{R}^{n\times n}, d\in\mathbb{R}^n_{\geq 0}} \|A\|_F^2 - \langle AA^\top, X\rangle$
s.t. $\mathrm{tr}(X) \leq r, \;\; 0 \preceq X \preceq I$	s.t. $\mathrm{tr}(X) \leq r, \;\; 0 \preceq X \preceq I$
$\max_{Y\in\mathcal{Q}} \langle X, Y\rangle \leq C_G\kappa^2$, where	$X \preceq \mathrm{diag}(d)$
$\mathcal{Q} = \{Y \in \mathbb{R}^{n\times n} : \; Y \succeq 0, \sum_{i=1}^n Y_{ii}^{q/2} \leq 1\}$	$\|d\|_{q/(q-2)} := \Big(\sum_{i=1}^n d_i^{q/(q-2)}\Big)^{(q-2)/q} \leq C_G\kappa^2.$

Figure 2: Two equivalent tractable convex relaxations **CP1** and **CP2** for problem (13). See Lemma 16 for proof of equivalence using convex duality.

Let X^* be the optimal solution to the convex program. We obtain the required robust low-rank projection matrix from X^* by a simple rounding procedure that focuses on the large singular values of X^*. The monotonicity property of the norm leads to an elegant analysis to guarantee that the resulting low-rank projection is $O(\kappa)$-robust, while also achieving small error. We now sketch the proof of Theorem 1 with the Frobnenius norm objective. Similar ideas also work when the projection error is measured in terms of the spectral norm. However, the objective function is instead rephrased as $\min\|(A^\top(I-X)A\|$ where $\|\cdot\|$ is the spectral norm; the algorithm and analyis, while slightly different again leverage the monotonicity property of the $q \to q^*$ norms.

Proof Sketch of Theorem 1. Assume $\|A\|_F = 1$ without loss of generality, and let $OPT = \varepsilon \in [0,1]$. It is easy to see any feasible projection matrix Π of rank r satisfying $\|\Pi\|_{q\to 2} \leq \kappa$ forms feasible solutions to **CP1** and **CP2** (for an appropriate feasible d) with the correct objective value. Moreover the relaxations **CP1** and **CP2** can be solved in polynomial time to arbitrary accuracy using the Ellipsoid algorithm. See Claim 17 for details.

Set $\delta := 1/(1+\gamma)$. Let $\widehat{X} = \sum_i \lambda_i v_i v_i^\top$ and $S = \{i : \lambda_i \geq 1-\delta\}$. Define for each $i \in [S]$, $\alpha_i := \langle v_i v_i^\top, AA^\top\rangle$. We form T from S by picking the $\min\{r, |S|\}$ ones with the largest $\{\alpha_i\}$ values. Let $\Pi_S = \sum_{i\in S} v_i v_i^\top$. Our projection matrix will be $\Pi_T = \sum_{i\in T} v_i v_i^\top$.

We use monotonicity of $\|\cdot\|_{q\to q^*}$ to show the operator norm constraint is satisfied:

$$\|\Pi_T\|_{q\to q^*} = \|\sum_{i\in T} v_i v_i^\top\|_{q\to q^*} \leq \frac{1}{1-\delta}\|\sum_{i:\lambda_i>1-\delta} \lambda_i v_i v_i^\top\|_{q\to q^*} \leq \frac{\|\widehat{X}\|_{q\to q^*}}{1-\delta} \leq \frac{\alpha\kappa}{1-\delta} = \frac{\alpha(1+\gamma)\kappa}{\gamma}$$

Also, since Π_S (and hence Π_T) projects onto the large eigenspace of $\widehat{X}$, we can prove that by truncating onto the large eigenvalues (see Lemma 14)

$$\langle I - \Pi_S, AA^\top \rangle = \sum_{i \notin S} \langle v_i v_i^\top, AA^\top \rangle \leq \frac{\varepsilon}{\delta}, \text{ and}$$
$$\sum_{i \in S} \langle (1-\lambda_i) v_i v_i^\top, AA^\top \rangle \leq \sum_i \langle (1-\lambda_i) v_i v_i^\top, AA^\top \rangle \leq 1 - \langle X, AA^\top \rangle \leq \varepsilon.$$

$$\text{Hence, } \sum_{i \in S} \lambda_i \alpha_i = \sum_{i \in S} \lambda_i \langle v_i v_i^\top, AA^\top \rangle \geq 1 - \varepsilon\Big(1 + \frac{1}{\delta}\Big) = 1 - (2+\gamma)\varepsilon,$$

for our choice of $\delta = 1/(1+\gamma)$. By our greedy choice of T, we have $\sum_{i \in T} \alpha_i \geq \sum_{i \in S} \lambda_i \alpha_i$, as $\sum_{i \in S} \lambda_i \leq \min\{\mathrm{tr}(X), |S|\} = |T|$, with each $\lambda_i \in [0,1]$. Thus $\|\Pi_T^\perp A\|_F^2 \leq (2+\gamma)\varepsilon$. This completes the proof. For the bicriteria guarantee with rank $r/(1-\delta)$ we output Π_S. The objective and $\|\cdot\|_{q \to q^*}$ bounds follow using similar arguments.

■

2.3.2. Technical Overview for Training-Time Adversarial Perturbations

Let $q = \infty$. Recall that our input instance $\tilde{A}$ is a δ-corrupted instance obtained from A by potentially corrupting every entry of it by a δ amount. Our goal is to output a robust low-rank projection matrix Π of rank at most r for the uncorrupted matrix A, that is not known to us. This question is interesting even from a purely statistical standpoint; but additionally, we would also like our algorithm to run in polynomial time.

Why should this be possible? Suppose the uncorrupted matrix A has a robust low-rank projection Π^* of small error i.e., $\|A - \Pi^* A\| < \varepsilon \|A\|$ (where $\|A\|$ is either the Frobenius norm or spectral norm). Also assume for just this discussion that the average column (Euclidean) length of A is 1, $\kappa = n^{0.1}$ say and $\delta = o(n^{-0.1})$. For any κ-robust projection Π, $\Pi A_j \approx \Pi \tilde{A}_j$ for each data point $j \in [m]$. So one could apply the worst-case algorithm on the corrupted input $\tilde{A}$, and hope to also get a robust projection of low-error for the unknown matrix A.

However, there are two major challenges in implementing this strategy. *(1) Solution value of $\tilde{A}$:* the robust projection Π^* may not achieve low error on $\tilde{A}$; in fact, $\tilde{A}$ may not have any good robust low-rank approximation – in this case the algorithm output may be useless. This is because the entry-wise perturbations could make A and $\tilde{A}$ far away in aggregate e.g., $\|A - \tilde{A}\|_F$ could be $\delta\sqrt{nm} \gg \sqrt{m} \approx \|A\|_F$.
(2) Identifiability issue: perhaps more importantly, even if the perturbation $\tilde{A}$ has a robust low-rank projection of small error, we need to argue that this subspace indeed attains small error on A! The second issue is crucial in resolving the purely information-theoretic aspect of the question; it involves ruling out the scenario where $\tilde{A}$ has good robust low-rank approximation that is very different from any robust low-rank approximation for A.

To address the second issue (identifiability), we prove that if the projection $\widehat{\Pi}$ gives a small error on $\tilde{A}$, it necessarily gives a low-error on A. Roughly speaking, if there are two data-matrices A and B with $\|A - B\|_\infty \leq \delta$, then for $\gamma \in (0,1)$

$$\|A - \Pi_1 A\|, \|B - \Pi_2 B\| < \gamma \|A\| \quad \Longrightarrow \quad \|A - \Pi_2 A\| \leq \gamma_1 \|A\| + \tfrac{1}{\gamma_2}\sqrt{m}\delta\kappa, \text{ (and similarly for } B),$$

where $\gamma_1 = \gamma_1(\gamma), \gamma_2 = \gamma_2(\gamma) \in (0,1)$. One can show that $\|\Pi_1 A - \Pi_1 B\|$ and $\|\Pi_2 A - \Pi_2 B\|$ are small since Π_1, Π_2 are robust (see Lemma 7); however this does not give a handle on $\|A - \Pi_2 A\|$. Note that the above statement does not follow from an application of the triangle inequality since we do not have any prior control on $\Pi_1 - \Pi_2$. This statement is particularly tricky to show for the spectral norm. A natural approach is to argue that $\Pi_1 A$ and $\Pi_2 B$ are close by arguing about their actions on any unit vector. We use a somewhat indirect proof; we show that for every direction $v \in \mathbb{S}^{n-1}$, (1) the lengths $\|Av\|_2$ and $\|Bv\|_2$ are similar and (2) the difference in the lengths $|\|Av\|_2 - \|Bv\|_2|$ is (approximately) lower bounded by $\|(A - \Pi_2 A)v\|_2$. This will allow us to conclude that $\|A - \Pi_2 A\|$ is small.

To tackle the first issue (solution value), we first preprocess (denoise) to find an alternate matrix A' with a good solution value. Suppose we have an algorithm to find

$$A' = \operatorname*{argmin}_{B:\|B-\tilde{A}\|_\infty \le \delta} \quad \min_{\Pi:\mathrm{rank}(\Pi)=r,\ \|\Pi\|_{\infty\to 2}\le\kappa} \|B - \Pi B\|^2. \tag{10}$$

We know that the uncorrupted matrix A is a feasible solution with good value. Hence the optimal solution A' of (10) has an even better solution. Moreover $\|A - A'\|_\infty \le 2\delta$. This reduces the first issue to a computational question of solving (10). For Frobenius norm error, we can obtain a good A' by instead solving a simple convex optimization problem.

For the spectral norm problem, we do not know of an efficient algorithm for (10). However by running our worst-case algorithm (for spectral norm error) on $\tilde{A}$, we will either find a good solution that also works for A, or we will certify that $\|A - \tilde{A}\|$ is too large i.e., the data was poisoned significantly. Finally we remark that we get the stronger computationally efficient guarantee for the spectral norm error (as for the Frobenius norm error) if we can resolve the spectral norm variant of (10), which is an open question.

2.4. Related and Concurrent Work on Training-time corruptions.

Subsequent work by Awasthi et al. (2020a), studies training time robustness in an average-case setting namely, the spiked covariance model where the goal is to recover the top principle subspace of the data distribution. They extend the algorithms developed in this work (Section B) to the average case setting, and in fact show that the $q \to 2$ operator norm of the principal subspace almost characterizes its robustness to adversarial perturbations at training time in the spiked covariance model. Very recently, d'Orsi et al. (2020) studies the problem of recovering an ℓ_0-sparse[3] principal component where there are adversarial perturbations in ℓ_∞ norm to the training data points, again focusing on the spiked covariance model. In contrast, our work studies the worst case formulation of the problem.

Comparison to the Huber contamination model and the robust PCA problem. There is a vast amount of literature in designing robust algorithms in a different model, the Huber's contamination model, where, unlike our setting, a small fraction of the data can be arbitrarily corrupted (Huber, 2011; Diakonikolas et al., 2018a; Lai et al., 2016; Diakonikolas and Kane, 2019). Our notion of training-time adversarial perturbations is very different in flavor – it involves bounded adversarial perturbations to potentially every training point. Another popular model is the robust PCA problem proposed in Candès et al. (2011). It assumes that a given corrupted matrix $\tilde{A}$ is a sum of two matrices, the true matrix A that is

3. Note that any ℓ_0 sparse unit vector is also ℓ_1 sparse; see Claim 6.

low-rank and a sparse corruption matrix S with sparsity pattern being essentially random. The corruptions although sparse can be unbounded in magnitude. This setting is again fundamentally different from ours. Recovery in this model necessitates incoherence type structural assumptions that the principal components of A are spread out, whereas in our setting *sparsity or localization* of the signal dictates the recovery error.

Please see Section G for more details, and comparison to other related work.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback. AV was supported by the National Science Foundation (NSF) under Grant No. CCF-1652491, CCF-1637585 and CCF 1934931.

References

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

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.

Noga Alon, Troy Lee, Adi Shraibman, and Santosh Vempala. The approximate rank of a matrix and its algorithmic applications: Approximate rank. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 675–684, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2029-0. doi: 10.1145/2488608.2488694. URL http://doi.acm.org/10.1145/2488608.2488694.

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

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 3 edition, 2003.

Sanjeev Arora and Ravi Kannan. Learning mixtures of separated nonspherical gaussians. *The Annals of Applied Probability*, 15(1A):69–92, 2005.

Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer, 2012.

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, 2019.

Pranjal Awasthi, Xue Chen, and Aravindan Vijayaraghavan. Estimating principal components under adversarial perturbations. In *Conference on Learning Theory, COLT*, pages 1 – 40, 2020a.

Pranjal Awasthi, Himanshu Jain, Ankit Singh Rawat, and Aravindan Vijayaraghavan. Adversarial robustness via robust low rank representations. *Neural Information Processing Systems (NeurIPS)*, 2020b.

Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45 (1):77–120, 2017.

Boaz Barak, Fernando GSL Brandao, Aram W Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 307–326. ACM, 2012.

Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 103–112. IEEE, 2010.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

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

Aditya Bhaskara and Aravindan Vijayaraghavan. Approximating matrix p-norms. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 497–511. Society for Industrial and Applied Mathematics, 2011.

Vijay Bhattiprolu, Mrinalkanti Ghosh, Venkatesan Guruswami, Euiwoong Lee, and Madhur Tulsiani. Approximating operator norms via generalized krivine rounding. *arXiv preprint arXiv:1804.03644*, 2018a.

Vijay Bhattiprolu, Mrinalkanti Ghosh, Venkatesan Guruswami, Euiwoong Lee, and Madhur Tulsiani. Inapproximability of matrix p to q norms. *arXiv preprint arXiv:1802.07425*, 2018b.

Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, October 1989.

Christos Boutsidis and David P Woodruff. Optimal cur matrix decompositions. *SIAM Journal on Computing*, 46(2):543–589, 2017.

Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 968–977. SIAM, 2009.

Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

S Charles Brubaker. Robust pca and clustering in noisy mixtures. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 1078–1087. SIAM, 2009.

S Charles Brubaker and Santosh S Vempala. Isotropic pca and affine-invariant clustering. In *Building Bridges*, pages 241–281. Springer, 2008.

Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from cryptographic pseudo-random generators. *arXiv preprint arXiv:1811.06418*, 2018a.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018b.

T Tony Cai, Zongming Ma, Yihong Wu, et al. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

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

Siu On Chan, Dimitris Papailliopoulos, and Aviad Rubinstein. On the approximability of sparse pca. In *Conference on Learning Theory*, pages 623–646, 2016.

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.

Agniva Chowdhury, Petros Drineas, David P. Woodruff, and Samson Zhou. Approximation algorithms for sparse principal component analysis, 2020.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of evasion adversaries. *arXiv preprint arXiv:1806.01471*, 2018.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

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.

Akshay Degwekar, Preetum Nakkiran, and Vinod Vaikuntanathan. Computational limitations in robust classification and win-win results. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 994–1028, 2019.

Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 329–338. IEEE, 2010.

Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics, 2019.

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, 2018a.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060. ACM, 2018b.

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.

Edgar Dobriban, William Leeb, Amit Singer, et al. Optimal prediction in the linearly transformed spiked model. *The Annals of Statistics*, 48(1):491–513, 2020.

Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

Tomasso d'Orsi, Pravesh K. Kothari, Gleb Novikov, and David Steurer. Sparse PCA: algorithms, adversarial perturbations and certificates. In *FOCS*, 2020.

Shivam Garg, Vatsal Sharan, Brian Zhang, and Gregory Valiant. A spectral view of adversarially robust features. In *Advances in Neural Information Processing Systems*, pages 10138–10148, 2018.

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.

Samuel B. Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pages 1021–1034, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5559-9. doi: 10.1145/3188745.3188748. URL http://doi.acm.org/10.1145/3188745.3188748.

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

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486): 682–693, 2009.

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.

Ravindran Kannan and Santosh Vempala. Randomized algorithms in numerical linear algebra. *Acta Numerica*, 26:95–135, 2017.

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

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

Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via gaussian surface area. In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008*, pages 541–550. IEEE Computer Society, 2008.

Pravesh K Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 2018.

Pravesh K Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE, 2010.

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.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

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

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.

Yifei Min, Lin Chen, and Amin Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. *arXiv preprint arXiv:2002.11080*, 2020.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*, 2019.

Boaz Nadler et al. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817, 2008.

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.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 755–764. ACM, 2010.

Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017.

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.

Dan Shen, Haipeng Shen, and James Stephen Marron. Consistency of sparse pca in high dimension, low sample size contexts. *Journal of Multivariate Analysis*, 115:317–333, 2013.

Roman Sloutsky, Nicolas Jimenez, S Joshua Swamidass, and Kristen M Naegle. Accounting for noise when clustering biological data. *Briefings in bioinformatics*, 14(4):423–436, 2013.

Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise l 1-norm error. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 688–701. ACM, 2017.

Suvrit Sra. Fast projections onto mixed-norm balls with applications. *Data Mining and Knowledge Discovery*, 25(2):358–377, 2012.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

Leslie G Valiant. Learning disjunction of conjunctions. In *IJCAI*, pages 560–566. Citeseer, 1985.

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

Santosh S. Vempala. Learning convex concepts from gaussian distributions with PCA. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010*, pages 124–130. IEEE Computer Society, 2010a.

Santosh S. Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *J. ACM*, 57(6), November 2010b. ISSN 0004-5411. doi: 10.1145/1857914.1857916. URL https://doi.org/10.1145/1857914.1857916.

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

Aravindan Vijayaraghavan, Abhratanu Dutta, and Alex Wang. Clustering stable instances of euclidean k-means. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6500–6509. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7228-clustering-stable-instances-of-euclidean-k-means.pdf.

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.

Henry Wolkowicz, Romesh Saigal, and Lieven Vandenberghe. *Handbook of semidefinite programming: theory, algorithms, and applications*, volume 27. Springer Science & Business Media, 2012.

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

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

Contents of Appendix

Appendix A. Notation and Preliminaries

Norms. For every $q \geq 1$ and $x \in \mathbb{R}^n$, we will use $\|x\|_q$ to denote the ℓ_q norm of the vector x i.e., $\|x\|_q^q = \sum_{i\in[n]} |x_i|^q$. The dual norm of ℓ_q is ℓ_{q^*} where $1/q^* + 1/q = 1$. We will heavily use Holder's inequality which states that

$$\text{(Hölder's inequality)} \quad |\langle u, v\rangle| \leq \|u\|_q \cdot \|v\|_{q^*} \quad \forall u, v \in \mathbb{R}^n. \tag{11}$$

When not specified, $\|x\|$ will denote the Euclidean norm of x. Further $\mathbb{S}^{n-1}$ will represent the unit sphere for the Euclidean norm. For convenience, we will use $\|x\|_0$ to denote the sparsity i.e., the size of the support of x (note that ℓ_0 is not a valid norm on vectors).

Operator Norms of Matrices. We will use the following matrix norms. For any $q, p \geq 1$ and any matrix $M \in \mathbb{R}^{n\times m}$, we will denote by $\|M\|_{q\to p} = \max_{y\in\mathbb{R}^m, \|y\|_q\leq 1} \|My\|_p$. By duality of vector norms, we have

$$\|M\|_{q\to p} = \max_{y\in\mathbb{R}^m, \|y\|_q\leq 1} \max_{z\in\mathbb{R}^n, \|z\|_{p^*}\leq 1} z^\top My = \max_{z\in\mathbb{R}^n, \|z\|_{p^*}\leq 1} \max_{y\in\mathbb{R}^m, \|y\|_q\leq 1} y^\top M^\top z = \|M^\top\|_{p^*\to q^*}.$$

When $p = q = 2$, this corresponds to the spectral norm of the matrix M i.e., the maximum singular value of M. When unspecified, we will use $\|M\|$ to denote the spectral norm of M. (Note that the above equalities from duality also show that the $\|A\| = \|A^\top\|$ i.e., the maximum right singular value is the same as the maximum left singular value). We will also make use of the following claim relating the $q \to 2$ and $q \to q^*$ norms of a matrix.

Claim 4 *For any projection matrix Π, and $q \geq 2$, $\|\Pi\|_{q\to q^*} = \|\Pi\|_{q\to 2}^2$ (this is crucially an equality, and not just an inequality). More generally, for any matrix B, we have $\|B^\top B\|_{q\to q^*} = \|B\|_{q\to 2}^2$.*

Proof Note that by duality of norms, we have for any matrix B we have:

$$\begin{aligned}\|B^\top B\|_{q\to q^*} &= \max_{\|y\|_q\leq 1} \|B^\top By\|_{q^*} = \max_{\|y\|_q\leq 1, \|z\|_q\leq 1} z^\top B^\top By = \max_{\|y\|_q\leq 1, \|z\|_q\leq 1} \langle Bz, By\rangle \\ &= \max_{\|y\|_q\leq 1} \langle By, By\rangle = \max_{\|y\|_q\leq 1} \|By\|_2^2 = \|B\|_{q\to 2}^2.\end{aligned}$$

For a projection matrix Π we also have $\Pi = \Pi^\top$ and $\Pi^2 = \Pi$. Hence the lemma follows. ■

Entry-wise Norms of Matrices. We will also consider various matrix norms obtained by considering a matrix $M \in \mathbb{R}^{m\times n}$ as a vector of size mn. In particular, for any $q \geq 1$ we will use $\|M\|_q$ to denote the ℓ_q norm of the "flattened" vector corresponding to M i.e., $\|M\|_q^q = \sum_{i=1,j=1}^{m,n} |M(i,j)|^q$. The Frobenius norm $\|M\|_F = \|M\|_2$. Moreover for matrices A, B, we use $\langle A, B\rangle := \text{tr}(A^\top B)$ to represent the trace inner product.

Monotonicity of Matrix Norms. The following property of certain matrix norms will be crucial in designing constant factor approximation algorithms for the low-rank approximations.

Definition 5 (Monotone matrix norm) *A matrix norm $|||\cdot|||$ is said to be monotone if and only if*

$$\forall A, B \succeq 0, \quad |||A + B||| \geq |||A|||. \tag{12}$$

Observe that it suffices to check the above condition for all rank-1 PSD matrices B i.e., $B = vv^\top$ for $v \in \mathbb{R}^n$. It is well known that all unitarily invariant matrix norms[4] are monotone (this is because unitarily invariant norms are just norms on the singular values). On the other hand, many other matrix norms including other entry-wise norms $\|X\|_q$ or general operator norms $\|X\|_{q\to p}$ are not necessarily monotone (see Claim 50 and Claim 51 for some counterexamples). Perhaps surprisingly, the $q \to q^*$ matrix operator norms are monotone (see Claim 15 for a simple proof of this fact)!

High probability bounds. We will say that an event holds *with high probability (w.h.p.)* if the probability of failure on a given instance is less than *any* polynomial of the input parameters e.g., the dimension n, and the number of data points m. We remark that in all our settings, one can amplify the success probability to $1-\eta$ for any small $\eta > 0$ by repeating the algorithm $\log(1/\eta)$ times.

A.1. Properties of Robust Projections.

Throughout the paper we will use the term projections and projection matrices to always refer to orthogonal projection matrices on to linear subspaces of $\mathbb{R}^n$. Next we list and prove some simple properties of subspaces with *robust* projection matrices i.e., subspaces with $\|\Pi\|_{\infty\to 2}$ (or more generally $q \to 2$ norm for some $q \geq 2$) that is upper bounded.

For any $q^* \in [1,2]$, the ratio of the ℓ_{q^*} vs ℓ_2 corresponds to an analytic notion of sparsity. The following claim gives an upper bound on the ℓ_{q^*} norm in terms of the sparsity.

Claim 6 (Analytic Sparsity) *Consider any vector $v \in \mathbb{R}^n$ of support size k. For any $q^* \in [1,2]$, we have*

$$\|v\|_{q^*} \leq k^{\frac{1}{q^*}-\frac{1}{2}}\|v\|_2.$$

In particular, $\|v\|_1 \leq \sqrt{k}\|v\|_2$ for vectors with support size at most k.

On the other hand, it is easy to see that the bound given here is tight for any vector that is equally spread out among its support of size k.

Proof Without loss of generality suppose $\|v\|_2 = 1$ (if $v = 0$ it holds trivially). Let v have support S of size k. Set $p := 2/q^*$, and let u be the vector such that $u_i = |v_i|^{q*}$ for each $i \in [n]$. By Holder's inequality

$$\|v\|_{q^*}^{q^*} = \sum_{i\in S} 1 \cdot u_i \leq \|\mathbf{1}_S\|_{p^*}\|u\|_p \leq k^{1/p^*}\Big(\sum_i |v_i|^{pq^*}\Big)^{1/p} \leq k^{1-q^*/2}\|v\|_2^{2/p} = k^{1-q^*/2},$$

hence establishing the lemma. ■

Recall that ℓ_{q^*} corresponds to the dual norm for ℓ_q, and $q^* \in [1,2]$ when $q \geq 2$. The following simple lemma proves two useful properties of robust subspaces i.e., subspaces having projection matrices with bounded $\infty \to 2$ norm (or more generally $q \to 2$ norm for $q > 2$). The first property shows that any two vectors that are close in ℓ_∞ norm will have nearby projections onto any subspace that is robust. The second property shows that a

4. A matrix norm $|||\cdot|||$ is unitarily invariant iff $|||A||| = |||UAV|||$ for all matrices A and all unitary matrices U, V.

subspace is robust (i.e., has a robust projection matrix) exactly when every vector in the subspace is *analytically* sparse.

Lemma 7 *[Properties of Robust Subspaces and Projections] Consider any subspace of $\mathcal{V} \subseteq \mathbb{R}^n$ with projection matrix $\Pi \in \mathbb{R}^{n\times n}$ satisfying $\|\Pi\|_{q\to 2} \le \kappa$. We have the following two properties:*

I. **Closeness of projections of pertubations:** *For any vector v and its perturbation $\tilde{v}$*

$$\|v - \tilde{v}\|_q \le \delta \quad \Longrightarrow \quad \|\Pi\tilde{v} - \Pi v\|_2 \le \kappa\delta.$$

II. **Analytic sparsity:** *For any $v \in \mathcal{V}$, we have $\|v\|_{q^*} \le \kappa\|v\|_2$, where $q^* = q/(q-1)$. Moreover, if every vector in $\mathcal{V}$ has $\|v\|_{q^*} \le \kappa\|v\|_2$, then $\|\Pi\|_{q\to 2} \le \kappa$. In particular $\|\Pi\|_{\infty\to 2} \le \kappa$ if and only if $\|v\|_1 \le \kappa$ for all unit vectors $v \in \mathbb{S}^{n-1} \cap \mathcal{V}$.*

Proof We first show property (I). Let $u := v - \tilde{v}$. Then

$$\|\Pi\tilde{v} - \Pi v\|_2 = \|\Pi u\| \le \|\Pi\|_{q\to 2}\|u\|_q \le \kappa\delta.$$

To show property (II), note that by duality of matrix operator norms we have $\|\Pi\|_{q\to 2} = \|\Pi^\top\|_{2\to q^*} = \|\Pi\|_{2\to q^*}$.

$$\text{Hence} \quad \forall v \in \mathbb{S}^{n-1} \cap \mathcal{V}, \quad \|v\|_{q^*} = \|\Pi v\|_{q^*} \le \|\Pi\|_{2\to q^*}\|v\|_2 \le \kappa.$$

For the converse, if there exists $v \in \mathbb{S}^{n-1} \cap \mathcal{V}$ s.t. $\|v\|_{q^*} > \kappa$, then by duality $\|\Pi\|_{2\to q^*} = \|\Pi\|_{q\to 2} > \kappa$.

■

Observe that the robustness condition on the subspace as captured by the $q \to 2$ operator norm bound of its projection matrix Π is basis independent.The following simple claim gives a simple sufficient condition on the basis of the subspace that implies robustness of the subspace spanned by them. This relates our robustness of the subspace to alternate notions of sparsity of subspaces that have been studied in the literature on sparse PCA (Vu and Lei, 2012, 2013).

Claim 8 *Given any orthonormal basis $v_1, v_2, \ldots, v_r$ for a subspace $\mathcal{V}$ such that $\|v_i\|_{q^*} \le \kappa$ for each $i \in [r]$, we have $\|\Pi\|_{q\to 2} \le \sqrt{r}\kappa$.*

Proof Firstly, $\Pi = \sum_{i=1}^r v_i v_i^\top$, and $\|\Pi\|_{q\to 2} = \|\Pi\|_{2\to q^*}$. We have

$$\|\Pi\|_{q\to 2} = \max_{u:\|u\|_q\le 1} \Big\| \sum_{i=1}^r \langle u, v_i\rangle v_i \Big\|_2 = \max_u \sqrt{\sum_{i=1}^r \langle u, v_i\rangle^2} \le \sqrt{r}\cdot \max_{u:\|u\|_q\le 1} \max_{v:\|v\|_{q^*}\le\kappa} |\langle u, v\rangle| \le \sqrt{r}\kappa.$$

■

For a given matrix $B \in \mathbb{R}^{n\times m}$, let us denote by $\Pi(B)$ to the projection matrix onto the column space of B. The following lemma shows that the best low-rank (κ, q)-robust projection objective (3) also finds the low-rank approximation that has smallest error among ones with a (κ, q) robust column space.

Claim 9 *Let $\mathcal{P}_r$ be the set of all rank-r projection matrices. Given a data matrix $A \in \mathbb{R}^{n\times m}$ and a given parameter $\kappa \geq 1$, $q > 0$, we have*

$$\min_{\substack{\Pi\in\mathcal{P}_r \\ \|\Pi\|_{q\to 2}\leq\kappa}} \|A - \Pi A\| = \min_{\substack{B:rank(B)\leq r, \\ \|\Pi(B)\|_{q\to 2}\leq\kappa}} \|A - B\|,$$

where $\|M\|$ here stands for the spectral norm. The above statement is also true for the Frobenius norm.

Proof Let B^* be the minimizer for the right minimization problem and let $\Pi_2 = \Pi(B^*)$ be its projection matrix, and let Π_1 be the minimizer for the left optimization problem. It is easy to see that $\|A - \Pi_1 A\| \geq \|A - B^*\|$, since $\Pi_1 A$ is also a feasible choice for B in the right minimization problem. The other direction follows from the fact that $\|A - \Pi(B)A\| \leq \|A - B\|$ holds for both Frobenius norm and the spectral norm (specifically, $\|Av - \Pi(B)Av\|_2 \leq \|Av - Bv\|_2$ for any $v \in \mathbb{R}^m$). ■

A.2. Approximation Algorithms for Operator Norms.

Here we briefly describe some known positive and negative results for approximating the $q \to p$ operator norm of a matrix (sometimes referred to as the (ℓ_q, ℓ_p)-Grothendieck problem). We will say that a randomized algorithm gives an *α-factor approximation* for the $q \to p$ operator norm (for some $\alpha \geq 1$) iff for any input matrix M the algorithm outputs with probability at least $(1 - n^{-\omega(1)})$ a vector $\widehat{x} \neq 0$ such that $\|M\widehat{x}\|_p / \|\widehat{x}\|_q \geq \frac{1}{\alpha} \cdot \|M\|_{q\to p}$. The $\infty \to 1$ norm is the well-known Grothendieck's problem (Grothendieck, 1952) (that is related to the cut-norm of a matrix (Alon and Naor, 2004) and has a rich history.

There is a lot of work on approximation algorithms and inapproximability results for computing these $q \to p$ norms (Nesterov, 1998; Alon and Naor, 2004; Bhaskara and Vijayaraghavan, 2011; Barak et al., 2012; Bhattiprolu et al., 2018a,b). Regarding approximation algorithms, the works of (Nesterov, 1998; Wolkowicz et al., 2012; Steinberg, 2005) provides a $1/(\frac{2\sqrt{3}}{\pi} - \frac{2}{3}) \approx 2.29$ approximation for when $1 \leq p \leq 2 \leq q \leq \infty$, and for the special case $p = 2$ or $q = 2$ the factor becomes $\sqrt{\pi/2} \approx 1.25$. Recently, improved upper and (almost matching) lower bounds were proved for many settings of q, p in (Bhattiprolu et al., 2018a,b). Formally, we have the following guarantee where $\gamma_{q^*}^{q^*}$ is the q^*th moment of a standard normal random variable.

Theorem 10 ((Bhattiprolu et al., 2018a,b; Nesterov, 1998; Steinberg, 2005)) *For computing the $\infty \to 2$ norm, there is a randomized polynomial time algorithm that gives a $\sqrt{\pi/2} \approx 1.25$-approximation, and for the $q \to 2$ norm there is a randomized polynomial time algorithm that gives a $1/\gamma_{q^*}$-factor approximation. Furthermore, when the input matrices are positive semidefinite, the integrality gap of the aforementioned SDP is $\pi/2$ for the $\infty \to 1$ norm, and $1/\gamma_{q^*}^2$ for the $q \to q^*$ operator norm respectively. Using a generalization of random hyperplane rounding, this SDP yields approximation algorithms that succeed with high probability for any given instance.*

Appendix B. Worst-case Approximation Guarantees

In this section, we show the approximation algorithm bounding the *Frobenius* norm error in Section B.1 and the approximation algorithm bounding the *spectral* norm error in Section B.2 separately.

B.1. Approximations in Frobenius Norm Error

We will aim to obtain a bicriteria approximation for the robust low-rank approximation problem given in (3) for the Frobenius norm error. In the rest of the section, we will focus on the the formulation where the objective is to minimize projection error, subject to a specified robustness requirement. It is easy to see that one can switch the role of the objective and constraint here and obtain similar guarantees for minimizing the robustness parameter κ, subject to a bound on the projection error. In what follows $q \in [2, \infty]$.

$$\min_{\Pi} \|\Pi^{\perp} A\|_F^2 = \min_{\Pi} \|A\|_F^2 - \langle AA^\top, \Pi \rangle \tag{13}$$

$$\text{s.t. } \Pi \text{ is a projection matrix of rank } \leq r, \text{ and } \|\Pi\|_{q\to 2} \leq \kappa. \tag{14}$$

We prove the following theorem.

Theorem 11 *Suppose the data matrix $A \in \mathbb{R}^{n\times m}$ has an (orthogonal) projection Π^* of rank at most r such that $\|\Pi^*\|_{q\to 2} \leq \kappa$ and the approximation error $OPT := \|(I - \Pi^*)A\|_F^2$. There exists a polynomial time algorithm such that given any $\gamma > 0$, it finds a projection matrix $\widehat{\Pi}$ of rank at most r satisfying*

$$\|\widehat{\Pi}\|_{q\to 2} \leq \sqrt{C_G(q)(1 + 1/\gamma)} \cdot \kappa, \text{ and } \|(I - \widehat{\Pi})A\|_F^2 \leq (2 + \gamma)OPT, \tag{15}$$

where $C_G(q) > 0$ is a constant that only depends on q as given in Theorem 10 (for $q = \infty$ this value is at most $\pi/2$). Moreover, for any $\gamma > 0$, there exists an algorithm that runs in polynomial time and finds an $r' \leq r(1+1/\gamma)$-dimensional orthogonal projection $\widehat{\Pi}$ such that

$$\|\widehat{\Pi}\|_{q\to 2} \leq \sqrt{C_G(q)(1 + 1/\gamma)} \cdot \kappa, \text{ and } \|(I - \widehat{\Pi})A\|_F^2 \leq (1 + \gamma)OPT. \tag{16}$$

The theorem above will be established by proving a statement about the more general problem of finding a low-rank projection under *any* monotone norm constraint that can be approximately certified. While the $q \to 2$ norm is not monotone as discussed in Section A, we will show that applying the more general guarantee on an appropriate monotone norm helps prove Theorem 11 above. Let $|||\cdot|||$ be a monotone matrix norm. Consider the following generalization of problem (1) that given a data matrix $A \in \mathbb{R}^{n\times m}$, and a parameter $\kappa \geq 1$, finds a projection

$$\min_{\Pi} \|A\|_F^2 - \langle AA^\top, \Pi \rangle \text{ s.t. } \Pi \text{ is a projection matrix of rank} \leq r, \text{ and } |||\Pi||| \leq \kappa. \tag{17}$$

Definition 12 *[α-approximately certifiable matrix norm] A matrix norm $|||\cdot|||$ over $\mathbb{R}^{n\times m}$ matrices is α-approximately certifiable for some $\alpha \geq 1$ iff there exists an algorithm that runs in time $\mathrm{poly}(n, m)$, and when given a PSD matrix $B \in \mathbb{R}^{n\times n}$ and a parameter κ as input will either certify that $|||B||| \leq \alpha\kappa$, or finds a $Z \in \mathbb{R}^{n\times n}$ such that (1) $\langle B, Z\rangle > \kappa$, and (2) $\langle M, Z\rangle \leq \kappa$ for all M s.t. $|||M||| \leq \kappa$.*

As we will see later the operator norms that we will consider (e.g., $q \to 2$ norm and the $q \to q^*$ norm for $q \geq 2$) will be $O(1)$-approximately certifiable.

The following general theorem gives an $O(1)$ bicriteria approximation for the problem assuming the monotone matrix norm $|||\cdot|||$ is approximately certifiable.

Theorem 13 *Let $|||\cdot|||$ be any matrix norm that is* monotone *and α-approximately certifiable for some $\alpha \geq 1$. Suppose we are given as input a data matrix $A \in \mathbb{R}^{n\times m}$ that has an (orthogonal) projection Π^* of rank at most r such that $|||\Pi^*||| \leq \kappa$ and the approximation error $OPT := \|(I-\Pi^*)A\|_F^2$. There exists a polynomial time algorithm such that given every $\gamma \in (0,1)$, it finds an orthogonal projection matrix $\widehat{\Pi}$ of rank at most r satisfying*

$$|||\widehat{\Pi}||| \leq \alpha\big(1+\tfrac{1}{\gamma}\big)\cdot\kappa, \text{ and } \|(I-\widehat{\Pi})A\|_F^2 \leq \Big(2+\gamma\Big)OPT. \tag{18}$$

Moreover, for any $\gamma \in (0,1)$, there exists a polynomial time algorithm that finds an $r' \leq (1+1/\gamma)r$-dimensional orthogonal projection $\widehat{\Pi}$ such that

$$|||\widehat{\Pi}||| \leq \alpha\big(1+\tfrac{1}{\gamma}\big)\cdot\kappa, \text{ and } \|(I-\widehat{\Pi})A\|_F^2 \leq (1+\gamma)\cdot OPT. \tag{19}$$

We consider the following mathematical programming relaxation for the problem. In the alternate formulation where we minimize the robustness parameter subject to an upper bound on projection error, the roles of (20) and (23) below is switched.

$$\min_X \|A\|_F^2 - \langle AA^\top, X\rangle \tag{20}$$
$$\text{s.t.} \operatorname{tr}(X) \leq r \tag{21}$$
$$0 \preceq X \preceq I \tag{22}$$
$$|||X||| \leq \kappa \tag{23}$$

First we observe that this is a valid convex relaxation to the problem. In fact any feasible projection matrix Π of rank at most r for (17) is a feasible solution to the above program (20)-(23) with the same value. The intended solution here is just $X = \Pi$. All the eigenvalues of Π are 0 or 1, since Π is a projection matrix; hence (21), (22) are satisfied. Moreover (23) is satisfied just because of the same constraint as in (17). Finally, the objective value is preserved since

$$\|A\|_F^2 - \langle AA^\top, \Pi\rangle = \|A\|_F^2 - \operatorname{tr}(AA^\top\Pi) = \|A\|_F^2 - \operatorname{tr}(\Pi A(\Pi A)^\top) = \|A\|_F^2 - \|\Pi A\|_F^2.$$

In the above program, the objective (20) and constraints (21)-(22) define a semi-definite program (SDP). Moreover, for (23), we see that for any $\lambda \in [0,1]$, by triangle inequality $|||\lambda X_1 + (1-\lambda)X_2||| \leq \lambda|||X_1||| + (1-\lambda)|||X_2|||$. Hence the set of all X that satisfies constraints (21) - (23) is convex. In general, constraint (23) may be NP-hard to verify for a given PSD matrix X. However, we can use the fact that $|||\cdot|||$ is approximately certifiable to get a approximately feasible solution to the program in polynomial time, using the Ellipsoid method.

The following lemma shows that by truncating a solution of the program (20)-(23) to just the terms corresponding to the large eigenvalues, we retain much of the objective.

Lemma 14 *Let $\varepsilon, \delta > 0$, and $M \succeq 0$. Suppose X satisfies the SDP constraints (22) and $\langle M, X\rangle \geq (1-\varepsilon)tr(M)$. Suppose $P_X^{1-\delta}$ is the projection operator onto the subspace spanned by eigenvectors of X with eigenvalues at least $(1-\delta)$. Then we have*

$$\langle I - P_X^{1-\delta}, M\rangle \leq \frac{\varepsilon}{\delta} \cdot tr(M). \tag{24}$$

Proof We can assume without loss of generality that $\mathrm{tr}(M) = 1$, since M can be scaled accordingly. Let $X = \sum_{i=1}^n \lambda_i v_i v_i^\top$ be the eigendecomposition of X (note that $\lambda_i \geq 0$ since X is p.s.d.), and let $S = \{i : \lambda_i \geq 1-\delta\}$. We have

$$(1-\varepsilon)\mathrm{tr}(M) \leq \langle M, X\rangle = \sum_i \lambda_i v_i^\top M v_i$$

$$\mathrm{tr}(M) = \sum_i v_i^\top M v_i, \text{ since } \{v_i : i \in [n]\} \text{ is an orthonormal basis.}$$

By subtracting the two inequalities, we get

$$\sum_{i=1}^n (1-\lambda_i) v_i^\top M v_i \leq \varepsilon \cdot \mathrm{tr}(M) = \varepsilon.$$

$$\sum_{i \notin S} \delta v_i^\top M v_i \leq \sum_{i \notin S} (1-\lambda_i) v_i^\top M v_i + \sum_{i \in S} (1-\lambda_i) v_i^\top M v_i \leq \varepsilon$$

from definition of $S, M \succeq 0$, and (22). Hence

$$\langle I - P_X^{1-\delta}, M\rangle = \sum_{i \notin S} v_i^\top M v_i \leq \frac{\varepsilon}{\delta} = \frac{\varepsilon}{\delta} \cdot \mathrm{tr}(M), \text{ as required.}$$

■

Proof of Theorem 13. We can scale the matrix A appropriately so that we can assume $\|A\|_F = 1$ without loss of generality. Let $OPT = \varepsilon$ for some $\varepsilon \in [0,1]$. We will use the Ellipsoid algorithm to approximately solve the relaxation in (20)-(23). As we have explained before, the feasible set is convex. We now show how to design an approximate hyperplane separation oracle for (23); the rest of the constraints just correspond to a simple SDP. Since $|||\cdot|||$ is α-approximately certifiable, we have a polynomial time algorithm that given a matrix $\widehat{X} \succeq 0$, either certifies that $|||\widehat{X}||| \leq \alpha\kappa$ (e.g., when the SDP value is at most $\alpha\kappa$), and otherwise produces a separating hyperplane of the form $\langle Z, X\rangle \leq \kappa$ that is not satisfied by $\widehat{X}$. We run the Ellipsoid algorithm to find a solution $\widehat{X}$ that satisfies $|||\widehat{X}||| \leq \alpha\kappa$, and has objective value that is arbitrarily close to OPT.

Set $\delta := 1/(1+\gamma)$. Let $\widehat{X} = \sum_i \lambda_i v_i v_i^\top$ and let $S = \{i : \lambda_i \geq 1-\delta\}$. Define for each $i \in [S]$, $\alpha_i := \langle v_i v_i^\top, M\rangle$, where $M = AA^\top$. We sort the elements of S based on $\{\alpha_i\}$, and pick greedily the first $\min\{r, |S|\}$ of them to form $T \subseteq S$. Our projection matrix will be $\Pi_T = \sum_{i \in T} v_i v_i^\top$.

We first argue that the operator norm constraint is approximately satisfied. By the monotonicity of the $|||\cdot|||$ we have

$$|||\Pi_T||| = |||\sum_{i \in T} v_i v_i^\top||| \leq \frac{1}{1-\delta} |||\sum_{i:\lambda_i > 1-\delta} \lambda_i v_i v_i^\top||| \leq \frac{|||\widehat{X}|||}{1-\delta} \leq \frac{\alpha\kappa}{1-\delta} = \frac{\alpha(1+\gamma)}{\gamma} \cdot \kappa \tag{25}$$

Also, from Lemma 14, we have

$$\sum_{i \notin S} \langle v_i v_i^\top, M \rangle = \langle I - P_{\widehat{X}}^{1-\delta}, M \rangle \leq \frac{\varepsilon}{\delta}, \text{ and}$$

$$\sum_{i \in S} \langle (1-\lambda_i) v_i v_i^\top, M \rangle \leq \sum_i \langle (1-\lambda_i) v_i v_i^\top, M \rangle \leq 1 - \langle X, M \rangle \leq \varepsilon.$$

$$\text{Hence, } \sum_{i \in S} \lambda_i \alpha_i = \sum_{i \in S} \lambda_i \langle v_i v_i^\top, M \rangle \geq 1 - \varepsilon\Big(1 + \frac{1}{\delta}\Big) = 1 - (2+\gamma)\varepsilon,$$

for our choice of $\delta = 1/(1+\gamma)$. By our greedy choice of T, we have $\sum_{i \in T} \alpha_i \geq \sum_{i \in S} \lambda_i \alpha_i$, as $\sum_{i \in S} \lambda_i \leq \min\{\mathrm{tr}(X), |S|\} = |T|$, with each $\lambda_i \in [0,1]$. Thus $\|\Pi_T^\perp A\|_F^2 \leq (2+\gamma)\varepsilon$.

The guarantee in (16) is obtained by returning the projection $\Pi_S = \sum_{i \in S} v_i v_i^\top$. Observe that $|S| \leq r/(1-\delta)$ from (21). The operator norm bounds follows using the same argument as (25) with $T = S$. Moreover, the objective value follows directly from Lemma 14. ■

Guarantees for the $q \to 2$ norm (Proof of Theorem 11). Our goal will be to apply Theorem 13 to obtain our required guarantee. However the $q \to 2$ operator norm is not monotone when $q > 2$; see Claim 51 for a counter-example. Our crucial insight is that we can instead use the $\|\cdot\|_{q \to q^*}$ norm which we show indeed satisfies the monotonicity property (Definition 5), so that we can apply Theorem 13.

Claim 15 (Monotonicity of $q \to q^*$ operator norm) *For any $q \geq 1$, the operator norm $\|\cdot\|_{q \to q^*}$ is monotone.*

Proof Let $B \in \mathbb{R}^{n \times n}$. It suffices to prove for any $B \succeq 0, v \in \mathbb{R}^n$, $\|B + vv^\top\|_{q \to q^*} \geq \|B\|_{q \to q^*}$.

$$\begin{aligned} \|B\|_{q \to q^*} &= \max_{\|y\|_q \leq 1, \|z\|_q \leq 1} z^\top B y = \max_{\|y\|_q \leq 1, \|z\|_q \leq 1} \langle B^{1/2} z, B^{1/2} y \rangle \\ &= \max_{\|y\|_q \leq 1} \langle B^{1/2} y, B^{1/2} y \rangle = \max_{\|y\|_q \leq 1} y^\top B y. \end{aligned}$$

In other words, the quadratic form is maximized by $y = z$. Moreover for every y,

$$y^T (B + vv^\top) y = y^T B y + \langle y, v \rangle^2 \geq y^\top B y.$$

Hence, $\|B + vv^\top\|_{q \to q^*} \geq \|B\|_{q \to q^*}$.

■

This gives the following mathematical programming relaxation for the problem, where the robustness constraint is captured by the $q \to q^*$ operator norm.

$$\min_X \|A\|_F^2 - \langle AA^\top, X \rangle \tag{26}$$

$$\text{s.t. } \mathrm{tr}(X) \leq r \text{ and } 0 \preceq X \preceq I \tag{27}$$

$$\|X\|_{q \to q^*} \leq \kappa^2 \tag{28}$$

In the above program, the objective (26) and constraints (27) define a semi-definite program (SDP). However, the $q \to q^*$ operator norm is NP-hard to compute. Instead we will use the standard SDP relaxation (Nesterov, 1998; Steinberg, 2005) for certifying the $q \to q^*$ norm of any PSD matrix X, where $C_G = C_G(q) > 0$ is the constant given by Theorem 10:

$$\|X\|_{q\to q^*} \leq \max_{\substack{Y\in\mathbb{R}^{n\times n}: Y\succeq 0 \\ \sum_i Y_{ii}^{q/2}\leq 1}} \langle X, Y\rangle \leq C_G \|X\|_{q\to q^*}, \tag{29}$$

This shows that $q \to q^*$ norm is $\alpha = C_G$ approximately certifiable. At this point, we have all the ingredients to prove Theorem 11 by using Theorem 13 as a black-box. However, let us first consider the convex relaxation(s) suggested by this certificate, with a view towards designing a more efficient algorithm based on convex relaxations.

Convex relaxations for the Frobenius norm objective. The program (26) along with the approximate certificate for $q \to q^*$ norm from (29) leads to the following two tractable relaxations for our problem (13) that are equivalent. **CP1** involves a universal quantifier over $Y \in \mathcal{Q}$. Fortunately, there exists an efficient hyperplane separation oracle for the constraint (32) (which by itself is an SDP relaxation). This can be used with the Ellipsoid algorithm to solve the above relaxation in polynomial time. We now instead give an equivalent convex relaxation for (30) that is much more efficient to solve. The main idea is to use Lagrangian duality to convert the universal quantifier in (32) into an existential quantifier. In what follows for a vector $d \in \mathbb{R}^n$, we will use $\mathrm{diag}(d)$ to denote the diagonal matrix in $\mathbb{R}^{n\times n}$ defined by d.

CP1 :

$$\min_X \|A\|_F^2 - \langle AA^\top, X\rangle \tag{30}$$

$$\text{s.t. } \mathrm{tr}(X) \leq r, \quad 0 \preceq X \preceq I \tag{31}$$

$$\max_{Y\in\mathcal{Q}} \langle X, Y\rangle \leq C_G\kappa^2, \text{ where}$$

$$\mathcal{Q} = \{Y \in \mathbb{R}^{n\times n} : \ Y \succeq 0, \sum_{i=1}^n Y_{ii}^{q/2} \leq 1\} \tag{32}$$

CP2 :

$$\min_{X\in\mathbb{R}^{n\times n}, d\in\mathbb{R}^n_{\geq 0}} \|A\|_F^2 - \langle AA^\top, X\rangle \tag{33}$$

$$\text{s.t. } \mathrm{tr}(X) \leq r, \quad 0 \preceq X \preceq I \tag{34}$$

$$X \preceq \mathrm{diag}(d) \tag{35}$$

$$\|d\|_{q/(q-2)} := \Big(\sum_{i=1}^n d_i^{q/(q-2)}\Big)^{(q-2)/q} \leq C_G\kappa^2. \tag{36}$$

Figure 3: Two tractable mathematical relaxations **CP1** and **CP2** for problem (13) with Frobenius norm objective.

The equivalence of **CP1** and **CP2** follows immediately from the following lemma that uses Lagrangian duality.

Lemma 16 *Consider the following two programs defined for $q \geq 2$ and input $C \in \mathbb{R}^{n\times n}$ with $C_{ii} \geq 0$ $\forall i \in [n]$.*

$$\textbf{\textit{primal:}}\ val_P := \max_{Y\in\mathbb{R}^{n\times n}} \langle C, Y\rangle \ \ s.t.\ \ \sum_{i=1}^{n} Y_{ii}^{q/2} \leq 1, \quad Y \succeq 0. \tag{37}$$

$$\textbf{\textit{dual:}}\ val_D := \min_{d\in\mathbb{R}^n} \|d\|_{q/(q-2)} = \Big(\sum_{i=1}^{n} d_i^{q/(q-2)}\Big)^{(q-2)/q} \ \ s.t. \quad \mathrm{diag}(d) \succeq C, \quad d \geq 0. \tag{38}$$

For any feasible solution Y of the primal, and any feasible solution d of the dual, we have $\langle C, Y\rangle \leq \|d\|_{q/(q-2)}$ i.e., weak duality holds. Moreover, the optimum values of the primal and the dual relaxations are equal $val_P = val_D$ (strong duality).

We remark that in the special case when $q = \infty$, the last constraint (36) becomes the simple linear constraint $\sum_i d_i \leq c_G\kappa^2$. The following claim shows that the above convex programs **CP1** and **CP2** are valid relaxations for (13) and can be solved in polynomial time.

Claim 17 *Any feasible projection matrix Π of rank r satisfying (14) forms feasible solutions to **CP1** , **CP2** (for an appropriate feasible d) and (26) with the same objective value as (13). Moreover the relaxations **CP1** and **CP2** can be solved in polynomial time to arbitrary accuracy.*

Proof First we argue about the feasibility of **CP1** and (26). The intended SDP solution here is just $X = \Pi$. All the eigenvalues of Π are 0 or 1, since Π is a projection matrix; hence (27) and the corresponding constraints in the programs (31) and (34) are satisfied. To verify (28), note that from Claim 4, $\|\Pi\|_{q\to q^*} = \|\Pi\|_{q\to 2}^2$. Moreover from (29), we also have that (32) is satisfied by X. Hence **CP1** (and (26)) are feasible. Moreover X is also feasible for **CP2** for an appropriate choice of d since the constraints (32) and (36) are equivalent from Lemma 16.

Finally, the objective value is preserved since

$$\|A\|_F^2 - \langle AA^\top, \Pi\rangle = \|A\|_F^2 - \mathrm{tr}(AA^\top\Pi) = \|A\|_F^2 - \mathrm{tr}(\Pi A(A\Pi)^\top) = \|A\|_F^2 - \|\Pi A\|_F^2.$$

It is easy to check that both relaxations **CP1** and **CP2** are convex. We now argue about the computational tractability of **CP2** . The relaxation is a semi-definite program (SDP) with an extra constraint (36). The constraint (36) is convex when $q \geq 2$. Let $q' = q/(q-2)$; the dual norm for $\ell_{q'}$ is $\ell_{q/2}$. Moreover there is a simple separation oracle for this constraint since by duality

$$\|d\|_{q/(q-2)} = \max_{y\in\mathbb{R}^n:\|y\|_{q/2}\leq 1} \langle y, d\rangle = \left\langle \frac{d^*}{\|d^*\|_{q/2}}, d \right\rangle, \text{ where } d_i^* = \mathrm{sign}(d_i)|d(i)|^{2/(q-2)} \ \ \forall i \in [n].$$

Hence by using the Ellipsoid algorithm, this problem can be solved in polynomial time. Finally **CP1** can also be solved in polynomial time since (32) is itself a semi-definite program: this can be solved in polynomial time (by Ellipsoid method for example) and can in turn be used as a hyperplane separation oracle for constraint (32) and the outer relaxation.

We remark that solving **CP1** is less computationally efficient compared to **CP2** due to the repeated use of the Ellipsoid method (to certify (32)). ■

Proof of Theorem 11. We will apply Theorem 13 with $|||\cdot||| := \|\cdot\|_{q\to q^*}$. From Lemma 15, we have that the $\|\cdot\|_{q\to q^*}$ is monotone. We now show that $\|\cdot\|_{q\to q^*}$ is $O(1)$-approximately certifiable. For this we use the constraint (32) of **CP1** or equivalently (36) of **CP2** . As shown in Claim 17 both of these constraints can be certified efficiently. By (29), this in turn shows that $\|\cdot\|_{q\to q^*}$ is $\alpha = C_G(q) = O_q(1)$-approximately certifiable (in particular, $\alpha = \pi/2$ for $\infty \to 1$). Hence, applying Theorem 13 we see that the algorithm outputs a projection matrix $\widehat{\Pi}$ with $\|\widehat{\Pi}\|_{q\to 2}^2 = \|\widehat{\Pi}\|_{q\to q^*} \le C_G(q)(1+\frac{1}{\gamma})\kappa^2$ which obtains an objective value of $(2+\gamma)OPT$. This completes the proof of Theorem 11.

We remark that applying Theorem 13 as a black-box involves solving an SDP to α-approximately certify the $\|\cdot\|_{q\to q^*}$ – this corresponds to solving the convex relaxation **CP1** . However, using the equivalence between **CP1** and **CP2** (from Lemma 16), we can also apply the same rounding algorithm from Theorem 13 to the solution X obtained from **CP2** to obtain the same guarantees. ■

We now complete the proof of Lemma 16 which establishes the equivalence of **CP1** and **CP2** .

Proof of Lemma 16. First we observe that the primal (37) is a conic program (PSD cone) with an additional convex constraint. Consider the following Lagrangian dual of the conic program

$$\min_{\lambda\ge 0, U\succeq 0} L(\lambda, U), \text{ where } L(\lambda,U) := \lambda + \max_{Y\in\mathbb{R}^{n\times n}} \langle C, Y\rangle - \lambda\sum_{i=1}^{n} Y_{ii}^{q/2} + \langle U, Y\rangle. \tag{39}$$

$$\text{Moreover by duality, } \ \text{val}_P = \min_{\lambda\ge 0, U\succeq 0} L(\lambda, U). \tag{40}$$

The second line follows from strong duality for the Lagrangian dual – it is easy to see that Slater's condition holds for the primal (there exists a primal feasible solution Y where all the constraints hold strictly). Note that $U \succeq 0$ (since the PSD cone is its own self dual). Note that for the optimal U, we have $C_{ij} + U_{ij} = 0$ for all $i \ne j \in [n]$ (otherwise Y can be chosen to make the objective go to ∞). By introducing the variable $d \in \mathbb{R}^n$ with $d_i = C_{ii} + U_{ii}$ for all $i \in [n]$, we can rewrite (39) as

$$\min_{\substack{\lambda\ge 0, d\in\mathbb{R}^n \\ \text{diag}(d)\succeq C}} L(\lambda, d), \text{ where } L(\lambda, d) = \lambda + \max_{Y\in\mathbb{R}^{n\times n}} \sum_{i=1}^{n} d_i Y_{ii} - \lambda Y_{ii}^{q/2}. \tag{41}$$

Note that the inner maximization objective is concave in the variables $\{Y_{ii}\}$; hence its maximum is obtained where the gradient vanishes i.e.,

$$\forall i \in [n],\ d_i = \lambda\Big(\frac{q}{2}\Big) Y_{ii}^{q/2-1} \quad \Longrightarrow \quad Y_{ii} = \Big(\frac{2d_i}{q}\Big)^{q/(q-2)}.$$

Further $d_i \geq 0$ for all $i \in [n]$. Substituting in (41) and simplifying we get that (41) is equivalent to

$$\min_{\substack{\lambda \geq 0, d \in \mathbb{R}^n_{\geq 0} \\ \mathrm{diag}(d) \succeq C}} L(\lambda, d), \text{ where } L(\lambda, d) = \lambda + \Big(\frac{q-2}{q}\Big)\Big(\frac{2}{q\lambda}\Big)^{2/(q-2)} \sum_{i=1}^{n} d_i^{q/(q-2)}. \tag{42}$$

Again $L(\lambda, d)$ is convex is λ; hence its minimum is attained at the critical point (that is strictly positive)

$$\lambda = \frac{2}{q}\Big(\sum_{i=1}^{n} d_i^{q/(q-2)}\Big)^{(q-2)/q}.$$

Substituting this value of λ in (42) proves that (39) is equivalent to the claimed dual formulation (38). Finally from Lagrangian duality and (40) the lemma follows. ■

B.2. Approximations in the Spectral Norm

We now show how techniques similar to those in Section B.2 can also be extended to robust low-rank approximations, when the error is measured in spectral norm as opposed to the Frobenius norm. In this section we will use $\|A\|$ to denote the spectral norm of matrix A. For convenience of exposition, we will measure the projection error in (3) using the spectral norm $\|A - \Pi A\|$ as opposed to the squared spectral norm.

Theorem 18 *Suppose the data matrix $A \in \mathbb{R}^{n \times m}$ has an (orthogonal) projection Π^* of rank at most r such that $\|\Pi^*\|_{q \to 2} \leq \kappa$ and the approximation error $OPT := \|(I - \Pi^*)A\|$. There exists a polynomial time algorithm such that given any $\gamma \in (0, 1)$, it finds an (orthogonal) projection matrix $\widehat{\Pi}$ of rank at most r satisfying*

$$\|\widehat{\Pi}\|_{q \to 2} \leq \sqrt{C_G(q)(1 + 2/\gamma)} \cdot \kappa, \text{ and } \|(I - \widehat{\Pi})A\| \leq \sqrt{(3 + \gamma)} \cdot OPT, \tag{43}$$

where $C_G(q) > 0$ is a constant that only depends on q as given in Theorem 10. For $q = \infty$ this value is known to be at most $\pi/2$.

Moreover, for any $\gamma \in (0, 1)$, there exists a polynomial time algorithm that finds an $r' \leq (1 + \frac{2}{\gamma})r$ dimensional orthogonal projection $\widehat{\Pi}$ such that

$$\|\widehat{\Pi}\|_{q \to 2} \leq \sqrt{C_G(q)(1 + 2/\gamma)} \cdot \kappa, \text{ and } \|(I - \widehat{\Pi})A\| \leq \sqrt{(1 + \gamma)} \cdot OPT. \tag{44}$$

Proof of Theorem 18. We will use the following mathematical relaxation for the problem.

$$\min \quad \lambda \tag{45}$$

$$\text{s.t. } A^\top (I - X)A \preceq \lambda I \tag{46}$$

$$\mathrm{tr}(X) \leq r, \quad \text{and } 0 \preceq X \preceq I \tag{47}$$

$$\|X\|_{q \to q^*} \leq \kappa^2 \tag{48}$$

The last constraint (48) is NP-hard to certify. So as in the previous section we will relax it and consider the following two convex relaxations **CP3** and **CP4** .

The following claim (which is analogous to Claim 17) shows that the above convex programs are valid relaxations and can be solved in polynomial time.

CP3 :

$$\min \ \lambda \tag{49}$$
$$\text{s.t. } A^\top(I-X)A \preceq \lambda I \tag{50}$$
$$\operatorname{tr}(X) \le r, \quad \text{and } 0 \preceq X \preceq I \tag{51}$$
$$\max_{Y\in\mathcal{Q}} \langle X, Y\rangle \le C_G\kappa^2, \text{ where}$$
$$\mathcal{Q} = \{Y \in \mathbb{R}^{n\times n} : \ Y \succeq 0, \sum_i Y_{ii}^{q/2} \le 1\} \tag{52}$$

CP4 :

$$\min \ \lambda \tag{53}$$
$$\text{s.t. } A^\top(I-X)A \preceq \lambda I \tag{54}$$
$$\operatorname{tr}(X) \le r, \ \ 0 \preceq X \preceq I \tag{55}$$
$$X \preceq \operatorname{diag}(d) \tag{56}$$
$$\|d\|_{q/(q-2)} := \Big(\sum_{i=1}^n d_i^{q/(q-2)}\Big)^{(q-2)/q} \le C_G\kappa^2 \tag{57}$$

Figure 4: Two tractable relaxations **CP3** and **CP4** for spectral norm objective.

Claim 19 *Any feasible projection matrix* Π *of rank* r *satisfying* $\|\Pi\|_{q\to 2} \le \kappa$ *forms feasible solutions to* ***CP3*** , ***CP4*** *(for an appropriate feasible* d*) and* (45) *with objective value* $\|A - \Pi A\|^2$*. Moreover the relaxations* ***CP3*** *and* ***CP4*** *can be solved in polynomial time to arbitrary accuracy.*

Proof The proof follows the same argument as Claim 17, with a small modification to account for the different objective. We first argue that **CP3** and **CP4** are valid relaxations. Consider any projection matrix Π of rank r satisfying $\|\Pi\|_{q\to 2} \le \kappa$. From Claim 17, we have that constraints (51) and (52) of **CP3** and (55) and (57) of **CP4** are satisfied. To see that the objective value is preserved, note that for any projection matrix Π,

$$\|A^\top(I-\Pi)A\| = \|A^\top\Pi^\perp\Pi^\perp A\| = \|\Pi^\perp A\|^2 = \|A - \Pi A\|^2,$$

as required. This establishes the first part of the claim.

We now show that **CP3** and **CP4** are polynomial time solvable up to accuracy $\eta > 0$ in time polynomial in the input size and $\log(1/\eta)$. We use the Ellipsoid algorithm to solve both the relaxations. It is easy to verify that the feasible sets are convex. The argument in Claim 17 proves that the constraints (51), (52), (55) and (57) are all efficiently separation. We now argue about the objective i.e., the constraint (50) and (54). Finally, given λ, X, (50) and (54) can also be efficiently separated by computing the maximum eigenvalue of $A^\top(I-X)A$. Let $v \in \mathbb{S}^{n-1}$ be the corresponding eigenvector. If the constraint is violated, the hyperplane separator is of the form

$$\langle vv^\top, A^\top A\rangle - \langle vv^\top, A^\top XA\rangle - \lambda \le 0, \text{ i.e., } \langle vv^\top, A^\top A\rangle - \langle Avv^\top A^\top, X\rangle - \lambda \le 0,$$

since $\operatorname{tr}(vv^\top A^\top XA) = \operatorname{tr}(Avv^\top A^\top X)$. This completes the proof. ■

The proof of Theorem 18 also crucially uses the monotonicity of $q \to q^*$ matrix operator norm, and also follows the same outline as the Frobenius norm objective. The primary difference arises in analyzing the objective.

Proof of Theorem 18. Let $OPT^2 := \varepsilon^2\|A\|^2$ for some $\varepsilon \in (0, 1]$. Set $\delta := 2/(2+\gamma)$. Claim 19 shows that in polynomial time we obtain a solution $X \succeq 0$ satisfying (47), (46) with $\lambda \le OPT$, and $\max_{Y\in\mathcal{Q}}\langle Y, Z\rangle \le C_G(q)\kappa^2$. From (29), this implies $\|X\|_{q\to q^*} \le C_G(q)\kappa^2$.

Let $X = \sum_i \lambda_i v_i v_i^\top$ and let $S = \{i : \lambda_i \geq 1 - \delta\}$. For the rest of the analysis we will assume without loss of generality that $\|A\| = 1$. We first show the guarantee in (44). The projection output is just $\Pi_S = \sum_{i \in S} v_i v_i^\top$. Observe that $|S| \leq r/(1-\delta)$ from (47). Since the projector we output is just Π_S, each of its associated eigenvalues are at least $1 - \delta$. Hence, the operator norm bounds follows using the monotonicity of the norm since $\Pi_S \preceq \frac{1}{1-\delta} X$. To verify the objective value we see that

$$\Big\| \sum_{i \in [n]} (1 - \lambda_i) A^\top v_i v_i^\top A \Big\| = \Big\| A^\top (I - X) A \Big\| \leq \varepsilon^2 \|A\|^2 = \varepsilon^2$$

$$\Big\| \sum_{i \notin S} \delta A^\top v_i v_i^\top A \Big\| \leq \Big\| \sum_{i \notin S} (1 - \lambda_i) A^\top v_i v_i^\top A \Big\| \leq \Big\| \sum_{i \in [n]} (1 - \lambda_i) A^\top v_i v_i^\top A \Big\| \leq \varepsilon^2$$

$$\text{Hence } \Big\| A^\top \Pi_S^\perp A \Big\| \leq \frac{\varepsilon^2}{\delta} = \Big(1 + \frac{\gamma}{2}\Big)\varepsilon^2,$$

as required. We now show the guarantee in (43) where we output a projection of rank at most r (with no slack). Let $M' := \sum_{i \in S} A^\top v_i v_i^\top A$. Let Π' be the projection matrix for the subspace corresponding to the best rank r projection of M'. The algorithm outputs Π'.

Note that $\Pi' \preceq \Pi_S$, hence by monotonicity, the $q \to q^*$ operator norm constraint is satisfied up to a $\alpha := C_G$ factor. Also note that if Π^* is the projection that gives the optimal solution to the problem,

$$\|M' - A^\top \Pi^* A\| \leq \|M' - A^\top A\| + \|A^\top A - A^\top \Pi^* A\| \leq \varepsilon^2 + \frac{\varepsilon^2}{\delta} = \varepsilon^2(1 + \tfrac{1}{\delta}) = (2 + \tfrac{\gamma}{2})\varepsilon^2.$$

But $A^\top \Pi^* A$ is a valid approximation of M' of rank at most r. Hence, we have that

$$\|M' - \Pi' M' \Pi'\| \leq \varepsilon^2(1 + \tfrac{1}{\delta})$$

$$\|A^\top A - \Pi' A^\top A \Pi'\| \leq \|A^\top A - \Pi' \Pi_S A^\top A \Pi_S \Pi'\| \leq \|A^\top A - M'\| + \|M' - \Pi' M' \Pi'\|$$

$$\leq \frac{2\varepsilon^2}{\delta} + \varepsilon^2 = (1 + \tfrac{2}{\delta})\varepsilon^2 = (3 + \gamma)\varepsilon^2,$$

as required.

■

As before the same ideas also give the following more general theorem for any monotone matrix norm $|||\cdot|||$ that is approximately certifiable.

Theorem 20 *Let $|||\cdot|||$ be any matrix norm that is* monotone *and α-approximately certifiable for some $\alpha \geq 1$. Suppose the data matrix $A \in \mathbb{R}^{n \times m}$ has a projection Π^* of rank at most r such that $|||\Pi^*||| \leq \kappa$ and the approximation error $OPT := \|(I - \Pi^*)A\|$. There exists a polynomial time algorithm such that given any $\gamma \in (0, 1)$, it finds an orthogonal projection $\widehat{\Pi}$ of dimension at most r satisfying*

$$|||\widehat{\Pi}||| \leq \sqrt{\alpha(1 + 2/\gamma)} \cdot \kappa, \text{ and } \|(I - \widehat{\Pi})A\| \leq \sqrt{(3 + \gamma)} \cdot OPT. \tag{58}$$

Moreover, for any $\gamma \in (0, 1)$, there exists a polynomial time algorithm that finds an orthogonal projection $\widehat{\Pi}$ of rank $r' \leq (1 + \frac{2}{\gamma})r$ such that

$$|||\widehat{\Pi}||| \leq \sqrt{\alpha(1 + 2/\gamma)} \cdot \kappa, \text{ and } \|(I - \widehat{\Pi})A\| \leq \sqrt{(1 + \gamma)} \cdot OPT. \tag{59}$$

We omit the proof, since the ideas are identical to Theorem 18.

B.3. Recovering the Optimal Projection Matrix

We now show that if the optimal robust low-rank projection has very small error compared to the rth smallest singular value of A, then we can in fact approximately recover the subspace itself up to small error measured in terms of the principal angles. For two subspaces with projection matrices Π_1, Π_2, the Sin of the canonical angles matrix is given by $\Pi_1^{\perp}\Pi_2$. These techniques will also be helpful for recovery in the Spiked Covariance model. The following simple corollary will work for both Frobenius norm error and spectral norm error. For this purpose, we will just use $|||A|||$ to denote the norm of A, where the unspecified matrix norm $|||\cdot|||$ is norm in which we are measuring the error – either Frobenius norm or spectral norm.

Corollary 21 *Suppose the data matrix $A \in \mathbb{R}^{n\times m}$ has an r-dimensional projection Π^* such that $\|\Pi^*\|_{q\to 2} \le \kappa$, the approximation error $OPT := |||(I-\Pi^*)A|||^2 < \varepsilon^2|||A|||^2$ and $\sigma_r(\Pi^* A) \ge \theta$. There exists a polynomial time algorithm that finds a projection $\widehat{\Pi}$ of rank at most r such that*

$$|||\Pi^{\perp}\widehat{\Pi}||| \le O(1+\alpha)\cdot\frac{\varepsilon|||A|||}{\theta}. \tag{60}$$

where the subspace corresponding to Π is a subset of the subspace given by Π^* and α is the approximation factor attained by the algorithm in Theorem 11 (or Theorem 18).

Note that the above bound holds for the spectral norm error and the Frobenius norm error.

Proof The algorithm is exactly the same algorithm used in Theorem 11. Let Π denote the best robust low-rank subspace for A. We will then use the Davis-Kahan $\sin\Theta$ theorem about perturbations of singular vectors to show that the subspaces given by Π_1 and Π_2 are close. Note that the Davis-Kahan theorem states that if Π_i is the projection matrix onto eigenspaces of $A_iA_i^{\top}$ respectively ($i \in \{1,2\}$) with the least singular values of $\Pi_1 A_1$ being at least $\delta > 0$ more than the singular values of $\Pi_2^{\perp}A_2$, then for any unitarily invariant norm $|||\cdot|||$,

$$|||\Pi_2^{\perp}\Pi_1||| \le \frac{|||A_1 - A_2|||}{\delta}.$$

We would like to apply it with $A_2 = \widehat{\Pi}A, A_1 = \Pi^* A$ and $\Pi_2 = \widehat{\Pi}, \Pi_1 = \Pi^*$. We know that by the triangle inequality, for some constant α given by the approximation ratio in Theorem 11 (or Theorem 18),

$$|||\Pi^* A - \widehat{\Pi}A||| \le |||\Pi^* A - A||| + |||A - \widehat{\Pi}A||| \le \varepsilon|||A||| + \alpha\varepsilon|||A||| \le (\alpha+1)\varepsilon|||A|||, \tag{61}$$

where we used the fact that $\widehat{\Pi}$ gives an α-factor approximation to the objective. Moreover, in our case $A_1 = \Pi^* A$ is itself of rank-r and $\Pi_2^{\perp}A_2 = 0$. Under the stronger assumption in (60), we have $\sigma_r(\Pi^* A) \ge \theta$. Hence we see that (60) holds since

$$|||\widehat{\Pi}^{\perp}\Pi^*||| \le \frac{|||\Pi^* A - \widehat{\Pi}A|||}{\theta} \le \frac{(1+\alpha)\varepsilon}{\theta}.$$

■

Appendix C. Data Poisoning and Robustness to Adversarial Perturbations at Training Time

In this section, we consider training-time robustness, where the input matrix $\tilde{A}$ is an adversarial perturbation of A and the goal is to recover the robust projection of A rather than $\tilde{A}$. In Section C.1, We will study the approximation algorithms in Frobenius norm error. In Section C.2, we will study its counterpart in spectral norm error. Finally we show a lower bound in Section C.3.

C.1. Training-Time Robustness: Approximations in Frobenius Norm Error

Theorem 22 *Suppose $q \geq 2$ and $A \in \mathbb{R}^{n\times m}$ is the (unknown) uncorrupted data matrix, with an (orthogonal) projection matrix Π^* of rank at most r that is robust i.e., $\|\Pi^*\|_{q\to 2} \leq \kappa$ satisfying $\|A - \Pi^* A\|_F^2 \leq \varepsilon \|A\|_F^2$ for some $\varepsilon \in [0,1]$. There exists a polynomial time algorithm that given as input any (adversarially perturbed) data matrix $\tilde{A}$ s.t. for each column $j \in [m]$, $\|\tilde{A}_j - A_j\|_q \leq \delta$, outputs an orthogonal projection $\widehat{\Pi}$ of rank at most r such that for any $\eta > 0$*

$$\|\widehat{\Pi}\|_{q\to 2} \leq O(\kappa), \text{ and } \|A - \widehat{\Pi}A\|_F^2 \leq O(\varepsilon + \eta) \cdot \|A\|_F^2 + O(\tfrac{1}{\eta}) \cdot \delta^2 \kappa^2 m. \tag{62}$$

To get a multiplicative approximation we will set $\eta = O(\varepsilon)$, and get an extra additive term of $\delta^2\kappa^2 m/\varepsilon$. Here think of $\delta^2\kappa^2 \ll \frac{1}{m}\cdot\varepsilon\|A\|_F^2$. Further we remark that the above guarantees are optimal up to constant factors; in particular, the additive factor of $O(m\delta^2\kappa^2)$ is unavoidable (see Proposition 28).

The main challenge here is that while A has a good low-rank projection (in fact a robust one), $\tilde{A}$ may be very far from a rank-r matrix (let alone having a robust rank-r approximation). Further, the best robust low-rank approximation of $\tilde{A}$ could be very different from the best robust low-rank projection of A. This is because the entry-wise perturbations of δ could be too large in aggregate; for instance, it could be the case that $\|\tilde{A}\|_F^2 \gg \|A\|_F^2$. Suppose Π^* is the best robust low-rank projection of A. We will run the algorithm in the previous section not on the given matrix $\tilde{A}$, but on a suitably modified matrix A'.

Lemma 23 *There is a polynomial time algorithm that given any matrix $M \in \mathbb{R}^{n\times m}$, can find*

$$\Gamma_q(M) = \min_{\substack{B\in\mathbb{R}^{n\times m} \text{ s.t.} \\ \|B_j - M_j\|_q \leq \delta, \forall j\in[m]}} \|B\|_F^2,$$

up to arbitrary accuracy.

Proof First we note that since $\|B\|_F^2 = \sum_j \|B_j\|_2^2$, the optimization problem is separable across each of the m samples i.e.,

$$\min_{\substack{B\in\mathbb{R}^{n\times m} \text{ s.t.} \\ \|B_j - M_j\|_q \leq \delta, \forall j\in[m]}} \|B\|_F^2 = \sum_{j\in[m]} \min_{\substack{B_j\in\mathbb{R}^{n} \text{ s.t.} \\ \|B_j - M_j\|_q \leq \delta}} \|B_j\|_2^2$$

Input: $\tilde{A}$, the corrupted $n \times m$ data matrix, rank r, robustness parameter $\kappa \geq 1$ and norm $q \geq 2$.

1. Compute A' (using Lemma 23) such that
$$A' = \operatorname*{argmin}_{\substack{B \in \mathbb{R}^{n \times m} \text{ s.t.} \\ \|B_j - \tilde{A}_j\|_q \leq \delta, \forall j \in [m]}} \|B\|_F.$$
2. Run the algorithm from Theorem 11 on A', to obtain a rank-r projection matrix $\widehat{\Pi}$.
3. Output $\widehat{\Pi}$.

Figure 5: Robust rank-r approximations in Frobenius norm under adversarial perturbations during training.

We now describe how to solve each of the m subinstances corresponding to the column $j \in [m]$, which for a given $b \in \mathbb{R}^n$ is of the form

$$\min_{\substack{z \in \mathbb{R}^n \\ \|z\|_q \leq \delta}} \|b - z\|_2^2.$$

Note that the least-squares objective $\|b - z\|_2^2$ is convex. Moreover the constraint $\|z\|_q \leq \delta$ is also convex; further there is a simple separation oracle for this constraint since by duality

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

Hence by using the Ellipsoid algorithm, this problem can be solved in polynomial time. [5] ■

Note that when $q = \infty$, it is easy to find the matrix A', by just setting

$$A'_{ij} = \text{sign}(M_{ij}) \cdot \max\{0, |M_{ij}| - \delta\}, \quad \forall i, j \in [n].$$

We will argue that Π^* also gives a good low-rank approximation to A'. This crucially uses the fact that Π^* has bounded $\ell_q \to 2$ norm, which implies the following useful lemma.

Lemma 24 *Suppose $A, B \in \mathbb{R}^{n \times m}$ are two matrices such that for each column $j \in [m]$, $\|A_j - B_j\|_q \leq \delta$, and let Π be any rank-r projection matrix such that $\|\Pi\|_{q \to 2} \leq \kappa$. Then for any $\eta \in (0, 1)$,*

$$(1 - \eta)\|\Pi A\|_F^2 - (\tfrac{1}{\eta} - 1)\delta^2\kappa^2 m \leq \|\Pi B\|_F^2 \leq (1 + \eta)\|\Pi A\|_F^2 + (\tfrac{1}{\eta} + 1)\delta^2\kappa^2 m.$$

5. In fact Projected Gradient Descent Algorithm can also be used here; see Sra (2012).

Proof For each $j \in [m]$, let A_j, B_j be the jth columns of A and B respectively. Then $\|\Pi(A_j - B_j)\|_2 \le \|\Pi\|_{q\to 2}\|A_j - B_j\|_q \le \delta\kappa$. Using this along with the triangle inequality we get,

$$\begin{aligned}\|\Pi B\|_F^2 &= \sum_{j=1}^m \|\Pi(B_j - A_j) + \Pi A_j\|_2^2 \ge \sum_{j=1}^m (\|\Pi A_j\|_2 - \delta\kappa)^2 \\ &\ge \sum_{j=1}^m \|\Pi A_j\|_2^2 - 2\Big(\frac{\delta\kappa}{\sqrt{\eta}}\Big)(\sqrt{\eta}\|\Pi A_j\|_2) + (\delta\kappa)^2 \\ &\ge (1-\eta)\|\Pi A\|_F^2 - (\tfrac{1}{\eta} - 1)\delta^2\kappa^2 m, \text{ for any } \eta \in (0,1).\end{aligned}$$

This proves the first inequality. A similar argument also shows the other inequality. ■

We now prove that Algorithm 5 finds an approximately optimal robust low-rank projection for unknown, uncorrupted data matrix A.

Proof of Theorem 22. The first step of the algorithm finds the matrix A' given by

$$A' = \operatorname*{argmin}_{\substack{B\in\mathbb{R}^{n\times m} \text{ s.t.} \\ \|B_j - \tilde{A}_j\|_q \le \delta, \forall j\in[m]}} \|B\|_F^2.$$

Note that $\|A'\|_F \le \|A\|_F$ since A is also a feasible solution for the above minimization. Moreover since $\|A_j - A'_j\|_q \le 2\delta$ for each $j \in [m]$, we get from Lemma 24,

$$\|\Pi^* A'\|_F^2 \ge (1-\eta)\|\Pi^* A\|_F^2 - 4(\tfrac{1}{\eta} - 1)\delta^2\kappa^2 m, \quad \text{for any } \eta \in (0,1). \tag{63}$$

Now we run the algorithm from the previous section (Theorem 11) on A'. From Theorem 11 (with $\delta = 1/2$ say), we find a rank-r projection matrix Π with $\|\Pi\|_{\infty\to 2} \le O(\kappa)$ such that

$$\begin{aligned}\|A' - \Pi A'\|_F^2 &\le 3\Big(\|A'\|_F^2 - (1-\eta)\|\Pi^* A\|_F^2 + 4(\tfrac{1}{\eta} - 1)\delta^2\kappa^2 m\Big) \\ &\le 3\Big(\|A - \Pi^* A\|_F^2\Big) + 3\eta\|\Pi^* A\|_F^2 + 12(\tfrac{1}{\eta} - 1)\delta^2\kappa^2 m \\ &\le 3(\varepsilon + \eta)\|A\|_F^2 + 12(\tfrac{1}{\eta} - 1)\delta^2\kappa^2 m.\end{aligned}$$

However we know that $\|A'\|_F^2 \ge \|\Pi^* A'\|_F^2$. Hence

$$\begin{aligned}\|\Pi A'\|_F^2 &\ge \|A'\|_F^2 - \|A' - \Pi A'\|_F^2 \ge \|\Pi^* A'\|_F^2 - \|A' - \Pi A'\|_F^2 \\ &\ge (1-\eta)\|\Pi^* A\|_F^2 - 3(\varepsilon + \eta)\|A\|_F^2 - 16(\tfrac{1}{\eta} - 1)\delta^2\kappa^2 m \\ \text{Hence, } \|A - \Pi A\|_F^2 &= \|A\|_F^2 - \|\Pi A\|_F^2 \le_{\text{Lemma 24}} \|A\|_F^2 - (1-\eta)\|\Pi A'\|_F^2 + (\tfrac{1}{\eta} + 1)\delta^2\kappa^2 m \\ &\le \|A\|_F^2 - (1-\eta)^2\|\Pi^* A\|_F^2 + 3(\varepsilon + \eta)(1-\eta)\|A\|_F^2 \\ &\quad + m\delta^2\kappa^2\Big(1 + \tfrac{1}{\eta} + 16(1-\eta)(\tfrac{1}{\eta} - 1)\Big) \\ &\le \|A - \Pi^* A\|_F^2 + (3\varepsilon + 5\eta)\|A\|_F^2 + \Big(1 + \tfrac{17}{\eta}\Big)\delta^2\kappa^2 m \\ &\le O(\eta)\|A\|_F^2 + O(\tfrac{1}{\eta})\delta^2\kappa^2 m,\end{aligned}$$

for any $\eta \ge 4\varepsilon$. ■

C.2. Training-Time Robustness: Approximations in Spectral Norm Error

We now show guarantees for low-rank approximations in spectral norm error that are similar to Theorem 22. However, there is a qualitative difference: we will either find a robust low-dimensional projection of the unknown dataset A, or we will certify that the dataset has been poisoned substantially. In particular, the algorithm will *never* output a low-dimensional representation that is bad for the unknown data matrix A. We will later see how these guarantees also imply training-time robustness for downstream unsupervised learning applications like spectral clustering, robust mean estimation and learning mixture models. In what follows $\|\cdot\|$ will refer to the spectral norm.

Theorem 25 *Suppose $q \geq 2$ and $A \in \mathbb{R}^{n\times m}$ is the (unknown) uncorrupted data matrix, and let Π^* have the smallest spectral norm error $\|A - \Pi A\|$ among (orthogonal) projections of rank at most r that are robust i.e., $\|\Pi\|_{q\to 2} \leq \kappa$. There exists a polynomial time algorithm (Alg. 6) that given as input any (adversarially perturbed) data matrix $\tilde{A}$ s.t. for each column $j \in [m]$, $\|\tilde{A}_j - A_j\|_q \leq \delta$ and a parameter $\tau > 0$, outputs either a projection matrix $\widehat{\Pi}$ of rank at most r or outputs* Bad Input *s.t.*

(I) if the algorithm outputs a projection $\widehat{\Pi}$ of rank at most r, then it is a near-optimal robust low-rank approximation for the unknown matrix A i.e., for some small universal constant $c \geq 1$,

$$\forall \eta > 0,\ \ \|\widehat{\Pi}\|_{q\to 2} \leq c_q\kappa,\ \text{and}\ \|(I-\widehat{\Pi})A\| \leq O\Big(1+\tfrac{1}{\eta}\Big)\Big(\tau+\|A-\Pi^*A\|+\sqrt{m}\delta\kappa\Big)+\sqrt{2\eta}\|A\|. \tag{64}$$

(II) if the algorithm outputs Bad Input *, then either the data was poisoned i.e., $\|A-\tilde{A}\| > \tau$, or there is no good robust spectral norm approximation for A i.e., $\|A - \Pi A\| > \tau$ for all rank-r projection matrices Π s.t. $\|\Pi\|_{q\to 2} \leq \kappa$.*

In particular, if we are promised that A has a good robust projection Π^* of value $\|A-\Pi^*A\| \leq \varepsilon\|A\|$, then the algorithm either finds an approximately optimal robust projection $\widehat{\Pi}$ of rank at most r for A with

$$\|\widehat{\Pi}\|_{q\to 2} \leq c_q\kappa,\ \text{and}\ \forall \eta > 0,\ \ \|(I-\widehat{\Pi})A\| \leq O\Big(1+\tfrac{1}{\eta}\Big)\Big(\|A-\Pi^*A\|+\sqrt{m}\delta\kappa)+\sqrt{2\eta}\|A\|, \tag{65}$$

or certifies that the data has been poisoned i.e., $\|\tilde{A} - A\| > \varepsilon\|A\|$.

Our algorithm just runs the worst-case approximation algorithm from Theorem 18 on $\tilde{A}$ to find a projection $\widehat{\Pi}$. If the error is less than τ, it outputs $\widehat{\Pi}$; else it certifies that the data is corrupt.

The main feature of the above algorithm is that it is *always correct*. The algorithm certifies that the input is Bad only when the data has been poisoned i.e., $\tilde{A}$ is substantially far from A, or A did not have a good robust low-rank approximation to begin with. More crucially, when it does output a projection matrix $\widehat{\Pi}$, it is guaranteed to be a valid robust projection[6] for the unknown matrix A. We remark that the additive error term of $\Omega(\delta\kappa\sqrt{m})$ is unavoidable here information-theoretically; see Proposition 28 for an example.

6. In particular it rules out the scenario where the algorithm finds a solution that it thinks is good (on $\tilde{A}$), but is in fact bad for the unknown, uncorrupted matrix A.

Input: $\tilde{A}$, the corrupted $n \times m$ data matrix, tolerance parameter $\tau > 0$, rank r, robustness parameter $\kappa \geq 1$ and norm $q \geq 2$.

1. Run the algorithm from Theorem 18 on $\tilde{A}$, to obtain a rank-r projection matrix $\widehat{\Pi}$.
2. If the robust low-rank approximation error on $\tilde{A}$, $\|\tilde{A} - \widehat{\Pi}\tilde{A}\| \leq \tau$, output $\widehat{\Pi}$.
3. Otherwise output Bad Input .

Figure 6: Robust rank-r approximations in Spectral norm error under adversarial perturbations in training.

The following is the key lemma that argues that if the projection $\widehat{\Pi}$ gives a small error on $\tilde{A}$, it necessarily gives a low-error on A.

Lemma 26 *Let $\delta \in \mathbb{R}_+$ and $A, B \in \mathbb{R}^{n \times m}$ such that $\|A-B\|_q \leq \delta$. Let Π_1, Π_2 be projection matrices such that $\|\Pi_1\|_{q\to 2}, \|\Pi_2\|_{q\to 2} \leq \kappa$, and $= \|A - \Pi_1 A\| \leq \varepsilon_1$ and $\|B - \Pi_2 B\| \leq \varepsilon_2$. Then we have that for any $\eta \in (0,1)$,*

$$\|A - \Pi_2 A\| \leq O\Big(1 + \tfrac{1}{\eta}\Big)\Big(\varepsilon_1 + \varepsilon_2 + \sqrt{m}\delta\kappa\Big) + \sqrt{2\eta}\|A\|, \tag{66}$$

$$\text{and } \|B - \Pi_1 B\| \leq O\Big(1 + \tfrac{1}{\eta}\Big)\Big(\varepsilon_1 + \varepsilon_2 + \sqrt{m}\delta\kappa\Big) + \sqrt{2\eta}\|B\|. \tag{67}$$

Proof The projection matrices Π_1, Π_2 are both robust. For $\ell \in \{1, 2\}$

$$\|\Pi_\ell A - \Pi_\ell B\|^2 \leq \|\Pi_\ell(A - B)\|_F^2 = \sum_{j \in [m]} \|\Pi_\ell(A_j - B_j)\|_2^2 \leq m\kappa^2\delta^2$$

$$\text{Hence } \Big|\|\Pi_\ell A\| - \|\Pi_\ell B\|\Big| \leq \sqrt{m}\kappa\delta. \tag{68}$$

Let $\gamma := \sqrt{m}\delta\kappa$. We also know that $\|A - \Pi_1 A\| \leq \varepsilon_1$.

$$\|A - \Pi_1 B\| \leq \|A - \Pi_1 A\| + \|\Pi_1 A - \Pi_1 B\| \leq \varepsilon_1 + \gamma$$

$$\text{Hence } \forall v \in \mathbb{S}^{n-1},\ \|Av - \Pi_1 Bv\|_2 \leq \varepsilon_1 + \gamma, \text{ and similarly } \|Bv - \Pi_2 Av\|_2 \leq \varepsilon_2 + \gamma. \tag{69}$$

But $Bv = \Pi_1 Bv + \Pi_1^\perp Bv$. We have for any $\eta \in (0,1)$

$$\begin{aligned}\|Bv\|_2^2 &= \|\Pi_1 Bv\|_2^2 + \|\Pi_1^\perp Bv\|_2^2 \geq (\|Av\|_2 - \varepsilon_1 - \gamma)^2 + \|\Pi_1^\perp Bv\|_2^2 \\ &\geq (1-\eta)\|Av\|_2^2 - (\varepsilon_1 + \gamma)^2(1 + \tfrac{1}{\eta}) + \|\Pi_1^\perp Bv\|_2^2 \\ \text{Similarly, } \|Av\|_2^2 &\geq (1-\eta)\|Bv\|_2^2 - (\varepsilon_2 + \gamma)^2(1 + \tfrac{1}{\eta}) + \|\Pi_2^\perp Av\|_2^2\end{aligned}$$

Combining the two, we get that

$$\begin{aligned}
\|Bv\|_2^2 &\geq (1-\eta)^2\|Bv\|_2^2 - (1+\tfrac{1}{\eta})\Big((\varepsilon_1+\gamma)^2 + (\varepsilon_1+\gamma)^2\Big) \\
&\quad + (1-\eta)\|\Pi_2^{\perp}Av\|_2^2 + \|\Pi_1^{\perp}Bv\|_2^2 \\
(1-\eta)\|\Pi_2^{\perp}Av\|_2^2 + \|\Pi_1^{\perp}Bv\|_2^2 &\leq (2\eta-\eta^2)\|Bv\|_2^2 + (1+\tfrac{1}{\eta})\Big((\varepsilon_1+\gamma)^2 + (\varepsilon_1+\gamma)^2\Big) \\
\forall v \in \mathbb{S}^{n-1}, \quad \|\Pi_1^{\perp}Bv\|_2^2 &\leq 2\eta\|Bv\|_2^2 + (1+\tfrac{1}{\eta})\Big((\varepsilon_1+\gamma)^2 + (\varepsilon_1+\gamma)^2\Big). \\
\text{Hence, } \|B-\Pi_1 B\|^2 &\leq 2\eta\|B\|^2 + (1+\tfrac{1}{\eta})(\varepsilon_1+2\gamma+\varepsilon_2)^2,
\end{aligned}$$

as required. A similar statement also follows for A using a symmetric proof. ■

Proof [Proof of Theorem 25] Firstly the algorithm from Theorem 18 runs on $\tilde{A}$ and produced a robust projection matrix $\widehat{\Pi}$. The proof consists of two parts. We first argue that if the algorithm outputs any robust rank-r projection matrix, then it has to be robust for A. Any such $\widehat{\Pi}$ satisfies $\|\tilde{A} - \widehat{\Pi}\tilde{A}\| \leq \tau$. Applying Lemma 26 with $\varepsilon_2 = \tau$ ($B = \tilde{A}$) and $\varepsilon_1 = \|A - \Pi^* A\|$, we have

$$\|A - \widehat{\Pi}A\| \leq O\Big(1+\tfrac{1}{\eta}\Big)\Big(\tau + \|A-\Pi^*A\| + \sqrt{m}\delta\kappa\Big) + \sqrt{2\eta}\|A\|.$$

On the other hand, if the input $\tilde{A}$ is not "Bad" i.e., (a) for the unknown matrix A, $\|A - \Pi^* A\| \leq \tau$, and (b) $\|A - \tilde{A}\| \leq \tau$, we now show that the algorithm outputs a good solution for A. In this case we have that $\|\tilde{A} - \Pi^* A\| \leq 2\tau$; hence,

$$\|\tilde{A} - \Pi^*\tilde{A}\| \leq \|\tilde{A} - \Pi^* A\| + \|\Pi^*\tilde{A} - \Pi^* A\| \leq 2\tau + \sqrt{\sum_{j\in[m]} \|\Pi^* A_j - \Pi^*\tilde{A}_j\|_2^2} \leq 2\tau + \sqrt{m}\kappa\delta.$$

Hence, by Lemma 26 applied with $\varepsilon_1 = \tau$ and $\varepsilon_2 = (B = \tilde{A})$, we have that

$$\|A - \widehat{\Pi}A\| \leq O\Big(1+\tfrac{1}{\eta}\Big)\Big(\tau + \sqrt{m}\delta\kappa\Big) + \sqrt{2\eta}\|A\|.$$

This proves the theorem. The moreover part follows by setting $\tau := \varepsilon\|A\|$.

■

In fact, Lemma 26 implies a stronger information-theoretic statement about finding a robust low-rank approximation of the unknown, uncorrupted matrix A with low spectral norm (just like Theorem 22 for Frobenius norm error). In fact we get a polynomial time algorithm assuming access to a polynomial time algorithm approximation algorithm for solving the following problem: given a matrix $\tilde{A} \in \mathbb{R}^{n\times m}$, find[7]

$$\min_{B:\|B_j - \tilde{A}_j\|_q \leq \delta \ \forall j\in[m]} \quad \min_{\Pi:\mathrm{rank}(\Pi)=r,\ \|\Pi\|_{q\to 2}\leq\kappa} \|B - \Pi B\|^2, \tag{70}$$

where $\|\cdot\|$ stands for the spectral norm.

7. This problem is reminiscent of the concept of ε-rank (Alon et al., 2013), that corresponds to the smallest rank attainable by changing every entry of the given matrix by at most δ.

Proposition 27 *Suppose $q \geq 2$ and $A \in \mathbb{R}^{n\times m}$ is the (unknown) uncorrupted data matrix, and let Π^* have the smallest spectral norm error $\|A - \Pi A\|$ among rank-r projections that are robust i.e., $\|\Pi\|_{q\to 2} \leq \kappa$. Suppose further that there is an efficient algorithm for finding an α-factor approximation algorithm for (70). Then there exists an algorithm that runs in polynomial time, and given as input any (adversarially perturbed) data matrix $\tilde{A}$ s.t. for each column $j \in [m]$, $\|\tilde{A}_j - A_j\|_q \leq \delta$ and a parameter $\tau > 0$, outputs a robust projection matrix $\widehat{\Pi}$ of rank at most r that is near optimal in approximation error for the unknown matrix A i.e., for some small universal constant $c \geq 1$,*

$$\forall \eta > 0, \quad \|(I - \widehat{\Pi})A\| \leq O\Big(1 + \tfrac{1}{\eta}\Big)\Big(\alpha\|A - \Pi^* A\| + \sqrt{m}\delta\kappa\Big) + \sqrt{2\eta}\|A\|. \tag{71}$$

Moreover, the above bound is achieved information-theoretically by an algorithm (that potentially does not have polynomial running time), by using an inefficient algorithm for problem (70).

We remark that the main difference between the above proposition and Theorem 25 is that Proposition 27 will always output a good robust projection for A (just like Theorem 22 for Frobenius norm error), but the algorithm is not computationally efficient unless (70) can be solved efficiently.

Proof Given $\tilde{A}$, the algorithm first runs the α-factor approximation algorithm for solving (70) on $\tilde{A}$. The uncorrupted matrix A is itself a feasible solution; hence the solution output by the algorithm A' has a robust low-rank approximation of error $O(\alpha)\|A - \Pi^* A\|$. Such a robust low-rank projection $\widehat{\Pi}$ for A' i.e., a projection for rank at most r with $\|\widehat{\Pi}\|_{q\to 2} \leq O(\kappa)$ and $\|A' - \widehat{\Pi}A'\| \leq O(\alpha)\|A - \Pi A\|$ can be found by running Theorem 18 on A'. Moreover A' and A are valid 2δ adversarial perturbations of each other. Now applying Lemma 26 with A, Π^* and $A', \widehat{\Pi}$ completes the proof. ■

C.3. Lower Bound for the Additive Error in Training with Adversarial Perturbations

We now show that the additive terms of $\Omega(m\delta^2\kappa^2)$ in Theorem 22 is unavoidable.

Proposition 28 *For any data matrix A with the following two properties:*

1. *Each column $\|A_j\|_2 \in [1/10, 10]$,*
2. *There exists Π^* of rank 1 and $\|\Pi^*\|_{\infty\to 2} \geq \kappa$ (which is at least 2) satisfying $\Pi^* A = A$,*

there exists δ_0 (depending on A) such that for any $\delta \leq \delta_0$, there exist A' as a δ-perturbation of A (i.e., $\|A - A'\|_\infty \leq \delta$) and a projection matrix Π' of rank 1 satisfying

1. *Π' is robust $\|\Pi'\|_{\infty\to 2} \leq \|\Pi^*\|_{\infty\to 2}$.*
2. *We still have $\Pi' A' = A'$ but $\|A - \Pi' A\|_F = \Omega(\delta\kappa\sqrt{m})$. Since $A - A'$ is of rank 2, this also implies a similar lower bound for the spectral norm.*

When A is a k-sparse flat matrix where every entry is either 0 or $\Theta(1/\sqrt{k})$, δ_0 is as large as $\Theta(1/\sqrt{k})$.

Proof Let v be the unit eigenvector of Π^* such that $\|v\|_1 \geq \kappa$. Without loss of generality, we assume $|v_1| \geq |v_2| \cdots \geq |v_n|$ and $\ell = \lfloor \text{supp}(v)/2 \rfloor$. Notice that $\text{supp}(v) \geq 2\ell$ by this definition such that $v_{2\ell} \neq 0$. At the same time, because of the Cauchy-Schwartz inequality, we have $\|v\|_1^2 \leq \text{supp}(v) \cdot \|v\|_2^2$ so that $\ell \geq \kappa^2/2 - 1$.

Then we set $\delta_0 = |v_{2\ell}|$ and consider any δ less than it. We perturb v to another "sparser" vector u whose coordinate-wise absolute values are given by

$$\Big(|v_1| + \delta, \ldots, |v_\ell| + \delta, |v_{\ell+1}| - \delta, \ldots, |v_{2\ell}| - \delta, |v_{2\ell+1}|, \ldots, |v_n|\Big).$$

However, since v_i may be negative or positive, we define u according to the sign function:

$$u = \Big(v_1 + \text{sign}(v_1)\cdot\delta, \ldots, v_\ell + \text{sign}(v_\ell)\cdot\delta, v_{\ell+1} - \text{sign}(v_{\ell+1})\cdot\delta, \ldots, v_{2\ell} - \text{sign}(v_{2\ell})\cdot\delta, v_{2\ell+1}, \ldots, v_n\Big).$$

We have $\|u\|_1 = \sum_{i=1}^{\ell} |v_i| + \delta + \sum_{i=\ell+1}^{2\ell} |v_i| - \delta + \sum_{i>2\ell} v_i = \|v\|_1$ and

$$\begin{aligned}\|u\|_2^2 &= (|v_1| + \delta)^2 + \cdots + (|v_\ell| + \delta)^2 + (|v_{\ell+1}| - \delta)^2 + \cdots + (|v_{2\ell}| - \delta)^2 + v_{2\ell+1}^2 + \cdots + v_n^2 \\ &= \sum_i v_i^2 + 2(\sum_{i=1}^{\ell} |v_i| - \sum_{i=\ell+1}^{2\ell} |v_i|)\delta + 2\ell\delta^2.\end{aligned}$$

Since $|v_1| \geq |v_2| \geq \cdots \geq |v_n|$, this is at least $\sum_i v_i^2 + 2\ell \cdot \delta^2 > \|v\|_2^2$. So let $\overline{u} = u/\|u\|_2$ with unit ℓ_2 norm and $\Pi' = \overline{u} \cdot \overline{u}^\top$. So $\|\Pi'\|_{\infty\to 2} = \|\overline{u}\|_1 < \|v\|_1 = \|\Pi^*\|_{\infty\to 2}$.

Next we consider A. Since $\Pi^* A = A$, we assume $A = [c_1 \cdot v, c_2 \cdot v, \ldots, c_m \cdot v]$ with coefficient $|c_i| \leq [1/10, 10]$. We set $A' = [c_1 \cdot u, \ldots, c_m \cdot u]$ such that $\|A - A'\|_\infty \leq 10\delta$ and $\Pi' A' = A'$.

Finally we lower bound $\|A - \Pi' A\|_F^2$. Notice that

$$\langle u, v\rangle = \sum_{i=1}^{\ell}(v_i^2 + |v_i|\delta) + \sum_{i=\ell+1}^{2\ell}(v_i^2 - |v_i|\delta) + \sum_{i>2\ell} v_i^2 = 1 + (\sum_{i=1}^{\ell}|v_i| - \sum_{i=\ell+1}^{2\ell}|v_i|)\delta.$$

Thus $\Pi' v = \langle u, v\rangle u/\|u\|_2^2 = \alpha u$ for $\alpha = \frac{1+(\sum_{i=1}^{\ell}|v_i| - \sum_{i=\ell+1}^{2\ell}|v_i|)\delta}{1+(\sum_{i=1}^{\ell}|v_i| - \sum_{i=\ell+1}^{2\ell}|v_i|)\delta + 2\ell\delta^2} < 1$. So we lower bound the distance between $v - \Pi' v$ by counting the ℓ entries from $v_{\ell+1}$ to $v_{2\ell}$:

$$\sum_{i=\ell+1}^{2\ell}[v_i - \alpha(v_i - \text{sign}(v_i)\delta)]^2 = \sum_{i=\ell+1}^{2\ell}[(1-\alpha)|v_i| + \alpha\delta]^2.$$

Since $\delta \leq |v_{2\ell}|$, each term in the summation is at least δ^2. So $\|A - \Pi' A\|_F = \sqrt{c_1^2 + \cdots + c_m^2} \cdot \delta \cdot \sqrt{\ell} = \Omega(\delta\kappa\sqrt{m})$. ■

Appendix D. Robustness to Adversarial Perturbations in Clustering

As a concrete application of our guarantees for resilience to adversarial perturbations, we study the problem of clustering under adversarial perturbations. Here our goal is to approximately recover the clusters of a well defined ground truth clustering, as well as good approximations to the cluster centers. Our main result is to apply the guarantee from Theorem 25 to show how to perform clustering of a well-clustered instance when every data point in the instance could be corrupted. Our guarantees will apply to clustering a mixture of well separated Gaussians and more general data distributions. In particular, we will show that a robust modification of the popular Lloyd's algorithm (Lloyd, 1982) (also known as the k-means algorithm) can be used to perform clustering in our model, thereby providing further evidence towards the widespread applicability of the algorithm. Existing guarantees for using Lloyd's algorithm (Kumar and Kannan, 2010; Awasthi and Sheffet, 2012) for clustering a mixture of Gaussians and general datasets assume that every pair of means μ_i, μ_j are separated by $\sim \sigma\sqrt{k}$, where σ is the maximum variance of the dataset around the mean and k is the number of clusters (see (72) for the formal condition). In the presence of adversarial perturbations of magnitude δ, even if the optimal clustering (according to unperturbed ground truth) of the perturbed data is provided to us, the best we can hope for is to estimate the cluster means up to an error that goes to zero with δ.

D.1. Overview of clustering results

Let $A \in \mathbb{R}^{n\times m}$ be clustered into k clusters of equal sizes with means $\mu_1, \mu_2, \ldots, \mu_k$. Furthermore, let $C \in \mathbb{R}^{n\times m}$ be the matrix of corresponding centers for each column of A and let σ be such that $\|A - C\| \leq \sigma\sqrt{m}$. Then A satisfies c-spectral stability if for each pair of optimal clusters, say, cluster r and s with means μ_r and μ_s, any point in cluster r, when projected onto the line joining μ_r and μ_s is closer to μ_r than μ_s by an additive amount of $\Delta_{r,s} := c\alpha k \cdot \sigma$. Here α is a quantity that captures the signal-to-noise ratio and the relative perturbation magnitude.[8] When A is a set of $m = \text{poly}(n, k)$ points drawn i.i.d. from a mixture of Gaussians with the variance of each Gaussian being bounded by σ^2, and with uniform mixture weight $1/k$ each, the separation condition becomes $\Delta_{r,s} = c\alpha k \cdot \text{polylog}(nk) \cdot \sigma$. Below we denote κ to be the robustness, as measured in $\|\cdot\|_{q\to 2}$, of the subspace spanned by the true means $\{\mu_1, \mu_2, \ldots, \mu_k\}$.

(Informal) Theorem 29 *[Robust Clustering] Fix $q \geq 2$, and let c_q be a constant that depends on q. Let $A \in \mathbb{R}^{n\times m}$ satisfy c-spectral stability, for $c > 200c_q$. Then given as input a δ-corrupted instance $\tilde{A}$ of A, there is a Lloyd's style algorithm that either certifies that the dataset is poisoned, i.e, $\|A - \tilde{A}\| = \Omega(\sigma\sqrt{m})$, or recovers each mean μ_r up to error $O(\alpha\sqrt{k}\sigma)$. Using the computed centers to cluster $\tilde{A}$, we obtain a clustering of $\tilde{A}$ such that the corresponding induced clustering on A that misclassifies $O(1/k)$- fraction of the points.*

In the special case of a mixture of Gaussians with equal mixing weights we recover the means upto error $\tilde{O}(\alpha\sigma)$, where we hide a polylog(m, n) factor in the $\tilde{O}$ notation. This implies $O(1/k^2)$-fraction clustering error.

8. We show that unlike standard clustering, the dependence on α is unavoidable with corruptions even for $k = 1$ (mean estimation).

See Theorem 31 and Theorem 33 for formal statements that also handle more general cluster sizes and mixing weights. Finally, as in Section 2.1.2 we can also prove that there is an algorithm (though computationally inefficient) that can cluster well-clustered instances up to the claimed error above, without the need for certification. Whether this can be achieved in polynomial time is an open question.

Our analysis proceeds in three stages: a) an initialization stage, b) a center improvement stage, and c) analyzing the robust Lloyd's updates. Each stage poses unique challenges arising from working with $\tilde{A}$ where each data point is potentially corrupted. The standard way to initialize Lloyd's algorithm via PCA[9] can be arbitrarily bad when every data point is corrupted. Using our algorithm for spectral norm error from Section 2.1.2 we instead project the data onto a robust k-dimensional subspace Π for $\tilde{A}$ with small error, or certify that the dataset has been poisoned. This then lets us compute initial centers that are $O(\alpha k\sigma)$-close to the true means.

In the second stage we improve the initial center estimates by a $\sqrt{k}$ factor. Our main technical contribution here is to establish a stronger version of the statements that appear in Kumar and Kannan (2010); Awasthi and Sheffet (2012) (see Lemma 35). This lemma simultaneously helps us argue about the clustering error, and also the variance of each current cluster around its mean, a quantity crucial to bound in order to analyze the iterative updates later. The first two stages together help us establish the guarantee for general well-clustered instances.

To establish the stronger guarantee for mixtures of Gaussians we first analyze the "ideal" iterative updates, as if we had access to the uncorrupted data. This largely follows the analysis in Kumar and Kannan (2010) and helps us argue that if the current center estimates are $\beta\alpha\sqrt{k}\sigma$ close to the corresponding means (where $\beta < 1$), then in the next step the ideal updates give estimates that are $\frac{\beta}{4}\alpha\sqrt{k}\sigma$ close. A key technical step is to show that when performing ideal updates, the variance of the formed clusters around their means is bounded even though the clusters themselves are impure! Using the bounded variance property, we next analyze the actual updates and use a specialized robust mean estimation procedure (Lemma 30) to get an estimate that is within $\tilde{O}(\alpha\sigma) + \frac{\beta}{2}\alpha\sqrt{k}\sigma$ of the true mean μ_r. Hence, the updates will keep improving until the unavoidable error of $\tilde{O}(\alpha\sigma)$.

Guarantees for mean estimation. For purpose of recovering the means we will crucially rely on a subroutine to robustly estimate the mean of a cluster. We sketch below this procedure and state the associated guarantee.

Lemma 30 *Let A be an $n \times m$ matrix representing m data points in n dimensions and let μ be a vector such that $\|\textsc{Mean}(A) - \mu\|_2 \leq \eta$. Let C be the $n \times m$ matrix with each column being μ. Let $\Pi^* = \mu\mu^\top/\|\mu\|^2$ be the one dimensional subspace denoting the projection onto μ and assume that $\|\Pi^*\|_{q\to 2} \leq \kappa$, for some $q \geq 2$. Let $\tilde{A}$ be the given input such that for every column $j \in [m]$ we have $\|A_j - \tilde{A}_j\|_q \leq \delta$. Furthermore, let $\sigma^2 > 0$ be a given upper bound on the variance of the data around μ, i.e., $\|A - C\| \leq \sigma\sqrt{m}$. Then the algorithm from Figure 7 when run on $\tilde{A}$, runs in polynomial time, and either certifies that the data has been poisoned, i.e., $\|\tilde{A} - A\| = \Omega(\sigma\sqrt{m})$, or outputs an estimate $\hat{\mu}$ of the true mean μ*

9. This initialization is needed for theoretical bounds. In practice, the initial centers are chosen as random data points.

RobustMean$(\tilde{A}, \kappa, \sigma^2)$
Input: The corrupted data matrix $\tilde{A} \in \mathbb{R}^{n \times m}$ with columns $\tilde{A}_i$ for $i \in [m]$, and the upper bound κ on the robustness of the subspace $\|\Pi^*\|_{q \to 2}$.

1. Run the algorithm from Figure 6 with $\tau = 2\sigma\sqrt{m}, r = 1$, and κ.
2. If the algorithm outputs Bad Input then terminate and certify that the data has been poisoned. Otherwise let Π be the 1-dimensional subspace output by the algorithm. Return $\hat{\mu} = \textsc{Mean}(\Pi\tilde{A})$.

Figure 7: Robust Mean Estimation.

such that

$$\|\hat{\mu} - \mu\|_2 \leq \eta + O(c_q)(1 + \frac{\kappa\delta}{\sigma}) \max\left(\sigma, \sqrt{\sigma\|\mu\|}\right),$$

where c_q is a constant that depends on q. In particular, the above implies a relative error guarantee of

$$\frac{\|\hat{\mu} - \mu\|_2}{\|\mu\|} \leq \eta + O(c_q)(1 + \frac{\kappa\delta}{\sigma}) \max\left(\frac{\sigma}{\|\mu\|}, \sqrt{\frac{\sigma}{\|\mu\|}}\right).$$

We provide the proof of the above lemma in Appendix H.4. In the appendix we also provide a matching lower bound stating that in general the above bound on estimation error cannot be improved. See also subsequent work (Awasthi et al., 2020a) for other algorithms for robust mean estimation, without the need for certification. However, the associated guarantees in Awasthi et al. (2020a) are incomparable and typically have a multiplicative dependence on η which is not desirable for the clustering application. In the above lemma we make use of the fact that the data has a small projection onto Π to get a stronger additive guarantee, or certify that the data has been poisoned.

Guarantees for k-means clustering. From the above discussion, in the context of clustering, even if one is given the original optimal clustering of the given perturbed dataset, we must incur a loss of $\Omega(\sigma \cdot \max(1, \sqrt{\|\mu\|/\sigma}))$ in simply estimating the true mean μ of a cluster. This suggests a separation condition of the type $\sim \alpha\sigma\sqrt{k}$, where α depends on $(1 + \frac{\mu_{\max}}{\sigma})$ and the guarantee to aim for is to estimate means upto error $O(\alpha\sigma)$ error. Here $\mu_{\max}$ is the maximum ℓ_2 norm of the any of the k mean vectors. In this section we will show that a modified Lloyd's combined with our certification procedure can indeed achieve this guarantee or certify that the dataset has been poisoned.

More formally, we will assume that there is a set of m points in $\mathbb{R}^n$ with ground truth clustering $C_1^*, C_2^*, \ldots, C_k^*$, and means $\mu_r = \textsc{Mean}(C_r^*)$ for $r \in [k]$ and $\mu_{\max} = \max_r \|\mu_r\|$. Let A be the $n \times m$ data matrix and C be the matrix of corresponding centers. We will assume that we have an upper bound σ^2 on the maximum variance of the data points around their mean, i.e. $\|A - C\|^2 \leq \sigma^2 m$ and define $\alpha = (1 + \frac{\kappa\delta}{\sigma})(1 + \frac{\mu_{\max}}{\sigma})^{2/3}$. We will enforce the spectral stability condition studied in Kumar and Kannan (2010) on our clustering instance.

This condition implies that for each pair of clusters C_r^*, C_s^* with means μ_r, μ_s and each point $x \in C_r^*$, $\bar{x}$ is closer to μ_r than to μ_s by a margin $\Delta_{r,s}$. Here $\bar{x}$ is the projection of x onto the line joining μ_r and μ_s. For a constant $c > 0$, the c-spectral stability condition requires that for each $r \neq s$,

$$\Delta_{r,s} \geq c\alpha\sigma\sqrt{k}\Big(\frac{\sqrt{m}}{\sqrt{|C_r^*|}} + \frac{\sqrt{m}}{\sqrt{|C_s^*|}}\Big) \tag{72}$$

Notice that the above also implies that every pair of means are separated i.e.,

$$\|\mu_r - \mu_s\| \geq c\alpha\sigma\sqrt{k}\Big(\frac{\sqrt{m}}{\sqrt{|C_r^*|}} + \frac{\sqrt{m}}{\sqrt{|C_s^*|}}\Big).$$

It is worth mentioning that in the typical analysis of Lloyd's algorithm (Kumar and Kannan, 2010; Awasthi and Sheffet, 2012) the dependence on α in the separation condition is not needed. However, as discussed before, in our noise model, some dependence on α is unavoidable to get a meaningful clustering guarantee.

Assumptions I: Fix $q \geq 2$. We will assume that we are given access to $\tilde{A}$ such that for every $j \in [m]$, $\|A_j - \tilde{A}_j\|_q \leq \delta$. Furthermore, define κ to be the robustness, of the subspace spanned by the means $\mu_1, \ldots, \mu_k$. Formally, let Π_C be the projection matrix for the orthogonal projection onto the space of the means. Then κ is such that $\|\Pi_C\|_{q\to 2} \leq \kappa$. Under Assumptions I, we prove the following theorem that applies to any stable dataset as defined in (72).

Theorem 31 *Fix $q \geq 2$, and let c_q be a constant that depends on q. Let A be a dataset that satisfies c-spectral stability for $c > 200c_q$. Under Assumptions I, there is a Lloyd's style algorithm that takes $\tilde{A}$ as input, runs in polynomial time, and either certifies that the dataset is poisoned, i.e, $\|A - \tilde{A}\| = \Omega(\sigma\sqrt{m})$, or outputs a clustering $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_k$ and means $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k$ such that*

$$\sum_{r=1}^{k} |C_r^* \triangle \hat{C}_{\pi(r)}| \leq O\Big(\frac{c_q^2 m}{k\alpha^2 c^2}\Big)$$

$$\|\mu_r - \hat{\mu}_{\pi(r)}\| \leq c_q\alpha\frac{\sigma\sqrt{m}}{\sqrt{|C_r^*|}}.$$

for an appropriately chosen bijection $\pi : [k] \to [k]$.

While the above theorem works for any data set that satisfies spectral stability, notice that it leads to a sub optimal mean estimation error of $O(\alpha\sigma\sqrt{m}/\sqrt{|C_r^*|})$ for each cluster r. For example, when the clusters are balanced, this will lead to a guarantee of $O(\alpha\sigma)\sqrt{k}$. Next, we show that for data sets that additionally satisfy Gaussian type concentration, we can indeed get $O(\alpha\sigma)$ estimation error even when each data point is corrupted.

Assumptions II: Let A be a given dataset with optimal clustering $C_1^*, C_2^*, \ldots, C_k^*$. We will assume that we are given $\tilde{A}$ that satisfies Assumptions I. Furthermore, we will assume that $\|C_r^*\| \geq n^3$ for each $r \in [k]$ and that for any subset $S \subset C_r^*$ of points such that $|S| > n \log n$, we have that $\|A_S - C_S\| \leq \sigma\sqrt{|S|} \cdot \operatorname{poly}\log(m, n)$. Here A_S, C_S are the matrices A and C restricted to the columns of the points in S. Additionally, we require a pointwise guarantee

that for each $r \in [k]$, and $A_i \in C_r^*$, $\|A_i - \mu_r\|^2 \leq 2\sigma^2 n \cdot \operatorname{poly}\log(m,n)$. It is easy to see that $m \geq \operatorname{poly}(m, 1/w_{\min})$ samples generated from a mixture of Gaussians with maximum variance σ^2 and minimum mixture weight $w_{\min}$ will, with high probability, satisfy the above assumptions. Under Assumptions II, we prove the following theorem that applies to any stable dataset as defined in (72).

Theorem 32 *Fix $q \geq 2$, and let c_q be a constant that depends on q. Let A be a dataset that satisfies c-spectral stability for $c > 200c_q$. Under Assumptions II, there is a Lloyd's style algorithm that takes $\tilde{A}$ as input, runs in polynomial time, and either certifies that the dataset is poisoned, i.e, $\|A - \tilde{A}\| = \Omega(\sigma\sqrt{m})$, or outputs a clustering $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_k$ and means $\hat{\mu_1}, \hat{\mu_2}, \ldots, \hat{\mu_k}$ such that*

$$\sum_{r=1}^{k} |C_r^* \triangle \hat{C}_{\pi(r)}| \leq O\Big(\frac{c_q^2 m}{k^2\alpha^2 c^2}\Big)$$
$$\|\mu_r - \hat{\mu}_{\pi(r)}\| \leq \tilde{O}(\alpha\sigma).$$

for an appropriately chosen bijection $\pi : [k] \to [k]$, where we hide a polylogarithmic (in n, m) factor in the $\tilde{O}$ notation.

As a corollary we get the following statement about robustly clustering a mixture of Gaussians.

Theorem 33 *Fix $q \geq 2$, and let c_q be a constant that depends on q. Define $\mathcal{M}$ to be a distribution that is a mixture of k Gaussians, i.e., $\mathcal{M} := \sum_{r=1}^{k} w_r \mathcal{N}(\mu_r, \Sigma_r)$. Furthermore, let $\Sigma_r \preceq \sigma^2 I$ and define $w_{\min} = \min_r w_r$ and $\mu_{\max} = \max_r \|\mu_r\|$, $\alpha = (1 + \frac{\kappa\delta}{\sigma})(1 + \frac{\mu_{\max}}{\sigma})^{2/3}$. Let A be a set $\operatorname{poly}(n, 1/w_{\min})$ samples generated i.i.d. from the mixture. If the mixture if well separated, i.e, $\|\mu_r - \mu_s\| \geq c\alpha\sigma\sqrt{k} \cdot \operatorname{poly}\log(n/w_{min})/\sqrt{w_{\min}}$ for $c > 200c_q$, and the means span a κ robust subspace, then given access to $\tilde{A}$ such that $\|\tilde{A}_j - A_j\|_q \leq \delta$, there is a Lloyd's style algorithm that, runs in polynomial time, and either certifies that the data is poisoned, i.e., $\|A - \tilde{A}\| \geq \Omega(\sigma\sqrt{m})$, or outputs a clustering $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_k$ and means $\hat{\mu_1}, \hat{\mu_2}, \ldots, \hat{\mu_k}$ such that*

$$\sum_{r=1}^{k} |C_r^* \triangle \hat{C}_{\pi(r)}| \leq O\Big(\frac{c_q^2 m}{k^2\alpha^2 c^2}\Big)$$
$$\|\mu_r - \hat{\mu}_{\pi(r)}\| \leq \tilde{O}(\alpha\sigma).$$

for an appropriately chosen bijection $\pi : [k] \to [k]$.

Computing Good Initial Centers. The first step in establishing the above theorems is to compute centers/means that are close to the true ones. A common approach for this step is to use PCA to project the data onto the top-k subspace of the input data matrix, and run any constant factor approximation algorithm for k-means clustering (Kumar and Kannan, 2010). However this can be arbitrarily bad if the data is corrupted as in our model. We instead show that by projecting the data onto a robust subspace as output by our guarantee from Theorem 25 and then using a k-means approximation algorithm, we do indeed recover good centers. Our initialization algorithm is shown in Figure 8. We next provide proofs for our claims.

Input: The corrupted data matrix $\tilde{A} \in \mathbb{R}^{n \times m}$ with columns $\tilde{A}_i$ for $i \in [m]$, upper bound κ on the robustness of the subspace $\|\Pi_C\|_{q \to 2}$, upper bound σ^2 on the data variance $\|A - C\|^2/m$.

1. Run the algorithm from Figure 6 with $\tau = 2\sigma\sqrt{m}, r = k$, and κ.
2. If the algorithm outputs Bad Input then certify that the data has been poisoned and terminate.
3. If the algorithm outputs a projection matrix Π then project $\tilde{A}$ onto Π. Denote $\hat{A} = \Pi\tilde{A}$ as the projected matrix.
4. Run a 10-approximation algorithm for k-means clustering on $\hat{A}$, and obtain k centers $\nu_1, \nu_2, \ldots, \nu_k$.
5. Output $\nu_1, \nu_2, \ldots, \nu_k$.

Figure 8: Computing initial center estimates.

Theorem 34 *Assume that the clustering instance A is c-stable for $c > 200c_q$. If Assumptions I hold, then the algorithm in Figure 8 runs in polynomial time, and either certifies that the data has been poisoned, i.e., $\|A - \tilde{A}\| > 2\sigma\sqrt{m}$, or the algorithm outputs centers $\nu_1, \nu_2, \ldots, \nu_k$ such that for all $r \in [k]$,*

$$\|\mu_r - \nu_{\pi(r)}\| \leq 30c_q\alpha\sqrt{k}\frac{\sigma\sqrt{m}}{\sqrt{|C_r^*|}}.$$

for an appropriately chosen bijection π.

Proof The proof will follow the general outline of Lemma 5.1 of Kumar and Kannan (2010), except that we need to argue following two stronger conditions. Firstly, we need to establish that $\tilde{A}$ projected on to Π has cost comparable to that of $k\sigma^2 m$. This will ensure that the approximation algorithm will output a clustering of low cost. Secondly, we also simultaneously need to establish that $\tilde{A}$ when projected on to Π has low cost clustering when true means C are used to cluster it. Together with the fact that $\tilde{A}$ and A are pointwise close in the projected space, we can then claim that missing out on a good approximation for even a single cluster center of C^* will incur a cost of $\Omega(k\sigma^2 m)$, thereby contradicting the approximation guarantee of the k-means algorithm used in step 2.

Establishing that $\Pi\tilde{A}$ has low cost with respect to C boils down to showing that Π is good for A given that it is good for $\tilde{A}$, a perturbation of A. This statement, established in Theorem 25 is the key in analyzing the initialization phase, and is a generalization of Lemma 56 to higher dimensional subspaces. Let's first establish that $\tilde{A}$ projected on to Π

has a low cost clustering. We have

$$\begin{aligned}\|\Pi\tilde{A} - \Pi C\|_F &\leq \sqrt{3k}\|\Pi\tilde{A} - \Pi C\| \text{ (since both } \Pi\tilde{A} \text{ and } \Pi C \text{ have rank at most } k)\\ &\leq \sqrt{3k}\Big(\|\Pi(\tilde{A} - A)\| + \|\Pi(A - C)\|\Big)\\ &\leq \sqrt{3k}(c_q\delta\kappa\sqrt{m} + \|A - C\|)\\ &\leq 2\sqrt{3}c_q\sqrt{k}\sigma\sqrt{m} \leq c_q(1 + \frac{\kappa\delta}{\sigma})\sqrt{12k}\sigma\sqrt{m}. \end{aligned} \tag{73}$$

Here the third inequality follows from the fact that for any $n\times m$ matrix M, $\|M\| \leq \ell_{\max}\sqrt{m}$, where $\ell_{\max}$ is the maximum ℓ_2 norm of a column of M. Furthermore, from the robustness of Π we know that for any $j \in [m]$,

$$\|\Pi(\tilde{A}_j - A_j)\| \leq c_q\kappa\delta.$$

Next, let's establish that $\|A - \Pi A\|$ is small. By triangle inequality we know that

$$\begin{aligned}\|A - \Pi_C A\| &\leq \|A - C\| + \|C - \Pi_C A\|\\ &= \|A - C\| + \|\Pi_C(C - A)\|\\ &\leq 2\|A - C\| \leq 2\sigma\sqrt{m}.\end{aligned}$$

Furthermore, from the guarantee of Theorem 25 we have that for any $\eta \in (0, 1)$,

$$\begin{aligned}\|A - \Pi A\| &\leq O(1 + \frac{1}{\eta})\Big(2\sigma\sqrt{m} + \|A - \Pi_C A\| + \kappa\delta\sqrt{m}\Big) + \sqrt{2\eta}\|A\|\\ &\leq O(1 + \frac{1}{\eta})\Big(4\sigma + \kappa\delta\Big)\sqrt{m} + \sqrt{2\eta}\|A\|.\end{aligned}$$

Setting $\eta = (5\sigma\sqrt{m}/\|A\|)^{2/3}$ we get that

$$\begin{aligned}\|A - \Pi A\| &\leq 4c_q(1 + \frac{\kappa\delta}{\sigma})\sigma\Big(1 + \Big(\frac{\|A\|}{\sigma\sqrt{m}}\Big)^{2/3}\Big)\sqrt{m}\\ &\leq 4c_q\alpha\sigma\sqrt{m}.\end{aligned} \tag{74}$$

The last inequality above follows from the fact that

$$\begin{aligned}\|A\| &\leq \|A - C\| + \|C\|\\ &\leq \sigma\sqrt{m} + \mu_{\max}\sqrt{m}.\end{aligned}$$

Next notice that

$$\begin{aligned}\|\Pi\tilde{A} - C\|_F &\leq \sqrt{3k}\|\Pi\tilde{A} - C\| \text{ (since both } \Pi\tilde{A} \text{ and } C \text{ have rank at most } k)\\ &\leq \sqrt{3k}\Big(\|\Pi(\tilde{A} - A)\| + \|\Pi A - A\| + \|A - C\|)\\ &\leq \sqrt{3k}(c_q\delta\kappa\sqrt{m} + 5c_q\alpha\sigma\sqrt{m} + \sigma\sqrt{m})\\ &\leq 6\sqrt{3k}c_q\alpha\sigma\sqrt{m}.\end{aligned} \tag{75}$$

Now we are ready to establish the claim of the theorem. From (73) we get that the centers $\nu_1, \nu_2, \ldots, \nu_k$ will have k-means cost at most $120k(1+\frac{\kappa\delta}{\sigma})^2 c_q^2\sigma^2 m$ on $\tilde{A}$. Furthermore, suppose that there exists a center μ_r such that every ν_s is far from it. For any point A_i, let $\nu_{c(i)}$ be the center in the set $\{\nu_1, \nu_2, \ldots, \nu_k\}$ that is closest to the projection of $\tilde{A}_i$ on to Π. Then we have that

$$\begin{aligned}
120kc_q^2(1+\frac{\kappa\delta}{\sigma})^2\sigma^2 m \geq \sum_{A_i\in C_r}\|\Pi\tilde{A}_i - \nu_{c(i)}\|^2 &= \sum_{A_i\in C_r}\|\Pi\tilde{A}_i - \mu_r + \mu_r - \nu_{c(i)}\|^2 \\
&\geq \frac{1}{2}|C_r|\|\mu_r - \nu_{c(i)}\|^2 - \sum_{A_i\in C_r}\|\Pi\tilde{A}_i - \mu_r\|^2 \\
&\geq \frac{1}{2}|C_r|\|\mu_r - \nu_{c(i)}\|^2 - \|\Pi\tilde{A} - C\|_F^2 \\
&\geq 450k\alpha^2 c_q^2\sigma^2 m - \|\Pi\tilde{A} - C\|_F^2 \\
&> 120kc_q^2\alpha^2\sigma^2 m. \qquad (76)
\end{aligned}$$

Noticing that $\alpha \geq (1+\frac{\kappa\delta}{\sigma})$, we get a contradiction to the fact that μ_r is far from every ν_s. This combined with the fact that the clustering instance is c-stable for $c > 200c_q$ implies that one can find a bijection $\pi : [k] \mapsto [k]$ between $\{\mu_1, \ldots, \mu_k\}$ and $\{\nu_1, \ldots, \nu_k\}$ such that each μ_i is close to a unique $\nu_{\pi(i)}$. ■

D.2. Analyzing Lloyd's Updates

Next we will use the obtained initial centers and run the robust Lloyd's algorithm starting with these centers as shown in Figure 9. Our goal in this section is to analyze the updates and establish Theorem 31 and Theorem 32.

Overview of Analysis and Challenges. Our analysis of the modified Lloyd's updates proceeds in two stages: a) a center improvement step, and b) analyzing robust Lloyd's updates. In (a), we first improve the initial center estimates obtained form the initialization phase to get estimates $\nu_1^{(1)}, \ldots, \nu_k^{(1)}$ such that each $\nu_r^{(1)}$ is $\sim \Delta_r/2$-close to the corresponding μ_r, where $\Delta_r = 40c_q\alpha\sigma\sqrt{m}/\sqrt{|C_r^*|}$. In other words, we get a factor $\sqrt{k}$ improvement over the initial estimates. This is sketched in step 3 of the algorithm in Figure 9. First, we motivate the need for this intermediate step, since it is not necessary in the analysis of Lloyd's algorithm for uncorrupted data.

Just as in standard analysis of Lloyd's updates, we would like to argue that if we have non-trivial estimates of the centers, as obtained from the initialization stage, forming clusters using these points and moving to the means of these clusters will improve the center estimates. To argue this we will crucially rely on the fact that when projected onto Π, A and $\tilde{A}$ are close pointwise. Hence, we can come up with a charging argument to assign mistakes made by the current centers on the uncorrupted points to the mistakes made by the centers on the corrupted points. We can then bound the number of such mistakes by observing that on $\Pi\tilde{A}$, the true means have a small k-means cost. This forces us to work in the projected space Π, but as a result inherently limits the accuracy to which we can obtain center estimates. Notice that if the initialization algorithm outputs Π, then Π is guaranteed to be good overall for $\tilde{A}$, in the sense that $\|\tilde{A} - \Pi\tilde{A}\| = O(\sigma\sqrt{m})$. However, Π

Input: The corrupted data matrix $\tilde{A} \in \mathbb{R}^{n \times m}$ with columns $\tilde{A}_i$ for $i \in [m]$, upper bound κ on the robustness of the subspace $\|\Pi_C\|_{q \to 2}$, upper bound σ^2 on the data variance $\|A - C\|^2/m$.

1. Let Π be the robust k dimensional projection matrix as computed by the initialization algorithm in Figure 8.
2. Define initial center estimates $\nu_1^{(0)}, \nu_2^{(0)}, \ldots, \nu_k^{(0)}$ to be the centers output by the algorithm in Figure 8.
3. For each $r \in [k]$, define $S_r = \{\Pi\tilde{A}_i : \|\Pi\tilde{A}_i - \nu_r\| \leq \frac{\|\Pi\tilde{A}_i - \nu_s\|}{3}, \forall s \neq r\}$. Update $\nu_r^{(1)} = \textsc{Mean}(S_r)$.
4. For $t = 2, 3, \ldots$ do:
   (a) For each $r \in [k]$, compute $S_r = \{\tilde{A}_i : \|\Pi\tilde{A}_i - \nu_r^{(t-1)}\| \leq \|\Pi\tilde{A}_i - \nu_s^{(t-1)}\|, \forall s \neq r\}$.
   (b) For each $r \in [k]$, update $\nu_r^{(t)} = \textsc{RobustMean}(S_r, \kappa, 4\sigma)$ //If RobustMean outputs Bad Input then certify that data is poisoned.

Figure 9: Iterative Updates of the Lloyd's Algorithm.

has no per cluster guarantee, and in general $\|\tilde{A}_r - \Pi\tilde{A}_r\|$ when restricted to a cluster C_r^* could be as large as $\sigma\sqrt{m}/\sqrt{|C_r^*|}$. Hence, to achieve our goal of estimating the centers upto $\tilde{O}(\alpha\sigma)$ accuracy, we also need to work outside of the projection Π at the same time. Due to these conflicting demands, notice that the Lloyd's updates we analyze in step 4 of the algorithm in Figure 9 perform clustering using current centers in the projected space, but perform robust mean estimation on the original input data.

Furthermore, from our guarantee on robust mean estimation in Theorem 30, we know that in the RobustMean step of the algorithm the centers will be accurate upto $\sim \alpha\sigma_S$, where σ_S is the standard deviation of the uncorrupted data points in S_r around the uncorrupted mean of S_r. As a result we need a stronger argument that not only shows that we have low clustering error given the current estimates, but also helps us argue about the variance of the formed clusters S_r at each step. Such an argument (Lemma 35) is a main technical contribution in the analysis.

Unfortunately, the argument (Lemma 35) only kicks in when we have much better center estimates than the one provided by the initialization stage, thereby requiring an additional center improvement stage. To argue about the center improvement stage, we use a trick from Awasthi and Sheffet (2012) and form sets S_r that correspond to points in $\Pi\tilde{A}$ that are significantly close to one of the centers $\nu_r^{(0)}$ than any other center $\nu_s^{(0)}$. Notice that these sets do not form a partitioning of the data. We then argue that any mistake made by this assignment must have also been made if one had used the true centers $\mu_1, \ldots, \mu_k$, to cluster $\Pi\tilde{A}$. Using the fact that the true means have small k-means cost on $\Pi\tilde{A}$ we can bound the number of such mistakes and hence get sets S_r that have low error, thereby helping us show that the means of these sets will be much closer to the true centers. This is established in

Theorem 36. The above arguments help us establish Theorem 31. We next state the key technical lemma for our analysis.

Lemma 35 *Let Π be the robust subspace computed in step (1) of the algorithm in Figure 9. For each cluster C_r^* in the optimal clustering of A, define $\Delta_r = 40c_q\alpha\sigma\sqrt{m}/\sqrt{|C_r^*|}$. Suppose we have center estimates $\{\nu_1, \nu_2, \ldots, \nu_k\}$ such that $\|\nu_r - \mu_r\| \leq \beta\Delta_r$, $\forall r \in [k]$, and some $\beta < 1$. When using ν_is to cluster $\Pi\tilde{A}$, define $T_{r,s}$ to be the set of points that are misclassified, w.r.t. the induced clustering on A, i.e., $T_{s\to r} = \{i : A_i \in C_r^*$ and $\|\Pi\tilde{A}_i - \nu_s\| \leq \|\Pi\tilde{A}_i - \nu_r\|\}$. There exists a constant $c_1 > 0$ depending on q such that if the clustering instance is c-stable for $c > 200c_q$ then we have that $|T_{s\to r}| \leq \frac{c_1\beta^2\sigma^2 m}{kc^2\|\mu_r - \mu_s\|^2}$.*

Proof Fix $s \neq r$ and let W be the subspace spanned by $\{\mu_r, \mu_s, \nu_r, \nu_s\}$ with Π_W being the projection matrix for the orthogonal projection on to the subspace. Define $\bar{A}_i$ to be the projection of A_i onto the line joining μ_r and μ_s. Since W contains μ_r, μ_s, this is also the same as the projection of $\Pi_W A_i$ on to the line joining μ_r and μ_s. Similarly, define $\bar{\tilde{A}}_i$ to be the projection of $\tilde{A}_i$ on to the line joining μ_r and μ_s, and again this is the same as the projection of $\Pi_W\tilde{A}_i$ on to the line joining μ_r and μ_s. We will crucially make use of the fact that

$$\|\bar{\tilde{A}}_i - \mu_s\| - \|\bar{\tilde{A}}_i - \mu_r\| \geq \Delta_{r,s} - O(\kappa\delta) \geq \Delta_{r,s}/2. \tag{77}$$

The above holds since from c-stability we know that $\|\bar{A}_i - \mu_s\| - \|\bar{A}_i - \mu_r\| \geq \Delta_{r,s}$. Furthermore, since $\|\tilde{A}_i - A_i\|_q \leq \delta$ and each of μ_r, μ_s is κ-sparse in ℓ_{q^*} norm, we have that $\|\bar{\tilde{A}}_i - \bar{A}_i\| \leq O(\kappa\delta)$. Here q^* is such that $1/q + 1/q^* = 1$. Next, let $v = \Pi_W\tilde{A}_i$. Then we have that

$$\begin{aligned}\|v - \mu_s\|^2 - \|v - \mu_r\|^2 &= \|\bar{\tilde{A}}_i - \mu_s\|^2 - \|\bar{\tilde{A}}_i - \mu_r\|^2 \\ &\geq \frac{\Delta_{r,s}\|\mu_r - \mu_s\|}{4} \text{ (using the fact that } \bar{\tilde{A}}_i \text{ lies on the line joining } \mu_r \text{ and } \mu_s).\end{aligned} \tag{78}$$

By triangle inequality we also have that,

$$\begin{aligned}\|v - \mu_s\|^2 - \|v - \mu_r\|^2 &\leq (\|v - \nu_s\| + \beta\Delta_s)^2 - (\|v - \nu_r\| - \beta\Delta_r)^2 \\ &\leq (\|v - \nu_r\| + \beta\Delta_s)^2 - (\|v - \nu_r\| - \beta\Delta_r)^2 \\ &\leq \beta(\Delta_s + \Delta_r)\|v - \nu_r\|.\end{aligned} \tag{79}$$

Here the first inequality uses the fact that ν_r, ν_s are close to μ_r, μ_s respectively and the second inequality uses the fact that $\tilde{A}_i$ is closer to ν_s than to ν_r, the same holds true for $\tilde{A}_i$ projected on to any subspace that contains ν_r and ν_s. From (78) and (79), and substituting the bound for $\Delta_{r,s}$ we get that $\|v - \nu_r\| \geq \frac{c\sqrt{k}\|\mu_r - \mu_s\|}{10c_q\beta}$, which in turn implies that $\|v - \mu_r\| \geq c\sqrt{k}\frac{\|\mu_r - \mu_s\|}{8c_q\beta}$. Hence we get that

$$\sum_{i\in T_{s\to r}} \|\Pi_W\tilde{A}_i - \mu_r\|^2 \geq |T_{s\to r}|\frac{c^2k\|\mu_r - \mu_s\|^2}{64c_q^2\beta^2}. \tag{80}$$

Combining with the fact that $\|\mu_r - \nu_r\| \leq \beta\Delta_r$ we get that

$$\sum_{i\in T_{s\to r}} \|\Pi_W \tilde{A}_i - \nu_r\|^2 \geq |T_{s\to r}| \frac{c^2 k \|\mu_r - \mu_s\|^2}{128 c_q^2 \beta^2}. \tag{81}$$

On the other hand we also have that

$$\begin{aligned}
\sum_{i\in T_{s\to r}} \|\Pi_W \tilde{A}_i - \nu_r\|^2 &\leq \sum_{i\in T_{s\to r}} 2\|\Pi_W \tilde{A}_i - \mu_r\|^2 + 2|T_{s\to r}|\|\mu_r - \nu_r\|^2 \text{ (by triangle inequality)} \\
&= \sum_{i\in T_{s\to r}} 2\|\Pi_W \tilde{A}_i - \Pi_W\mu_r\|^2 + 2|T_{s\to r}|\|\mu_r - \nu_r\|^2 \text{ (since } \mu_r \text{ lies in } \Pi_W) \\
&\leq \sum_{i\in T_{s\to r}} 2\|\Pi_W \tilde{A}_i - \Pi_W\mu_r\|^2 + 2|T_{s\to r}|\beta^2\Delta_r^2 \\
&\leq \sum_{i\in T_{s\to r}} 4\|\Pi_W \tilde{A}_i - \Pi_W\Pi\mu_r\|^2 + \sum_{i\in T_{s\to r}} 4\|\Pi_W(\Pi\mu_r - \mu_r)\|^2 + 2|T_{s\to r}|\beta^2\Delta_r^2 \\
&\leq \sum_{i\in T_{s\to r}} 4\|\Pi_W \tilde{A}_i - \Pi_W\Pi\mu_r\|^2 + 4|T_{s\to r}|\|\Pi\mu_r - \mu_r\|^2 + 2|T_{s\to r}|\beta^2\Delta_r^2,
\end{aligned} \tag{82}$$

where the last but one line also uses triangle inequality. Next notice that

$$\begin{aligned}
\|\Pi\mu_r - \mu_r\| &= \frac{1}{|C_r^*|}\|\sum_{A_i\in C_r}(\Pi A_i - A_i)\| = \frac{1}{|C_r^*|}\|\mathbf{1}^\top(\Pi A - A)\| \\
&\leq \frac{1}{\sqrt{|C_r^*|}}\|\Pi A - A\| \leq \frac{4c_q\alpha\sigma\sqrt{m}}{\sqrt{|C_r^*|}} = \frac{2}{5}\Delta_r.
\end{aligned}$$

Substituting into (82) we get that

$$\begin{aligned}
\sum_{i\in T_{s\to r}} \|\Pi_W \tilde{A}_i - \nu_r\|^2 &\leq \sum_{i\in T_{s\to r}} 4\|\Pi_W \tilde{A}_i - \Pi_W\Pi\mu_r\|^2 + |T_{s\to r}|(\frac{16}{5} + 2\beta^2)\Delta_r^2 \\
&= \sum_{i\in T_{s\to r}} 4\|\Pi_W\Pi^\top\Pi\tilde{A}_i - \Pi_W\Pi^\top\Pi^2\mu_r\|^2 + |T_{s\to r}|(\frac{16}{5} + 2\beta^2)\Delta_r^2 \\
&= \sum_{i\in T_{s\to r}} 4\|\Pi_W\Pi^\top(\Pi\tilde{A}_i - \Pi\mu_r)\|^2 + |T_{s\to r}|(\frac{16}{5} + 2\beta^2)\Delta_r^2 \\
&\leq 4\|\Pi_W\Pi^\top(\Pi\tilde{A} - \Pi C)\|_F^2 + |T_{s\to r}|(\frac{16}{5} + 2\beta^2)\Delta_r^2 \\
&\leq 16\|\Pi\tilde{A} - \Pi C\|^2 + |T_{s\to r}|(\frac{16}{5} + 2\beta^2)\Delta_r^2
\end{aligned}$$

since $\Pi_W\Pi^\top(\Pi\tilde{A} - \Pi C)$ has rank at most 4. Hence

$$\sum_{i\in T_{s\to r}} \|\Pi_W \tilde{A}_i - \nu_r\|^2 \leq 32c_q^2(1 + \frac{\kappa\delta}{\sigma})^2\sigma^2 m + |T_{s\to r}|(\frac{16}{5} + 2\beta^2)\Delta_r^2. \tag{83}$$

The last inequality uses the fact that

$$\|\Pi\tilde{A} - \Pi C\| \leq \|\Pi(\tilde{A} - A)\| + \|\Pi A - \Pi C\| \leq \kappa\delta\sqrt{m} + \|A - C\| \leq \kappa\delta\sqrt{m} + \sigma\sqrt{m}.$$

Combining, (81) and (83) we get the desired claim. ■

In order to apply Lemma 35 iteratively we need initial centers such that $\|\mu_r - \nu_r\| \leq \beta\Delta_r$, for $\beta \leq \frac{1}{4}$. However, notice that the the initialization procedure of Figure 8 only guarantees $\beta \leq 30c_q\sqrt{k}$. We next argue that step (3) of the algorithm in Figure 9 provides center estimates that are much closer to the true means, thereby allowing us to analyze the iterative Lloyd's updates in step (4) of the algorithm.

Theorem 36 *If the clustering instance A is c-stable as defined in Theorem 34, then given $\tilde{A}$ as input, steps* 1-3 *of the Algorithm in Figure 9, run in polynomial time, and output centers $\nu_1^{(1)}, \ldots, \nu_k^{(1)}$ such that*

$$\forall r \in [k], \quad \|\mu_r - \nu_{\sigma(r)}^{(1)}\| \leq \beta\Delta_r,$$

for an appropriately chosen bijection σ. Here $\Delta_r = 40c_q\alpha\sigma\sqrt{m}/\sqrt{|C_r^*|}$ and $\beta < 1$.

Proof The proof strategy closely follows the one in Awasthi and Sheffet (2012) and consists of three main steps. We first define clusters T_r for $r \in [k]$ such that T_r consists of points $\Pi\tilde{A}_j$ for $A_j \in C_r^*$. In other words, $\{T_1, T_2, \ldots, T_k\}$ is the clustering induced on the data set $\Pi\tilde{A}$ by the optimal clustering $\{C_1^*, C_2^*, \ldots, C_k^*\}$. We first argue that S_r is pure w.r.t. T_r i.e., at most $O(\frac{1}{c^2}|C_r^*|)$ points of T_r do not belong to S_r and in total at most $O(\frac{1}{k}|C_r^*|)$ points from T_s, for $s \neq r$, end up belonging to S_r. Next use the fact that any points that belongs to $|S_r \cap T_s|$ for $s \neq r$, will also be misclassified when using centers $\Pi\tilde{\mu}_1, \ldots, \Pi\tilde{\mu}_k$ instead of centers $\nu_1^{(0)}, \ldots, \nu_k^{(0)}$. Here $\tilde{\mu}_r = \textsc{Mean}(T_r)$. Now each projected center $\Pi\tilde{\mu}_r$ is much closer to the corresponding true center μ_r. To see this notice that

$$\begin{aligned}
\|\Pi\tilde{\mu}_r - \mu_r\| &= \frac{1}{|C_r^*|}\Big\|\sum_{\Pi\tilde{A}_i \in T_r}(\Pi\tilde{A}_i - A_i)\Big\| \\
&\leq \frac{1}{|C_r^*|}\Big\|\sum_{\Pi\tilde{A}_i \in T_r}(\Pi(\tilde{A}_i - A_i))\Big\| + \frac{1}{|C_r^*|}\Big\|\sum_{\Pi\tilde{A}_i \in T_r}(\Pi A_i - A_i)\Big\| \\
&\leq c_q\kappa\delta + \frac{1}{|C_r^*|}\|\mathbf{1}^\top(\Pi A - A)\| \leq c_q\kappa\delta + \frac{1}{\sqrt{|C_r^*|}}\|\Pi A - A\| \\
&\leq c_q\kappa\delta + \frac{4c_q\alpha\sigma\sqrt{m}}{\sqrt{|C_r^*|}} \leq \frac{\Delta_r}{9}.
\end{aligned}$$

With the above idea, arguing that $|T_s \cap S_r|$ is small and T_r has large overlap with S_r follows verbatim from Lemmas 4.2 and 4.3 of Awasthi and Sheffet (2012) by substituting $\Pi\tilde{A}_i$ instead of $\hat{A}_i$ in the proofs. In the final step we use the following standard fact stated in Lemma 37 below and adapted from its original version in Awasthi and Sheffet (2012); Kumar and Kannan (2010). From the guarantees on $|T_s \cap S_r|$ and $|T_r \cap S_r|$ we can set $\rho_{out} = \frac{1}{8}$ and $\rho_{in} = c/10k$ to get that

$$\|\textsc{Mean}(S_r) - \textsc{Mean}(\hat{C}_r)\| \leq 2\frac{\sigma\sqrt{m}}{\sqrt{|C_r^*|}}.$$

Furthermore we also have that

$$\begin{aligned}\|\textsc{Mean}(\hat{C}_r) - \mu_r\| &= \frac{1}{|C_r^*|}\|\sum_{\Pi A_i \in C_r^*}(\Pi A_i - A_i)\| \\ &= \frac{1}{|C_r^*|}\|\mathbf{1}^\top(\Pi A - A)\| \leq \frac{1}{\sqrt{|C_r^*|}}\|\Pi A - A\| \\ &\leq \frac{4c_q\alpha\sigma\sqrt{m}}{\sqrt{|C_r^*|}} \leq \frac{\Delta_r}{10}.\end{aligned}$$

Combining the above two we get that

$$\|\mu_r - \nu_r^{(1)}\| \leq O(\beta\Delta_r),$$

for $\beta < 1$. ■

Lemma 37 (Fact 1.3 from Awasthi and Sheffet (2012)) *Fix a target cluster C_r^* and let $\hat{C}_r$ be the projection of points in C_r^* onto Π. Let S_r be a set of points created by removing $\rho_{out}|C_r^*|$ points from $\hat{C}_r$ and adding $\rho_{in}|C_s^*|$ points from each cluster $\hat{C}_s$ for $s \neq r$, s.t. every added point x satisfies $\|x - \Pi\mu_s\| \geq \frac{2}{3}\|x - \Pi\mu_r\|$. If $\rho_{out} < 1/4$ and $\rho_{in} := \sum_{s\neq r}\rho_{in} < 1/4$ then we have that*

$$\|\textsc{Mean}(S_r) - \textsc{Mean}(\hat{C}_r)\| \leq 2\Big(\sqrt{\frac{\rho_{out}}{|C_r^*|}} + \frac{3\sqrt{k}}{2}\sqrt{\frac{\rho_{in}}{|C_r^*|}}\Big)\sigma\sqrt{m} \tag{84}$$

Proof [Proof of Theorem 31] The theorem follows from using steps 1-3 of the algorithm in Figure 9 and from the guarantees in Lemma 35 and Lemma 36. ■

Achieving $\tilde{O}(\alpha\sigma)$ Guarantee for Mean Estimation.

Proof [Proof of Theorem 32] Notice that Theorem 36 gives us centers $\nu_1, \ldots, \nu_k$ that are $\beta\Delta_r$ close to the corresponding true centers $\mu_1, \ldots, \mu_k$. We start with these centers and perform Lloyd's updates as shown in step 4 of the algorithm in Figure 9. Next suppose that at iteration t we have centers $\nu_1^{(t)}, \ldots, \nu_k^{(t)}$ such that $\|\nu_r^{(t)} - \mu_r\| \leq \beta\Delta_r$ for $r \in [k]$. We will argue that using $\nu_1^{(t)}, \ldots, \nu_k^{(t)}$ to form clusters $S_1, S_2, \ldots, S_r$ and computing new means by calling the RobustMean procedure on the sets S_r, either leads to a certification that the dataset is poisoned or leads to new centers estimates $\nu_1^{(t+1)}, \ldots, \nu_k^{(t+1)}$ that satisfy $\|\nu_r^{(t+1)} - \mu_r\| \leq \Big(\frac{\beta}{2}\Delta_r + \tilde{O}(\alpha\sigma)\Big)$. Hence the estimates will improve until the unavoidable error of $\tilde{O}(\alpha\sigma)$. We will prove the claim in two steps. First we analyze the "ideal" updates. For each S_r define S_r^* to the set S_r with corrupted points replace by the original points, i.e., $S_r^* = \{A_i : \tilde{A}_i \in S_r\}$. We next show that the mean of S_r^* is close to μ_r upto $\frac{\beta}{2}\Delta_r$ error. As in Lemma 35 define $T_{r\to r} = S_r^* \cap C_r^*$ and for $s \neq r$, define $T_{r\to s} = S_r^* \cap C_s^*$. Then we have

by triangle inequality that,

$$\begin{aligned}
\|\textsc{Mean}(S_r^*) - \mu_r\| &= \Big\| \frac{|T_{r\to r}|}{|S_r^*|}(\textsc{Mean}(T_{r\to r}) - \mu_r) + \sum_{s\neq r} \frac{|T_{r\to s}|}{|S_r^*|}(\textsc{Mean}(T_{r\to s}) - \mu_r) \Big\| \\
&\leq \frac{|T_{r\to r}|}{|S_r^*|}\Big\|\textsc{Mean}(T_{r\to r}) - \mu_r\Big\| + \sum_{s\neq r} \frac{|T_{r\to s}|}{|S_r^*|}\Big\|\textsc{Mean}(T_{r\to s}) - \mu_r\Big\| \\
&\leq \frac{|T_{r\to r}|}{|S_r^*|}\Big\|\textsc{Mean}(T_{r\to r}) - \mu_r\Big\| + \sum_{s\neq r} \frac{|T_{r\to s}|}{|S_r^*|}\Big\|\textsc{Mean}(T_{r\to s}) - \mu_s\Big\| \quad (85) \\
&\quad + \sum_{s\neq r} \frac{|T_{r\to s}|}{|S_r^*|}\Big\|\mu_r - \mu_s\Big\| \quad (86)
\end{aligned}$$

Next we notice that

$$\begin{aligned}
\Big\|\textsc{Mean}(T_{r\to r}) - \mu_r\Big\| &= \frac{|C_r^* \setminus T_{r\to r}|}{|T_{r\to r}|}\Big\| \sum_{A_i \in C_r^* \setminus T_{r\to r}} (A_i - \mu_r)\Big\| \\
&\leq \frac{\sqrt{|C_r^* \setminus T_{r\to r}|}}{|T_{r\to r}|}\sigma\sqrt{m} \\
&= \frac{\sqrt{|C_r^*| - |T_{r\to r}|}}{|T_{r\to r}|}\sigma\sqrt{m}.
\end{aligned}$$

The first inequality above follows from the fact that

$$\begin{aligned}
\Big\| \sum_{A_i \in C_r^* \setminus T_{r\to r}} (A_i - \mu_r)\Big\| &= \|\mathbf{1}_S^\top (A - C)\| \ (\mathbf{1}_S \text{ is the indicator vector for points in } C_r^* \setminus T_{r\to r}) \\
&\leq \frac{\sigma\sqrt{m}}{\sqrt{|C_r^* \setminus T_{r\to r}|}}.
\end{aligned}$$

Next, by Assumptions II regarding large subsets of optimal clusters we have that for sets $T_{r\to s}$ either $|T_{r\to s}| \leq n \log n$ or

$$\begin{aligned}
\Big\|\textsc{Mean}(T_{r\to s}) - \mu_s\Big\| &= \frac{1}{|T_{r\to s}|}\Big\| \sum_{A_i \in T_{r\to s}} (A_i - \mu_s)\Big\| \\
&\leq \frac{1}{\sqrt{|T_{r\to s}|}}\sigma \text{poly}\log(m,n).
\end{aligned}$$

Furthermore, we also have the pointwise guarantee that for every $A_i \in T_{r\to s}$, $\|A_i - \mu_s\| \leq 2\sigma\sqrt{n} \cdot \text{poly}\log(m,n)$. Hence we get that

$$\frac{|T_{r\to s}|}{|S_r^*|}\Big\|\textsc{Mean}(T_{r\to s}) - \mu_r\Big\| \leq \max\Big(\frac{\sqrt{|T_{r\to s}|}}{|S_r^*|}\sigma\text{poly}\log(m,n), \frac{\sigma\text{poly}\log(m,n)}{n}\Big).$$

Substituting back into (85) we get that

$$\begin{aligned}
\|\textsc{Mean}(S_r^*) - \mu_r\| &\leq \frac{\sqrt{|C_r^*| - |T_{r,r}|}}{|S_r^*|}\sigma\sqrt{m} + \sum_{s\neq r}\frac{\sqrt{|T_{r\to s}|}}{|S_r^*|}\sigma\text{polylog}(m,n) + \sum_{s\neq r}\frac{|T_{r\to s}|}{|S_r^*|}\left\|\mu_r - \mu_s\right\| + \sigma \\
&\leq \frac{\sqrt{|S_r| - |T_{r,r}|}}{|S_r^*|}\sigma\sqrt{m} + \frac{\sqrt{|S_r \triangle C_r^*|}}{|S_r^*|}\sigma\sqrt{m} \\
&\quad + \sum_{s\neq r}\frac{\sqrt{|T_{r\to s}|}}{|S_r^*|}\sigma\text{polylog}(m,n) + \sum_{s\neq r}\frac{|T_{r\to s}|}{|S_r^*|}\left\|\mu_r - \mu_s\right\| + \sigma \\
&= \frac{\sigma\sqrt{m}\sqrt{\sum_{s\neq r}|T_{r\to s}|}}{|S_r^*|} + \frac{\sigma\sqrt{m}\sqrt{\sum_{s\neq r}|T_{r\to s}| + |T_{s\to r}|}}{|S_r^*|} \\
&\quad + \sum_{s\neq r}\frac{\sigma\text{polylog}(m,n)\sqrt{|T_{r\to s}|}}{|S_r^*|} + \sum_{s\neq r}\frac{|T_{r\to s}|}{|S_r^*|}\left\|\mu_r - \mu_s\right\| + \sigma \\
&\leq 4\sigma\sqrt{m}\sum_{s\neq r}\frac{\sqrt{|T_{r\to s}|}}{|S_r^*|} + 4\sigma\sqrt{m}\sum_{s\neq r}\frac{\sqrt{|T_{s\to r}|}}{|S_r^*|} \\
&\quad + \sum_{s\neq r}\frac{|T_{r\to s}|}{|S_r^*|}\left\|\mu_r - \mu_s\right\| + \sigma
\end{aligned}$$

Noticing that $|S_r^*| > |C_r^*|/2$ we get that

$$\|\textsc{Mean}(S_r^*) - \mu_r\| \leq 4\sigma\sqrt{m}\sum_{s\neq r}\frac{\sqrt{|T_{r\to s}|}}{|C_r^*|} + 4\sigma\sqrt{m}\sum_{s\neq r}\frac{\sqrt{|T_{s\to r}|}}{|C_r^*|} + \sum_{s\neq r}\frac{2|T_{r\to s}|}{|C_r^*|}\left\|\mu_r - \mu_s\right\| + \sigma$$

Substituting the bound on $T_{r\to s}$ from Lemma 35 we get that

$$\|\textsc{Mean}(S_r^*) - \mu_r\| \leq \frac{8c_1\sigma\sqrt{m}}{|C_r^*|}\sum_{s\neq r}\frac{\beta\sigma\sqrt{m}}{c\sqrt{k}\|\mu_r - \mu_s\|} + \sum_{s\neq r}\frac{2|T_{r\to s}|}{|C_r^*|}\left\|\mu_r - \mu_s\right\| + \sigma$$

where c_1 is an absolute constant depending on q. Substituting the lower bound on $\|\mu_r - \mu_s\|$ and using the definition of Δ_r we get that

$$\begin{aligned}
\|\textsc{Mean}(S_r^*) - \mu_r\| &\leq \frac{\beta\Delta_r}{4}\sum_{s\neq r}\frac{1}{c^2 k} + \sum_{s\neq r}\frac{2|T_{r\to s}|}{|C_r^*|}\left\|\mu_r - \mu_s\right\| + \sigma \\
&\leq \frac{\beta\Delta_r}{4} + \sum_{s\neq r}\frac{2|T_{r\to s}|}{|C_r^*|}\left\|\mu_r - \mu_s\right\| + \sigma.
\end{aligned}$$

To bound the second term, we again substitute the guarantee on $|T_{r\to s}|$ from Lemma 35 and get that

$$\begin{aligned}\sum_{s\neq r}\frac{2|T_{r\to s}|}{|C_r^*|}\Big\|\mu_r-\mu_s\Big\| &\le \sum_{s\neq r}\frac{2c_1\beta^2\sigma^2 m}{ck|C_r^*|\|\mu_r-\mu_s\|}\\ &\le \sum_{s\neq r}\frac{2c_1\beta^2\sigma\sqrt{m}\min(\sqrt{|C_r^*|},\sqrt{|C_s^*|})}{\alpha c^2 k\sqrt{k}|C_r^*|}\\ &\le \frac{\beta^2\Delta_r}{\alpha}\sum_{s\neq r}\frac{1}{c^2k\sqrt{k}} \le \frac{\beta^2\Delta_r}{\alpha c^2\sqrt{k}}.\end{aligned}$$

Combining the above we get that

$$\|\textsc{Mean}(S_r^*)-\mu_r\| \le \frac{\beta\Delta_r}{3}. \tag{87}$$

Next we analyze the true updates that correspond to running the RobustMean procedure on the set S_r. Notice from the guarantee of Theorem 30, when run on S_r, that either the algorithm will certify that the dataset if poisoned or will output an approximation to $\textsc{Mean}(S_r^*)$ upto a factor of $O(\alpha\sigma_{S_r^*})$ where $\sigma_{S_r^*}$ is the variance of the set S_r^* around μ_r. We next bound this value.

$$\begin{aligned}\sigma_{S_r^*}^2 &= \max_{v:\|v\|=1}\frac{1}{|S_r^*|}\sum_{A_i\in S_r^*}\Big((A_i-\mu_r)\cdot v\Big)^2\\ &\le \max_{v:\|v\|=1}\frac{1}{|S_r^*|}\sum_{A_i\in T_{r\to r}}\Big((A_i-\mu_r)\cdot v\Big)^2+\sum_{s\neq r}\max_{v:\|v\|=1}\frac{1}{|S_r^*|}\sum_{A_i\in T_{r\to s}}\Big((A_i-\mu_r)\cdot v\Big)^2 \qquad (88)\end{aligned}$$

Since $|T_{r\to r}|\ge\frac{7}{8}|C_r^*|\ge n^2$, from Assumptions II regarding large subsets of clusters we can bound the first term by

$$\max_{v:\|v\|=1}\frac{1}{|S_r^*|}\sum_{A_i\in S_r^*}\Big((A_i-\mu_r)\cdot v\Big)^2 \le O(\sigma^2\text{poly}\log(m,n)). \tag{89}$$

To bound the second term we have by triangle inequality that

$$\max_{v:\|v\|=1}\frac{1}{|S_r^*|}\sum_{A_i\in T_{r\to s}}\Big((A_i-\mu_r)\cdot v\Big)^2 \le \max_{v:\|v\|=1}\frac{1}{|S_r^*|}\sum_{A_i\in T_{r\to s}}\Big((A_i-\mu_s)\cdot v\Big)^2+\frac{|T_{r\to s}|}{|S_r^*|}\|\mu_r-\mu_s\|^2.$$

Here again the first term is either small due to $|T_{r\to s}|$ being small or is bounded due to Assumptions II about variance of large subsets. In particular, we have that

$$\max_{v:\|v\|=1}\frac{1}{|S_r^*|}\sum_{A_i\in T_{r\to s}}\Big((A_i-\mu_s)\cdot v\Big)^2 \le \max\Big(\frac{2\sigma^2 n\log n}{|S_r^*|},2\sigma^2\text{poly}\log(m,n)\frac{|T_{r\to s}|}{|S_r^*|}\Big). \tag{90}$$

Finally, using the bound on $|T_{r\to s}|$ from Lemma 35 we have that

$$\frac{|T_{r\to s}|}{|S_r^*|}\|\mu_r-\mu_s\|^2 \le \frac{2\beta^2\Delta_r^2}{\alpha^2 ck}. \tag{91}$$

Combining (89), (90), and (91) we get that

$$\sigma^2_{S_r^*} = \max_{v:\|v\|=1} \frac{1}{|S_r^*|} \sum_{A_i \in S_r^*} \Big((A_i - \mu_r)\cdot v\Big)^2$$
$$\leq O(\sigma^2 \text{poly}\log(m,n)) + 2\frac{\beta^2\Delta_r^2}{\alpha^2 c}.$$

Hence, at each step the RobustMean procedure will either certify that the dataset is poisoned or will find estimates $\nu_1^{(t+1)}, \ldots, \nu_k^{(t+1)}$ such that

$$\|\nu_r^{(t+1)} - \text{Mean}(S_r^*)\| \leq \tilde{O}(\alpha\sigma) + \frac{\beta\Delta_r}{4}.$$

Combining with (87) we get that at iteration $t+1$

$$\|\nu_r^{(t+1)} - \mu_r\| \leq \tilde{O}(\alpha\sigma) + \frac{\beta\Delta_r}{2}.$$

Hence, the updates will keep improving until the unavoidable error of $\tilde{O}(\alpha\sigma)$. ■

Information Theoretic Upper Bounds (Computationally Inefficient Algorithms). Finally, we would like to mention that using Proposition 27, via an (inefficient) algorithm we can get the same guarantees as in this section on clustering without the need for certification. In other words, if exponential time is allowed, then there exist algorithms for robust mean estimation and robust clustering that, given any δ-corrupted instance of the problem, will *always* output solutions achieving the error guarantees in Theorem 30, Theorem 31, Theorem 32 and Theorem 33 from this section. In order to achieve this, we simply use the (inefficient) robust mean estimation procedure from the guarantee of Theorem 58 when performing the modified Lloyd's updates and we use the guarantee of Proposition 27 to always compute good initial centers without the need for certification.

Appendix E. Learning Intersection of Halfspaces

We next demonstrate the applicability of our primitives in supervised learning as well. We will consider the problem of learning an intersection of k halfspaces over the Gaussian distribution on $\mathbb{R}^n$ in the presence of adversarial perturbations to the samples, both at testing-time and training-time. We will represent an intersection of halfspaces by a Boolean function $h : \mathbb{R}^n \to \{0,1\}$ denoted by $h(x) = \prod_{i=1}^k \mathbf{1}(w_i^\top x \geq \theta_i)$, where $\forall i \in [k]$, $\|w_i\|_2 = 1$ and $\theta_i \in \mathbb{R}$ and where $\mathbf{1}(\cdot)$ denotes the indicator function. Let $\mathcal{H}_k$ represent the hypothesis class of all intersections of at most k halfspaces. We will also refer to '1' as the positive label, and '0' as the negative label.

In the uncorrupted setting, the training points $x_1, \ldots, x_m \in \mathbb{R}^n$ are drawn i.i.d. from a Gaussian distribution, and their corresponding labels $y_i = h^*(x_i)$ for some $h^* \in \mathcal{H}_k$ (this corresponds to the realizable setting). The special case of $k = 1$ corresponds to standard linear classification. A series of well-known results (Vempala, 2010b,a; Klivans et al., 2008) starting with Vempala (2010b) shows that when we are given access to uncorrupted training samples in $\mathbb{R}^n$ drawn from a Gaussian distribution , one can PAC-learn an intersection of

half-spaces in time $f(k)\cdot \mathrm{poly}(n)$, where $f(k)$ has a super-polynomial dependence on k. Our algorithmic techniques will be used to learn an intersection of $k = O(1)$ half-spaces even when there are adversarial perturbations *both* at *training-time* and *test-time*. For simplicity we will focus on the case when the uncorrupted points are drawn from a spherical Gaussian $N(0, \sigma^2 I)$. We believe that the same ideas should also extend to general Gaussians, and other convex geometrical concepts as in Vempala (2010a).

Consider a classifier $h \in \mathcal{H}_2$ that is adversarially robust i.e., suppose $h(x) = \mathbf{1}(w_1^\top x \geq 0) \cdot \mathbf{1}(w_2^\top x \geq 0)$ is robust to adversarial δ-perturbations at test-time measured in ℓ_q norm, and let $b := \|w_1 - w_2\|_2 \in (0, 2)$. It is easy to show that $\max\{\|w_1\|_{q^*}, \|w_2\|_{q^*}\} \leq O(\sigma)/\delta$, otherwise for most positive examples there exists δ-adversarial perturbation that h misclassifies w.h.p! Moreover for such an adversarially robust classifier, we can assume that the subspace Π^* spanned by w_1, w_2 satisfies $\kappa := \|\Pi^*\|_{q\to 2} \leq O(\sigma/(\delta b))$ (see Claim 45). For general k, if the labels are generated by an intersection of k-halfspaces represented by $h^*(x) := \prod_{i=1}^k \mathbf{1}(w_i^\top x \geq \theta_i)$ with $\|w_i\|_2 = 1\ \forall i \in [k]$, we assume that the projection matrix Π^* onto the span of the normals $w_1, \ldots, w_k$ satisfies $\|\Pi^*\|_{q\to 2} \leq \kappa$.

We consider the following natural model, where each of the samples can be corrupted adversarially up to δ measured in ℓ_q norm for $q \geq 2$:

- Samples $x_1, x_2, \ldots, x_m \in \mathbb{R}^n$ are drawn i.i.d from $N(0, \sigma^2 I)$. The labels $y_1 = h^*(x_1), \ldots, y_m = h^*(x_m)$.
- For each $j \in [m]$, an adversary corrupts (corruptions could be dependent) the points to produce $\tilde{x}_1, \ldots, \tilde{x}_m \in \mathbb{R}^n$ such that $\forall j \in [m],\ \|\tilde{x}_j - x_j\|_q \leq \delta$.
- The input consists of $\{(\tilde{x}_1, y_1), (\tilde{x}_2, y_2), \ldots, (\tilde{x}_m, y_m)\}$.

The goal is to find an intersection of k half-spaces that achieves low-error and is adversarially robust to δ-perturbations at test-time (this is sometimes referred to as robust accuracy). Now we state our main result in this section.

Theorem 38 *Suppose $\kappa > 0, q \geq 2, \delta > 0$, and $k \leq n^{1/2}$. Let $h^* = \prod_{i=1}^k \mathbf{1}(w_i^\top x \geq \theta_i)$ with the normal vectors $w_1, \ldots, w_k$ spanning a (κ, q)-robust subspace. For convenience, let $\varepsilon = O\big(k^{4/3} \cdot (\kappa\delta\sqrt{k}/\sigma + \kappa^2\delta^2/\sigma^2)^{1/3}\big)$ denote the desired learning error rate. Suppose we are given $m = \mathrm{poly}(n, 1/\varepsilon)$ samples $\{(\widetilde{x}_i, y_i) : i \in [m]\}$ where $\widetilde{x}_i$ is a δ-perturbation (under ℓ_q norm) of $x_i \sim N(0, \sigma^2 I)$ and $y_i = h^*(x_i)$. There exists an algorithm that runs in time $\mathrm{poly}(n) \cdot (\frac{k}{\varepsilon})^{O(k^2)}$ to output $\widetilde{h} = \prod_{i=1}^k \mathbf{1}\big((w_i')^\top x \geq \theta'\big)$ such that with probability 0.9,*

$$\mathbb{P}_{x\sim N(0,\sigma^2 I)}[\widetilde{h}(x) = h^*(x)] \geq 1-\varepsilon, \text{ and } \mathbb{P}_{x\sim N(0,\sigma^2 I)}[\forall z \text{ s.t. } \|z\|_q \leq \delta,\ \widetilde{h}(x+z) = h^*(x)] \geq 1-2\varepsilon.$$

The above algorithm runs in polynomial time and returns an intersection of k half-spaces that achieves error $\varepsilon = o(1)$ as long as $\kappa\delta = o(\sigma)$. For example, when $\kappa \approx n^{0.1}$ this allows us to tolerate $\delta = 1/\kappa = o(n^{-0.1})$ as opposed to a tolerance of $\delta = o(n^{-1/2})$ for the naive approach. Recall from the earlier discussion, that such a condition is necessary qualitatively: even a single half-space $\mathbf{1}(w_1^\top x \geq 0)$ is not robust when $\|w_1\|_{q^*} = \kappa$ and $\kappa\delta \gg \sigma$.

Notation. We will use the following notation specific to this problem. Let $X \in \mathbb{R}^{n\times m}$ be the uncorrupted points, and $\tilde{X} \in \mathbb{R}^{n\times m}$ be the points obtained after adversarial perturbations. In particular, let m_+ denote the number of positive labels and $X_+, \tilde{X}_+ \in \mathbb{R}^{n\times m_+}$ correspond to the positive examples. In what follows $\mathbb{1} = (1, 1, \dots)$ will represent the all-ones vector of appropriate dimension. Let $B = \tilde{X}_+ - \frac{1}{m_+}\tilde{X}_+\mathbb{1}\mathbb{1}^\top$ be the centered input matrix corresponding to the (corrupted) positive examples. Hence, we can construct the co-variance matrices, *uncorrupted* and *corrupted* by $M_+ = \mathbb{E}\left[(x-\mu_+)(x-\mu_+)^\top \mid h^*(x) = +1\right]$, and $\tilde{M}_+ = \frac{1}{m_+}BB^\top$.

We will assume without loss of generality that $m_+ \geq (\kappa\delta/\sigma)\cdot m$. Otherwise, we can output the trivial hypothesis $x_1 > 0 \wedge (-x_1 > 0)$ that achieves an accuracy of $1 - O(\kappa\delta/\sigma)$ with high probability.

Finally, we will say that an intersection of halfspaces h is in a subspace $S \subset \mathbb{R}^n$ iff h can be represented as $h(x) = \prod_{i=1}^k \mathbf{1}(w_i^\top x \geq \theta_i)$, where $w_1, \dots, w_k \in S$.

Algorithm description and overview. The algorithm (Algorithm 10) follows the same general approach as Vempala (2010a). The main idea in Vempala (2010a) is to consider the co-variance matrix of just the positive examples X_+. With infinite samples, the (population) variance of X_+ in all the directions orthogonal to the span of $w_1, \dots, w_k$ in h^* is σ^2. On the other hand, the variance along directions in $\text{span}\{w_1, \dots, w_k\}$ is less than σ^2 because any thresholding (or any convex restriction) can only make the variance smaller; quantitative bounds on the gap are given in Lemma 39! Suppose the data is uncorrupted i.e., we are given X, we can just find the eigenspace corresponding to the k smallest eigenvalues of the covariance matrix corresponding to X_+, and learn the hypothesis in the k-dimensional subspace.

Lemma 39 (Lemma 4.8 in Vempala (2010a)) *Let g be the standard Gaussian density function in $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}_+$ be any logconcave function. Define the function h to the density $h(x) = f(x)g(x)/\beta$ where $\beta = \int_{\mathbb{R}^n} f(x)g(x)dx$. Then for any unit vector $u \in \mathbb{R}^n$,*

$$\text{Var}_h(u^\top x) \leq 1 - \frac{e^{-b^2}}{2\pi},$$

where the support of f along u is $[a_0, a_1]$ and $b = \min\{|a_0|, |a_1|\}$. In the particular the above statement also holds when f corresponds to the indicator function over any convex set.

We are given a matrix $\tilde{X}$ that corresponds to a δ-perturbation of X (a training time perturbation). We will use a convex-programming approach as in Section B, but to find the robust analog of a *least* singular subspace for the covariance matrix corresponding to $\tilde{X}_+$ i.e., our goal is to find an (orthogonal) projection matrix Π of rank r that is (κ, q) robust i.e., $\|\Pi\|_{q\to 2} \leq \kappa$ and that minimizes $\|\Pi B\|_F^2$.

We consider the following mathematical programming relaxation for the problem.

$$\min_Y \langle BB^\top, Y\rangle \tag{92}$$

$$\text{s.t.} \quad \text{tr}(Y) = r \tag{93}$$

$$0 \preceq Y \preceq I \tag{94}$$

$$\|Y\|_{q\to q^*} \leq \kappa^2 \tag{95}$$

Input: Samples $\tilde{X} \in \mathbb{R}^{n\times m}$ with labels $y_1, \ldots, y_m \in \{0,1\}$, σ, robustness parameter $\kappa \geq 1$, and the perturbation parameters δ and q. Set $\tau = k/(k+1)$.

1. Split the samples into two parts of T_1, T_2 where $|T_2| = \mathrm{poly}(k, \sigma/\kappa\delta)$.
2. Let $\tilde{X}_+$ be the positive examples in T_1 and let m_+ be the number of positive examples. Set $B = \tilde{X}_+ - \frac{1}{m_+}\tilde{X}_+ \mathbb{1}\mathbb{1}^\top$.
3. If $m_+ < (\kappa\delta/\sigma)m$, output the trivial hypothesis $h(x) = (x_1 > 0) \wedge (-x_1 > 0)$.
4. Else solve the convex program (92) on input B, with parameters κ, q to get a PSD matrix Y.
5. Let t be the number of eigenvalues of Y that are at least τ. Let $\widehat{\Pi}$ be the orthogonal projection given by the top $\min\{t, k\}$ and S_1' be this subspace.
6. We run a net argument on S_1' to find a hypothesis h :
   (a) Project the samples in T_2 onto the subspace S_1' to get samples $\{(x_j', y_j) : j \in T_2\}$ where $x_j' = \widehat{\Pi} \cdot \tilde{x}_j$.
   (b) Set $\varepsilon = 0.01 \cdot \max\left\{(\kappa\delta/\sigma)^3, (k+1)\eta\right\}$ where η is the same parameter in Lemma 40.
   (c) Let $\mathcal{W}$ be an ε-net of unit vectors in $S_1' : \forall v \in S_1', \exists u \in W$ s.t. $\|v - u\|_2 \leq \varepsilon$. Let T be an $\varepsilon\sigma$-net of thresholds in $[-5\sigma \cdot \log 1/\varepsilon, 5\sigma \cdot \log 1/\varepsilon]$.
   (d) Output any $h = \prod_{i=1}^k \mathbf{1}(w_i^\top x \geq \theta_i)$ with each $w_i \in \mathcal{W}$ and $\theta_i \in T$ satisfying

$$\sum_{j\in T_2} \mathbf{1}\Big(h(x_j') = y_j\Big) \geq |T_2| \cdot (1 - C \cdot k^{4/3}\eta^{1/3}).$$

Figure 10: Learning a robust intersection of halfspaces with training corruptions.

Note that the above program is a relaxation where for any (κ, q)-robust (orthogonal) projection matrix Π of rank r, $Y = \Pi$ is a feasible solution. Moreover as in Theorem 11, we can use the Ellipsoid algorithm along with a $O(1)$ factor separation oracle for the constraint (95), to find in polynomial time $\widehat{Y}$ such that $\|\widehat{Y}\|_{q\to q^*} \leq c_q\kappa^2$ (for some absolute constant $c_q > 0$ and the objective value attained by $\widehat{Y}$ is at most the optimum solution value of the above program (up to arbitrarily small accuracy).

The following lemma shows that we will recover a $(O(\kappa), q)$-robust projection matrix that captures all the directions where M_+ takes value significantly smaller than σ^2.

Lemma 40 *There exists a constant $c_1 > 0$ such that the projection matrix $\widehat{\Pi}$ output after step 5 of the algorithm satisfies $\|\widehat{\Pi}\|_{q\to 2} \le c_1\kappa$ and for any $\lambda \in (0,1)$*

$$\forall v \in \mathbf{S}^{n-1} \text{ s.t. } \Pi^* v = v \text{ and } v^\top M_+ v < \sigma^2(1-\lambda), \quad \text{we have } \|\widehat{\Pi} v\|_2^2 \ge 1 - \frac{2(r+1)\eta}{\lambda}$$
$$\text{where } \eta = 2\sqrt{k}\cdot\frac{\kappa\delta}{\sigma} + \frac{\kappa^2\delta^2}{\sigma^2} + O\Big(\frac{n\log n}{\sqrt{m_+}} + \frac{\kappa\delta}{\sigma}\cdot\frac{\sqrt{n}\log n}{\sqrt{m_+}}\Big).$$

The error η in Theorem 38 inherits the same η from the above lemma where we simplify the last two terms to $\kappa^2\delta^2/\sigma^2$ by assuming $m_+ \ge (\kappa\delta/\sigma)m$ and $m = \mathrm{poly}(n, k, \sigma/\kappa\delta)$ is sufficiently large. We defer the proof of Lemma 40 to Section E.1 and finish the proof of Theorem 38 in the rest of this section.

Lemma 39 implies the following claim (see the proof of Theorem 1.3 in Vempala (2010a)).

Claim 41 *Let $\{\lambda_i : i \in [n]\}$ and $\{v_i : i \in [n]\}$ be the eigenvalues and eigenvectors of M_+. Given any $\gamma \in (0,1)$, we set a subspace $S_1 := \mathrm{span}\{v_i | \lambda_i < (1-\gamma)\sigma^2\}$. Then there exists $h_1 \in \mathcal{H}_k$ in the subspace S_1, that agrees with h^* with probability at least $1 - O(\gamma\cdot k)$ over the uncorrupted samples:*

$$\mathop{\mathbb{P}}_{x\sim N(0,\sigma^2 I)}[h^*(x) = h_1(x)] > 1 - O(\gamma\cdot k).$$

We use the following claim along with Lemma 40 to show that the algorithm (up to step 5) recovers a subspace S_1' very close to S_1.

Claim 42 *Given $\gamma < 1/2$ and a subspace S_1 of dimension k with projection Π^*, let S_1' be a subspace whose projection matrix $\widetilde{\Pi}$ satisfies the following property: for every unit vector v in S_1, $\|\widetilde{\Pi} v\|_2 \ge 1-\gamma$. Then for any $h \in \mathcal{H}_k$ in subspace S_1, there exists another $h' \in \mathcal{H}_k$ in subspace S_1' (given by a natural projection onto S_1') that agrees with h with probability at least $1 - O(k\cdot\sqrt{\gamma\log\frac{1}{\gamma}})$ on an uncorrupted sample.*

Proof For an intersection $h(x) = \prod_{i=1}^k \mathbf{1}(w_i^\top x \ge \theta_i)$ in S_1, we have $\Pi^* w_i = w_i$. Without loss of generality, we assume $\|w_i\|_2 = 1$. Let $w_i' = \widetilde{\Pi} w_i$ for each i and $h'(x) = \prod_{i=1}^k \mathbf{1}((w_i')^\top x \ge \theta_i)$ be the projection of $h(x)$ into S_1'. So

$$\mathbb{P}[h(x) \ne h'(x)] \le \sum_{i=1}^k \mathbb{P}\left[\mathbf{1}\big(w_i^\top x \ge \theta_i\big) \ne \mathbf{1}\big((w_i')^\top x \ge \theta_i\big)\right].$$

Next we bound each probability. Notice that $(w_i')^\top x$ is a random variable drawn from $\sigma\cdot N(0, \|w_i'\|_2^2)$ and $w_i^\top x$ is drawn from $(w_i')^\top x + \sigma\cdot N(0, \|w_i\|_2^2 - \|w_i'\|_2^2)$. Note that $\|w_i'\|_2^2 \ge (1-\gamma)^2$ and $\|w_i\|_2^2 = 1$. So for $c := 6\sigma\sqrt{\gamma\log 1/\gamma}$

$$\mathbb{P}\left[\mathbf{1}\big(w_i^\top x \ge \theta_i\big) \ne \mathbf{1}\big((w_i')^\top x \ge \theta_i\big)\right] \le \mathop{\mathbb{P}}_{g\sim\sigma\cdot N(0,\|w_i'\|_2^2)}\Big[|g-\theta_i| \le c\Big] + \mathop{\mathbb{P}}_{\substack{g'\sim\\ \sigma\cdot N(0,\|w_i\|_2^2-\|w_i'\|_2^2)}}\Big[|g'| \ge c\Big]$$
$$\le O(\sqrt{\gamma\log 1/\gamma}).$$

We get the last inequality by using Gaussian anticoncentration for the first term, and standard Gaussian tail bounds for the second term: since $\|w_i\|_2^2 - \|w_i'\|_2^2 \le 2\gamma$ from the guarantee of S_1', the second term is bounded by $O(\sqrt{\gamma \log(1/\gamma)})$. Finally note that one can always rescale $\{w_i'\}$ (along with the thresholds) to be unit vectors without changing h'. ■

Next we use the VC dimension to bound the empirical risk error.

Claim 43 *Let S_1' be a subspace of dimension ℓ with projection matrix $\widetilde{\Pi}$ and consider a fixed $h^* \in \mathcal{H}_k$ in subspace S_1'. Then $m = O(k \cdot l \log k/\varepsilon)^2$ random Gaussian points $x_1, \ldots, x_m$ satisfy that with probability 0.99, any h, an intersection of k halfspaces in S_1', will have*

$$\mathbb{P}_{x \sim N(0,\sigma^2 I)} \big[h(\widetilde{\Pi}x) = h^*(\widetilde{\Pi}x)\big] = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\big(h(\widetilde{\Pi}x) = h^*(\widetilde{\Pi}x)\big) \pm \varepsilon.$$

Proof We first bound the VC dimension of intersections of at most k halfspaces by $2k(\ell+1)\log 5k$. The VC dimension of halfspaces in S_1' is $\ell+1$. Then the intersections have VC dimension at most $2k(\ell+1)\cdot\log(5k)$. So by the learnability of VC dimension (Blumer et al., 1989), the empirical error is ε for any h with probability at least 0.99. ■

Finally we finish the proof of Theorem 38 assuming Lemma 40 and the above claims.

Proof of Theorem 38 Set $\eta := O\big(\sqrt{k}\cdot\frac{\kappa\delta}{\sigma} + \frac{\kappa^2\delta^2}{\sigma^2}\big)$ and $\gamma := ((k+1)\eta)^{1/3}$. Let S^* be the subspace spanned by $w_1, \ldots, w_k$; by assumption Π^* is its projector. Let S_1 be the subspace spanned by the eigenvectors of M_+ whose corresponding eigenvalues are at most $\sigma^2(1-\gamma)$; note that the dimension of S_1 is at most k and $S_1 \subset S^*$. Let $\bar{h}_1 \in \mathcal{H}_k$ be the classifier in the subspace S_1 given by Claim 41 that approximates h^* up to error $O(k\cdot\gamma)$.

Then we apply Lemma 40 to obtain S_1' and $\widetilde{\Pi}$ such that $\|\widetilde{\Pi}v\|_2^2 \ge 1 - 2(k+1)\eta/\gamma$ for any unit vector v in S_1, since for our choice of η, it holds that $\eta \ge 2\kappa\delta\sqrt{k}/\sigma + \kappa^2\delta^2/\sigma^2 + O(\frac{n\log n}{\sqrt{m_+}} + \kappa\delta/\sigma \cdot \frac{\sqrt{n}\log n}{\sqrt{m_+}})$ and $m_+ \ge (\kappa\delta/\sigma)m$. For convenience, let h_1^* be the projection of $\bar{h}_1$ in S_1' from Claim 42. From the guarantees in Claim 41 and 42, we have

$$\mathbb{P}_{x\sim N(0,\sigma^2 I)}[h_1^*(x) = h^*(x)] \ge 1 - O(\gamma k) - O\Big(k\cdot\sqrt{\tfrac{(k+1)\eta}{\gamma}\cdot\log\tfrac{\gamma}{(k+1)\eta}}\Big). \tag{96}$$

In the rest of this proof, we will focus on learning h_1^*, or a good proxy for it, using the second part of samples in T_2. For convenience, we use $m_2 = |T_2|$ and $\big(\widetilde{x}_i, h(x_i)\big)_{i\in[m_2]}$ to denote the input. Since h^* and h_1^* are fixed and $m_2 > (k^2\log k/\varepsilon)^2$, we have with probability at least 0.99,

$$\frac{1}{m_2}\sum_{i=1}^{m_2}\mathbf{1}\big(h_1^*(x_i) \ne h^*(x_i)\big) \le O\Big(\gamma k + k\cdot\sqrt{\tfrac{(k+1)\eta}{\gamma}\cdot\log(\tfrac{\gamma}{(k+1)\eta})}\Big). \tag{97}$$

Recall that $\mathcal{W}$ is an ε-net of unit vectors in $S_1' : \forall v \in S_1', \exists u \in W$ s.t. $\|v-u\|_2 \le \varepsilon$. Let T be an $(\varepsilon\sigma)$-net of thresholds in $[-5\sigma\log 1/\varepsilon, 5\sigma\log 1/\varepsilon]$. So $\mathcal{W}\times T$ is a net for halfspaces in S_1': for any halfspace $\mathbf{1}(w^\top x \ge \theta)$ with $w \in S_1'$, exist $w' \in \mathcal{W}$ and $\theta' \in T$ such that $\mathbb{P}[\mathbf{1}\big(w^\top x \ge \theta\big) = \mathbf{1}\big((w')^\top x \ge \theta'\big)] \ge 1 - O(\sqrt{\varepsilon\log 1/\varepsilon})$. Similarly, $(\mathcal{W}\times T)^{\otimes k}$ gives a net

for intersections of k halfspaces in subspace S_1': for any intersection h of k halfspaces, exist $w_1, \ldots, w_k \in \mathcal{W}$ and $\theta_1, \ldots, \theta_k \in T$ such that

$$\mathop{\mathbb{P}}_{x \sim N(0,\sigma^2 I)}[h(x) = \prod_{j=1}^{k} \mathbf{1}(w_j^\top x \geq \theta_j)] \geq 1 - O(k\sqrt{\varepsilon \log 1/\varepsilon}).$$

Next we consider the empirical estimation over m_2 Gaussian points for $m_2 > (k^2 \log k/\gamma)^2$. By Claim 43, we have

$$\mathop{\mathbb{P}}_{x \sim N(0,\sigma^2 I)}\big[h_1^*(x) = h(x)\big] = \frac{1}{m_2}\sum_{i=1}^{m_2} \mathbf{1}\Big(h_1^*(x_i) = h(x_i)\Big) \pm \gamma. \tag{98}$$

Now we consider the empirical estimation after *perturbations*. For adversarial perturbations $\widetilde{x}_1, \ldots, \widetilde{x}_m$, we use $\|\widetilde{\Pi}\|_{q\to 2} = O(\kappa)$ to bound $\|\widetilde{\Pi}(x_i - \widetilde{x}_i)\|_2 \leq O(\kappa\delta)$. Let us fix $h \in \mathcal{H}_k$ in subspace S_1'; this could be h_1^* or any classifier in the net $(\mathcal{W}\times T)^{\otimes k}$. For a random $x_i \sim N(0, I)$, its adversarial perturbation $\widetilde{x}_i$ changes its label in h only if $w_i^\top \cdot (\widetilde{\Pi} x_i) \geq \theta_i$ and $w_i^\top \cdot (\widetilde{\Pi}\widetilde{x}_i) < \theta_i$ for some $i \in [k]$. Since their difference $w^\top \cdot (\widetilde{\Pi}X_i - \widetilde{\Pi}\widetilde{X}_i)$ is always bounded by $\|w\|_2 \cdot \|\widetilde{\Pi}X_i - \widetilde{\Pi}\widetilde{X}_i\|_2 \leq \kappa\delta$ in absolute value. So $\widetilde{x}_i$ changes its label in h with probability at most

$$k \cdot \mathop{\mathbb{P}}_{x_i}\Big[\big|w^\top(\widetilde{\Pi}x_i)\big| \leq \kappa\delta\Big] \leq k \cdot \frac{2\kappa\delta}{\sigma\cdot\sqrt{2\pi}}.$$

For m_2 random points $x_1, \ldots, x_{m_2}$, we have by standard concentration bounds,

$$\mathbb{P}\left[\sum_{i=1}^{m_2} \mathbf{1}\big(h(\widetilde{\Pi}\cdot x_i) \neq h(\widetilde{\Pi}\cdot\widetilde{x}_i)\big) \geq k \cdot \frac{2\kappa\delta}{\sigma\sqrt{2\pi}} \cdot m_2 + 5\sqrt{m_2 \cdot k \log|\mathcal{W}\times T|}\right] \leq \frac{1}{200\cdot|\mathcal{W}\times T|^k}.$$

For convenience, let err denote the normalized error $k \cdot \frac{2\kappa\delta}{\sigma\sqrt{2\pi}} + 5\sqrt{\frac{k\log|\mathcal{W}\times T|}{m_2}}$, which is $O(k\cdot\kappa\delta/\sigma + k^2 \log\frac{1}{\varepsilon\sigma}/\sqrt{m_2})$. We apply a union bound over the net of classifiers to claim that with probability at least 0.99,

$$\frac{1}{m_2}\sum_{i=1}^{m_2} \mathbf{1}\big(h(\widetilde{\Pi}\cdot x_i) \neq h(\widetilde{\Pi}\cdot\widetilde{x}_i)\big) \leq err \text{ for any } h \in \Big\{(\mathcal{W}\times T)^{\otimes k}, h_1^*\Big\}. \tag{99}$$

Plug this into Equation (98), we have

$$\mathop{\mathbb{P}}_{x}\big[h_1^*(x) = h(x)\big] = \frac{1}{m_2}\sum_{i=1}^{m_2} \mathbf{1}\big(h_1^*(\widetilde{\Pi}\cdot\widetilde{x}_i) = h(\widetilde{\Pi}\cdot\widetilde{x}_i)\big) \pm \gamma \pm 2err. \tag{100}$$

We are ready to show the correctness of Algorithm 10. It will output a solution because there exists $h \in (\mathcal{W}\times T)^{\otimes k}$ very close to h_1^*:

$$\begin{aligned}\frac{1}{m_2}\sum_{i=1}^{m_2} \mathbf{1}\big(h_1^*(\widetilde{\Pi}\widetilde{x}_i) = h(\widetilde{\Pi}\widetilde{x}_i)\big) &\geq \mathop{\mathbb{P}}_{x}[h_1^*(x) = h(x)] - \gamma - 2err && \text{(Equation (100))}\\ &\geq 1 - O(k\sqrt{\varepsilon\log 1/\varepsilon}) - \gamma - 2err. && \text{(from the property of the net)}\end{aligned}$$

At the same time, by equation (97), we also have

$$\frac{1}{m_2}\sum_{i=1}^{m_2}\mathbf{1}\big(h_1^*(\widetilde{\Pi}\widetilde{x}_i)=h^*(x_i)\big)\geq\frac{1}{m_2}\sum_{i=1}^{m_2}\mathbf{1}\big(h_1^*(\widetilde{\Pi}x_i)=h^*(x_i)\big)-err\geq 1-O(\gamma k+k\sqrt{(k+1)\eta/\gamma})-err.$$

Thus with probability 0.9, we have

$$\frac{1}{m_2}\sum_{i=1}^{m_2}\mathbf{1}\big(h(\widetilde{\Pi}\widetilde{x}_i)=h^*(x_i)\big)\geq 1-O(\gamma k+k\sqrt{(k+1)\eta/\gamma}+k\sqrt{\varepsilon\log 1/\varepsilon}+err).$$

We set the parameters and simplify the error: Let $\gamma=(k+1)^{1/3}\eta^{1/3}$ such that $\gamma=\sqrt{(k+1)\eta/\gamma}$ and $\varepsilon=0.01\cdot\max(\gamma^3,\kappa^3\delta^3/\sigma^3)$ s.t. $\sqrt{\varepsilon\log 1/\varepsilon}\leq\max\{\gamma,\kappa\delta\}$. Since $k<n^{1/2}$, the error becomes

$$1-O(k\gamma+err)=1-O\Big(k^{4/3}\cdot\eta^{1/3}+k\cdot\kappa\delta/\sigma+k^2\log(\tfrac{1}{\varepsilon})/\sqrt{m_2}\Big)=1-O(k^{4/3}\cdot\eta^{1/3}).$$

We finish the proof by showing any h in the net satisfying $\frac{1}{m_2}\sum_{i=1}^{m_2}\mathbf{1}\big(h(\widetilde{\Pi}\widetilde{x}_i)=h^*(x_i)\big)\geq 1-c\cdot k^{4/3}\cdot\eta^{1/3}$ is close to h^*: $\mathbb{P}_{X\sim N(0,I)}[h(X)=h^*(X)]\geq 1-O(k^{4/3}\cdot\eta^{1/3})$. By equation (99), we rewrite the guarantee of h into

$$\frac{1}{m_2}\sum_{i=1}^{m_2}\mathbf{1}\big(h(\widetilde{\Pi}x_i)=h^*(x_i)\big)\geq 1-ck^{4/3}\cdot\eta^{1/3}-err.$$

Furthermore, we use (97) to rewrite it as

$$\frac{1}{m_2}\sum_{i=1}^{m_2}\mathbf{1}\big(h(\widetilde{\Pi}x_i)=h_1^*(x_i)\big)\geq 1-ck^{4/3}\cdot\eta^{1/3}-err-O(\gamma k+k\cdot\sqrt{(k+1)\eta/\gamma\cdot\log\gamma/(k+1)\eta}).$$

By Equation (98), we have

$$\mathbb{P}[h(X)=h_1^*(X)]\geq 1-c\cdot k^{4/3}\cdot\eta^{1/3}-err-O(\gamma k+k\cdot\sqrt{(k+1)\eta/\gamma\cdot\log\gamma/(k+1)\eta})-\gamma.$$

We combine it with (96) using triangle inequality to obtain

$$\mathbb{P}[h(X)=h^*(X)]\geq 1-c\cdot k^{4/3}\cdot\eta^{1/3}-err-\gamma-O(\gamma k+k\cdot\sqrt{(k+1)\eta/\gamma\cdot\log\gamma/(k+1)\eta})=1-O(k^{4/3}\cdot\eta^{1/3}).$$

This establishes the required (natural) accuracy bound on the classifier $\widetilde{h}$. We now show the classifier $\widetilde{h}$ that we output is adversarially robust to test-time perturbations. Since $\|\widetilde{\Pi}\|_{\infty\to 2}=O(\kappa)$, we know the subspace spanned by the normals of halfspaces in $\widetilde{h}$ also has $\|\cdot\|_{\infty\to 2}\leq\|\widetilde{\Pi}\|_{\infty\to 2}=O(\kappa)$. Since each w_i is in the subspace with $\|w_i\|_{q^*}=O(\kappa)$, we know $w_i\cdot(x-\widetilde{x})=O(\kappa\delta)$ for any δ perturbation in ℓ_q norm. Using the Gaussian anti-concentration again, we bound the probability of a flip in the label from x to $\widetilde{x}$ for all valid δ-perturbations $\widetilde{x}$ by $\mathbb{P}[\widetilde{h}(x)\neq\widetilde{h}(\widetilde{x})]\leq O(k\kappa\delta/\sigma)$. By triangle inequality, this implies that the required robust error is at most $\varepsilon+O(k\kappa\delta/\sigma)\leq 2\varepsilon$. ■

E.1. Proof of Lemma 40

We first prove the following lemma which shows that that the robustness constraint ensures that the value on M_+ and $\tilde{M}_+$ are close.

Lemma 44 *For any projection matrix Y of rank r with $\|Y\|_{q\to q^*} \le \kappa^2$, we have with probability 0.99*

$$\left|\langle Y, M_+ - \tilde{M}_+\rangle\right| \le 2\kappa\delta\sqrt{r}\sigma + \kappa^2\delta^2 + \gamma(m_+), \text{ where } \gamma(m_+) = O\Big(\frac{\sigma^2 n\log n}{\sqrt{m_+}} + \frac{\sigma\kappa\delta\sqrt{n}\log n}{\sqrt{m_+}}\Big) \tag{101}$$

goes to 0 as the number of positive samples $m_+ \to \infty$.

Proof For convenience, let $\mu_+ = \mathbb{E}[x \mid h^*(x) = +]$, $\tilde{C}_+ = \tilde{X}_+ \mathbb{1}\mathbb{1}^\top$, $C_+ = X\mathbb{1}\mathbb{1}^\top$. Let $Z_+^\top = \tilde{X}_+ - X_+ - (\tilde{C}_+ - C_+)$ (and similarly for Z). Consider the following shifted covariance matrices for the positive examples for the corrupted

$$\tilde{M}_+ = \frac{1}{m_+}BB^\top = \frac{1}{m_+}\Big((\tilde{X}_+ - \tilde{C}_+)(\tilde{X}_+ - \tilde{C}_+)^\top\Big) \tag{102}$$

$$= \frac{1}{m_+}\Big((X_+ - C_+)(X_+ - C_+)^\top + Z_+Z_+^\top + (X_+ - C_+)Z_+^\top + Z_+(X_+ - C_+)^\top\Big),$$

$$= M_+ + E + \frac{1}{m_+}\Big(Z_+Z_+^\top + (X_+ - C_+)Z_+^\top + Z_+(X_+ - C_+)^\top\Big), \tag{103}$$

$$\text{where } E = \frac{1}{m_+}(X_+ - C_+)(X_+ - C_+)^\top - \mathbb{E}\left[(x-\mu_+)(x-\mu_+)^\top \mid h^*(x) = +\right] \tag{104}$$

represents the sampling error in the covariance matrix among positive examples, even in the absence of any corruptions. Moreover the samples (columns) of $X_+ - C_+$ are distributed according to a restriction of a spherical Gaussian onto a convex set. From Lemma 39, it follows that the (population) variance of X_+ in every direction is at most σ^2. Hence, the operator norm

$$\|\tfrac{1}{m_+}(X_+ - C_+)(X_+ - C_+)^\top\| \le \|M_+\| + \left\|\tfrac{1}{m_+}(X_+ - C_+)(X_+ - C_+)^\top - M_+\right\|$$
$$\le \sigma^2 + O\Big(\frac{\sigma^2\sqrt{n}\log n}{\sqrt{m_+}}\Big).$$

Recall that $\|Y^{1/2}\|_{q\to 2} = \sqrt{\|Y\|_{q\to q^*}} \le \kappa$. Hence, for any PSD matrix Y satisfying $\mathrm{tr}(Y) = r, 0 \preceq Y \preceq I$ and $\|Y\|_{q\to q^*} \le \kappa^2$,

$$\left|\langle Y, \tilde{M}_+ - M_+\rangle\right| \le \left\langle Y, \tfrac{1}{m_+}Z_+Z_+^\top\right\rangle + |\langle Y, E\rangle| + 2\left\langle Y, \tfrac{1}{m_+}(X_+ - C_+)Z_+^\top\right\rangle\Big|$$
$$\le \mathrm{tr}(Y)\|E\| + \frac{1}{m_+}\|Y^{1/2}Z_+\|_F^2 + \frac{2}{m_+}\langle Y^{1/2}Z_+, Y^{1/2}(X_+ - C_+)\rangle$$
$$\le r\|E\| + \frac{1}{m_+}\underbrace{\|Y^{1/2}Z_+\|_F^2}_{\text{using } \|Y^{1/2}\|_{q\to 2}\le\kappa} + 2\frac{1}{\sqrt{m_+}}\|Y^{1/2}Z_+\|_F \cdot \frac{1}{\sqrt{m_+}}\underbrace{\|Y^{1/2}(X_+ - C_+)\|_F}_{\le\|Y^{1/2}\|_F\cdot\|(X_+ - C_+)\|}$$
$$\le O\Big(\frac{r\sigma^2 n\log n}{\sqrt{m_+}} + \frac{\sigma\kappa\delta\sqrt{n}\log n}{\sqrt{m_+}}\Big) + 2\kappa\delta\cdot\sqrt{r}\sigma + \kappa^2\delta^2,$$

with probability 0.99. ■

We now prove Lemma 40.

Proof [Proof of Lemma 40] Let $\eta := 2\kappa\delta\sqrt{r}\sigma + \kappa^2\delta^2 + \gamma(m_+)$ be the error in Lemma 44. Also let $\Pi^* M_+ \Pi^* = \Sigma'$; by Lemma 39 we have $\Sigma' \preceq \sigma^2 I$. The optimal objective value of the relaxation (92)

$$\begin{aligned}
\langle \Pi^*, \tilde{M}_+ \rangle &\geq \langle Y, \tilde{M}_+ \rangle \\
\langle \Pi^*, \Sigma' \rangle = \langle \Pi^*, M_+ \rangle &\geq \langle Y, M_+ \rangle - 2\eta = \langle Y, \sigma^2 I - \sigma^2 \Pi^* + \Sigma' \rangle - 2\eta \\
&= r - \langle Y, \Pi^*(\sigma^2 I - \Sigma')\Pi^* \rangle - 2\eta, \quad \text{using Lemma 44.} \\
\text{Hence } \langle Y, \Pi^*(\sigma^2 I - \Sigma')\Pi^* \rangle &\geq \langle \Pi^*, \Pi^*(\sigma^2 I - \Sigma')\Pi^* \rangle - 2\eta \\
\text{i.e., } \langle \Pi^*(I - Y)\Pi^*, \Pi^*(\sigma^2 I - \Sigma')\Pi^* \rangle &\leq 2\eta.
\end{aligned}$$

We have a unit vector v in the subspace given by Π^* with $v^\top(\sigma^2 I - \Sigma')v \geq \sigma^2\lambda$. Moreover $0 \preceq Y \preceq I, \Sigma' \preceq \sigma^2 I$.

$$\begin{aligned}
(v^\top(I - Y)v)(v^\top(\sigma^2 I - \Sigma')v) &\leq \langle \Pi^*(I - Y)\Pi^*, \Pi^*(\sigma^2 I - \Sigma')\Pi^* \rangle \leq 2\eta \\
(1 - v^\top Y v) \cdot \sigma^2\lambda \leq 2\eta. \text{ Hence } \quad v^\top Y v &\geq 1 - \frac{2\eta}{\lambda\sigma^2}. \qquad (105)
\end{aligned}$$

Let $Y = \sum_{i=1}^n \lambda_i u_i u_i^\top$ be the eigendecomposition of Y with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, and $\lambda_1 \leq 1$. As in the algorithm let t be the number of eigenvalues that are larger than $r/(r+1)$. Since $\mathrm{tr}(Y) = r$, we have $t \leq r$. Note that $\widehat{\Pi} = \sum_{i=1}^t u_i u_i^\top$. By monotonicity of matrix norm $\|\widehat{\Pi^*}\|_{q\to q^*} \leq c_q(1 + 1/r)\kappa^2$. Finally,

$$\begin{aligned}
1 - \frac{2\eta}{\sigma^2\lambda} \leq v^\top Y v = \sum_{i=1}^n \lambda_i \langle u_i, v\rangle^2 &\leq \sum_{i=1}^t \langle u_i, v\rangle^2 + \lambda_{t+1} \sum_{i=t+1}^n \langle u_i, v\rangle^2 \leq 1 - \|\Pi^\perp v\|_2^2 + \frac{r}{r+1}\|\Pi^\perp v\|_2^2 \\
\|\Pi^\perp v\|_2^2 \leq \frac{2(r+1)\eta}{\sigma^2\lambda}, &\quad \text{hence proving the lemma.}
\end{aligned}$$

■

E.2. Properties of test-time robust classifiers

The following claim shows that for any test-time robust classifier given by the intersection of $k = 2$ halfspaces, the normals of the halfspaces are sparse. Moreover the subspace spanned by the normals is robust. We remark that a simple statement holds for general k with a dependence on the least (non-trivial) singular value of the matrix given by normals.

Claim 45 *Let $h : \mathbb{R}^n \to \{0, 1\}$ represent a classifier given by $h(x) = \mathbf{1}(w_1^\top x \geq 0) \cdot 1(w_2^\top x \geq 0)$, where $\|w_1\|_2 = \|w_2\|_2 = 1$ and $\|w_1 - w_2\|_2 = \gamma \in (0, 2)$. Suppose for $x \sim N(0, \sigma^2 I)$ with $h(x) = 1$ $\big($note that this happens with probability at least $\Omega(1 - \gamma/2)\big)$, we have with probability at least $2/3$ that*

$$\forall x' \in \mathbb{R}^n \ \text{s.t.} \ \|x - x'\|_q \leq \delta, \quad h(\tilde{x}) = h(x).$$

Then there exists some universal constant $c > 0$ such that we have $\max\{\|w_1\|_{q^*}, \|w_2\|_{q^*}\} \leq c\sigma/\delta$. *Moreover if* Π^* *is the projection matrix onto the span of* w_1, w_2, *we have that* $\|\Pi^*\|_{q\to 2} \leq (c/\gamma) \cdot \sigma/\delta$.

Proof For $x \sim N(0, \sigma^2 I)$ (even conditioned on $h(x) = 1$), we have that with probability at least 0.9, $|w_1^\top x|, |w_2^\top x| \leq O(\sigma)$. Let $\kappa' = \|w_1\|_{q^*} \geq \|w_2\|_{q^*}$ without loss of generality. Hence by norm duality, there exists a z with $\|z\|_q = \delta$ such that $w_1^\top z = \kappa\delta$. Thus if $\kappa\delta > c\sigma$ (for a large enough $c > 0$), we have $w_1^\top(x - z) < 0$. Hence the adversarial perturbation $\tilde{x} = x - z$ misclassifies the point.

Let Π^* be the projection matrix onto the span of w_1, w_2. Let u be any vector in the subspace given by Π^*. It is easy to see that since $\|w_1 - w_2\|_2 \geq \gamma$, u can be expressed as a linear combination $u = \alpha_1 w_1 + \alpha_2 w_2$ where $\alpha_1^2 + \alpha_2^2 = O(1/\gamma^2)$. This is because the minimum singular value of the matrix with columns w_1, w_2 is at least $\Omega(\gamma)$. Hence $\|u\|_{q^*} \leq |\alpha_1|\|w_1\|_{q^*} + |\alpha_2|\|w_2\|_{q^*} \leq O(\kappa/\gamma)$. This proves the lemma. ■

Appendix F. Trading off Natural Accuracy for Adversarial Robustness in Classification via Robust Projections

In many other natural scenarios it might be desirable to trade off natural accuracy for significant robustness to test-time perturbations. In this section we demonstrate how our techniques can be used for this purpose.

We study the simple binary classification setting under a natural Gaussian model (Anderson, 2003) for data generation. that was studied in recent works (Tsipras et al., 2018; Schmidt et al., 2018). In this model positive examples are drawn from a Gaussian distribution with mean μ_1 and covariance matrix Σ, whereas the negative examples are drawn from a Gaussian with mean μ_2 and the same covariance Σ. It is easy to see that if the means are well separated, e.g., if $\|\mu_1 - \mu_2\|_2 \geq \Omega(\sqrt{\log 1/\varepsilon})\sqrt{\|\Sigma\|}$, then the Bayes optimal classifier will have error at most ε. The works of Tsipras et al. (2018); Schmidt et al. (2018) used this simple model to study adversarial robustness and demonstrate that in many settings there is a natural tradeoff between the error and the robust error of any classifier (at test time); in particular there are settings where no robust classifier can achieve high natural accuracy. We continue this line of investigation and demonstrate that in many natural settings there do exists adversarially robust classifiers that also have small (natural) error. Furthermore, our algorithmic techniques can be used to learn such classifiers whereas standard approaches will fail.

To demonstrate this, we will consider settings where the means μ_1, μ_2 remain reasonable well separated when projected onto a robust subspace Π^*, albeit by a smaller amount. Furthermore, the component of vector $\mu_1 - \mu_2$ orthogonal to Π^* will be spread out, as measured by the analytic sparsity ℓ_{q^*}/ℓ_2 (where $q = \infty$, this is ℓ_1/ℓ_2 sparsity). In such a setting, an adversary can make a small ℓ_q perturbation to a fresh test example and make the error of the Bayes optimal classifier close to half. On the other hand, projecting the data onto the robust subspace Π^* first and performing classification in the subspace will let us tradeoff a small amount of natural accuracy for significant robustness gains. We now formally define the Gaussian model.

Gaussian Data Model $\mathcal{M}(\mu_1, \mu_2, \Sigma)$. In the Gaussian model an example label pair $(x, y) \in \mathbb{R}^n \times \{-1, +1\}$ is generated as follows. Pick $y = 1$ with probability $1/2$ and $y = -1$ with probability $1/2$. Conditioned on y generate x as

$$x|y \sim \left\{ \begin{array}{l} N(\mu_1, \Sigma), \text{ if } y = 1 \\ N(\mu_2, \Sigma), \text{ if } y = -1 \end{array} \right\}$$

We will use $(x, y) \sim \mathcal{M}(\mu_1, \mu_2, \Sigma)$ to denote a labeled example generated via the above process. Given a positive-definite matrix A, we will use A^+ to denote the Penrose-Moore pseudoinverse of A. Given a classifier $f : \mathbb{R}^n \to \{-1, +1\}$ we define the error of f to be the standard classification error given by

$$err(f) = \mathbb{P}_{(x,y)\sim\mathcal{M}(\mu_1,\mu_2,\Sigma)}[f(x) \neq y]. \tag{106}$$

We will measure perturbations in ℓ_q norm for $q > 2$ and for a given perturbation radius $\delta > 0$, we define the robust error of f to be

$$err_{rob}(f) = \mathbb{P}_{(x,y)\sim\mathcal{M}(\mu_1,\mu_2,\Sigma)}\Big[\exists z \in B_\delta(0) : f(x + z) \neq y\Big], \text{ where } B_\delta(0) = \{z : \|z\|_q \leq \delta\}. \tag{107}$$

To demonstrate this, we make the following natural assumptions on the structure of the above Gaussian model. We suspect that these results will hold in more general settings as well. For notational convenience, $\|\cdot\|$ will be used by default to denote the ℓ_2 norm for vectors, and the spectral norm for matrices, in the rest of the section.

Assumptions I.

1. *Mean Separation.* For a fixed $\varepsilon, \varepsilon_1 \in (0, 1)$ and a constant $c \geq 1$, there exists a rank-r projection Π^* with $\|\Pi^*\|_{\infty\to 2} \leq \kappa$, $\delta > 0$, and $\alpha, \beta \in (0, 1)$, such that

$$\|(\mu_1 - \mu_2)\| \geq c\sqrt{\log\big(\frac{1}{\varepsilon}\big)}\sqrt{\|\Sigma\|} \tag{108}$$

$$\|\Pi^*(\mu_1 - \mu_2)\| \geq c\sqrt{\log\big(\frac{1}{\varepsilon + \varepsilon_1}\big)}\sqrt{\|\Sigma\|} + \frac{\kappa\delta\beta}{\alpha} \tag{109}$$

$$\alpha^2\|\Sigma\| \leq \sigma_r(\Pi^*\Sigma\Pi^*) \leq \beta^2\|\Sigma\|. \tag{110}$$

2. *Spread Condition.* Let v be the vector defined as $v = \Sigma^+(\mu_1 - \mu_2)$. Then we have that

$$\|v\|_{q^*} \geq n^{0.1(\frac{1}{2}-\frac{1}{q})}\|v\|_2. \tag{111}$$

3. δ satisfies the following bound

$$\delta \geq \frac{\sqrt{\|\Sigma\|}\|(\Sigma^+)^{\frac{1}{2}}(\mu_1 - \mu_2)\|}{n^{0.1(\frac{1}{2}-\frac{1}{q})}}. \tag{112}$$

The constant 0.1 in (111) above is chosen for ease of exposition and in general one can define a similar condition in terms of n^c for a small constant $c > 0$.

Bayes Optimal Classifier: For the Gaussian model above, the Bayes optimal classification for a point x is obtained by comparing the density functions $p(x|y = 1)$ and $p(x|y = -1)$. In particular we have that

$$\frac{p(x|y=1)}{p(x|y=-1)} = \frac{e^{-(x-\mu_1)^\top \Sigma^+ (x-\mu_1)}}{e^{-(x-\mu_2)^\top \Sigma^+ (x-\mu_2)}}.$$

The above corresponds to the classifier

$$\begin{aligned} f^*(x) &= sgn\Big((x-\mu_2)^\top \Sigma^+ (x-\mu_2) - (x-\mu_1)^\top \Sigma^+ (x-\mu_1)\Big) \\ &= sgn\Big(\|(\Sigma^+)^{\frac{1}{2}}(x-\mu_2)\|^2 - \|(\Sigma^+)^{\frac{1}{2}}(x-\mu_1)\|^2\Big). \end{aligned} \tag{113}$$

Robust Projection-based Classifier: In a similar manner we define the robust classifier that performs classification after projection onto Π^*. When projected onto Π^*, the conditional distribution of x is again Gaussian with means either $\Pi^*\mu_1$ or $\Pi^*\mu_2$ depending on y, and the covariance matrix being $\Gamma = \Pi^*\Sigma\Pi^*$. Then we define the robust classifier as

$$f_{\Pi^*}(x) = sgn\Big(\|(\Gamma^+)^{\frac{1}{2}}\Pi^*(x-\mu_2)\|^2 - \|(\Gamma^+)^{\frac{1}{2}}\Pi^*(x-\mu_1)\|^2\Big). \tag{114}$$

The assumptions in (108) and (109) ensure that the means are separated in the ambient space as well as when projected onto the robust subspace Π^*. This will ensure that $f^*(x)$ has error at most ε and at the same time $f_{\Pi^*}(x)$ is not too much worse and has error at most $\varepsilon + \varepsilon_1$. The assumption in (110) will be used to argue that adversarial perturbations do not hurt the robust classifier $f_{\Pi^*}(x)$ and its robust error also remains at most $\varepsilon + \varepsilon_1$. Finally, the assumptions in (111) and (112) will be used to show that orthogonal to Π^* the vector $\mu_1 - \mu_2$ is spread out and an adversary can take advantage of this fact to design test-time perturbations that make the robust error of the Bayes optimal classifier close to half. Notice that the assumptions allow for a fairly large range of the perturbation radius δ as stated in (112). We next formalize these arguments.

F.1. Trading off Natural Accuracy for More Robustness Statistically

Proposition 46 (Non-robustness of the Bayes-optimal classifier) *If the Gaussian model $\mathcal{M}(\mu_1, \mu_2, \Sigma)$ satisfies Assumptions I then for the Bayes optimal classifier $f^*(x)$ defined in (113) it holds that*

$$\begin{aligned} err(f^*) &\leq \varepsilon \\ err_{rob}(f^*) &\geq \frac{1}{2}. \end{aligned}$$

Proof We first analyze the error of f^*. From (113) we have

$$\begin{aligned} err(f^*) = &\frac{1}{2}\mathbb{P}_{x\sim N(\mu_1,\Sigma)}\Big[\|(\Sigma^+)^{\frac{1}{2}}(x-\mu_2)\|^2 \leq \|(\Sigma^+)^{\frac{1}{2}}(x-\mu_1)\|^2\Big] \\ &+ \frac{1}{2}\mathbb{P}_{x\sim N(\mu_2,\Sigma)}\Big[\|(\Sigma^+)^{\frac{1}{2}}(x-\mu_1)\|^2 \leq \|(\Sigma^+)^{\frac{1}{2}}(x-\mu_2)\|^2\Big] \end{aligned}$$

Using the fact that when $x \sim N(\mu, \Sigma)$ then we have that $(\Sigma^+)^{\frac{1}{2}}(x-\mu)$ is distributed as Πg where $g \sim N(0, I)$ and Π is the orthogonal projection onto the subspace spanned by the singular vectors of Σ. Then we get have

$$\begin{aligned}
err(f^*) &= \frac{1}{2}\mathbb{P}_{g\sim N(0,I)}\Big[\|\Pi g + (\Sigma^+)^{\frac{1}{2}}(\mu_1-\mu_2)\|^2 \le \|\Pi g\|^2\Big] \\
&\quad + \frac{1}{2}\mathbb{P}_{g\sim N(0,I)}\Big[\|\Pi g + (\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\|^2 \le \|\Pi g\|^2\Big] \\
&= \mathbb{P}_{g\sim N(0,I)}\Big[\|\Pi g + (\Sigma^+)^{\frac{1}{2}}(\mu_1-\mu_2)\|^2 \le \|\Pi g\|^2\Big] \\
&= \mathbb{P}_{g\sim N(0,I)}\Big[\langle \Pi g, (\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\rangle \ge \frac{1}{2}\|(\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\|^2\Big].
\end{aligned}$$

Since $\langle \Pi g, (\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\rangle$ is a Gaussian with mean zero and variance $\|(\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\|^2$, we know from standard facts about Gaussian distribution that with probability at least $1-\varepsilon$, it holds that

$$|\langle \Pi g, (\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\rangle| \le c'\sqrt{\log(\frac{1}{\varepsilon})}\|(\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\|,$$

for a universal constant $c' > 0$. Furthermore, (108) implies that

$$\|(\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\| \ge \frac{1}{\sqrt{\|\Sigma\|}}\|\mu_2-\mu_1\| > c\sqrt{\log(\frac{1}{\varepsilon})}.$$

Setting $c > 2c'$ we get that

$$\begin{aligned}
err(f^*) &= \mathbb{P}_{g\sim N(0,I)}\Big[\langle \Pi g, (\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\rangle \ge \frac{1}{2}\|(\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\|^2\Big] \\
&\le \varepsilon.
\end{aligned}$$

Next we analyze the robust error of f^*. We have

$$\begin{aligned}
err_{rob}(f^*) &= \frac{1}{2}\mathbb{P}_{x\sim N(\mu_1,\Sigma)}\Big[\sup_{z\in B_\delta(0)}\|(\Sigma^+)^{\frac{1}{2}}(x+z-\mu_1)\|^2 - \|(\Sigma^+)^{\frac{1}{2}}(x+z-\mu_2)\|^2 \ge 0\Big] \\
&\quad + \frac{1}{2}\mathbb{P}_{x\sim N(\mu_2,\Sigma)}\Big[\sup_{z\in B_\delta(0)}\|(\Sigma^+)^{\frac{1}{2}}(x+z-\mu_2)\|^2 - \|(\Sigma^+)^{\frac{1}{2}}(x+z-\mu_1)\|^2 \ge 0\Big]
\end{aligned}$$

Again from symmetry we can rewrite this as

$$\begin{aligned}
err_{rob}(f^*) &= \mathbb{P}_{g\sim N(0,I)}\Big[\sup_{z\in B_\delta(0)}\|\Pi g + (\Sigma^+)^{\frac{1}{2}}z\|^2 - \|\Pi g + (\Sigma^+)^{\frac{1}{2}}(\mu_1-\mu_2) + \Sigma^{-\frac{1}{2}}z\|^2 \ge 0\Big] \\
&= \mathbb{P}_{g\sim N(0,I)}\Big[\langle \Pi g, (\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\rangle + \sup_{z\in B_\delta(0)}\langle (\Sigma^+)^{\frac{1}{2}}z, (\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\rangle \ge \frac{1}{2}\|(\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\|^2\Big] \\
&= \mathbb{P}_{g\sim N(0,I)}\Big[\langle \Pi g, (\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\rangle + \delta\|\Sigma^+(\mu_2-\mu_1)\|_{q^*} \ge \frac{1}{2}\|(\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\|^2\Big].
\end{aligned}$$

We will next show that $\delta\|\Sigma^+(\mu_2-\mu_1)\|_{q^*} \geq \frac{1}{2}\|(\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\|^2$, thereby establishing that $err_{rob}(f^*) \geq 1/2$. From (111) we have that

$$\begin{aligned}\delta\|\Sigma^+(\mu_2-\mu_1)\|_{q^*} &\geq \delta n^{0.1(\frac{1}{2}-\frac{1}{q})}\|\Sigma^+(\mu_2-\mu_1)\|_2 \\ &\geq \sqrt{\|\Sigma\|}\|(\Sigma^+)^{\frac{1}{2}}(\mu_1-\mu_2)\|\|\Sigma^+(\mu_2-\mu_1)\|_2, \text{ [using the lower bound on } \delta \text{ in (112)]} \\ &\geq \|(\Sigma^+)^{\frac{1}{2}}(\mu_2-\mu_1)\|^2.\end{aligned}$$

■

Proposition 47 (Guarantees for the Robust Projection-based classifier) *If the Gaussian model $\mathcal{M}(\mu_1,\mu_2,\Sigma)$ satisfies Assumptions I then for the robust classifier $f_{\Pi^*}(x)$ defined in (114) it holds that*

$$\begin{aligned}err(f_{\Pi^*}) &\leq \varepsilon+\varepsilon_1 \\ err_{rob}(f_{\Pi^*}) &\leq \varepsilon+\varepsilon_1.\end{aligned}$$

Proof Notice that Π^*x is distributed as $N(\Pi^*\mu,\Gamma)$ when $x\sim N(\mu,\Sigma)$, where $\Gamma=\Pi^*\Sigma\Pi^*$. Hence similar to the calculation in Proposition 46 we have that $f_{\Pi^*}(x)$ has error at most $\varepsilon+\varepsilon_1$ provided

$$\|\Pi^*(\mu_2-\mu_1)\| \geq c\sqrt{\log\big(\frac{1}{\varepsilon+\varepsilon_1}\big)}\sqrt{\|\Gamma\|}.$$

From (109) and noticing that $\|\Gamma\|\leq\|\Sigma\|$ we get that $err(f_{\Pi^*})\leq\varepsilon+\varepsilon_1$. Next we analyze the robust error of the classifier. Again from the calculations in the previous lemma we have $err_{rob}(f_{\Pi^*})$:

$$= \mathbb{P}_{g\sim N(0,I)}\Big[\langle\Pi^*g,(\Gamma^+)^{\frac{1}{2}}\Pi^*(\mu_2-\mu_1)\rangle + \sup_{z\in B_\delta(0)}\langle\Pi^*z,\Gamma^+\Pi^*(\mu_2-\mu_1)\rangle \geq \frac{1}{2}\|(\Gamma^+)^{\frac{1}{2}}\Pi^*(\mu_2-\mu_1)\|^2\Big].$$

Next notice that

$$\begin{aligned}\sup_{z\in B_\delta(0)}\langle\Pi^*z,\Gamma^+\Pi^*(\mu_2-\mu_1)\rangle &\leq \sup_{z\in B_\delta(0)}\|\Pi^*z\|\|\Gamma^+\Pi^*(\mu_2-\mu_1)\| \\ &\leq \kappa\delta\frac{\|(\Gamma^+)^{\frac{1}{2}}\Pi^*(\mu_2-\mu_1)\|}{\alpha\sqrt{\|\Sigma\|}}. \text{ [from (110)]}\end{aligned}$$

Notice that $\langle\Pi^*g,(\Gamma^+)^{\frac{1}{2}}\Pi^*(\mu_2-\mu_1)\rangle$ is a Gaussian that with probability at least $1-\varepsilon-\varepsilon_1$ takes a value at most $c'\|(\Gamma^+)^{\frac{1}{2}}\Pi^*(\mu_2-\mu_1)\|\sqrt{\log\big(\frac{1}{\varepsilon+\varepsilon_1}\big)}$. Hence to ensure that $err_{rob}(f_{\Pi^*})\leq\varepsilon+\varepsilon_1$ it is enough to have

$$\frac{1}{2}\|u\|^2 \geq c'\sqrt{\log\big(\frac{1}{\varepsilon+\varepsilon_1}\big)}\|u\| + \frac{\kappa\delta}{\alpha\sqrt{\|\Sigma\|}}\|u\|,$$

where $u=(\Gamma^+)^{\frac{1}{2}}\Pi^*(\mu_2-\mu_1)$. The bound follows from (109) and noticing that (110) implies that

$$\|u\| = \|(\Gamma^+)^{\frac{1}{2}}\Pi^*(\mu_2-\mu_1)\| \geq \frac{1}{\beta\sqrt{\|\Sigma\|}}\|\Pi^*(\mu_2-\mu_1)\|.$$

■

F.2. Efficient Algorithms for Finding a Robust Classifier

Next we discuss how our techniques from Section B.3 can be used to find a robust classifier in the Gaussian model discussed above. Given labeled examples, the first step of the learning algorithm is to use standard estimators for mean, covariance of single Gaussians on both the positive and negative examples separately, to find approximate parameters $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}$. Our recovery guarantee will depend on γ_1, γ_2 such that $\|(I-\Pi^*)(\Sigma^+)^{\frac{1}{2}}\|_2 \leq \gamma_1 \|(\Sigma^+)^{\frac{1}{2}}\|$ and $\sigma_r(\Pi^*(\Sigma^+)^{\frac{1}{2}}\Pi^*) \geq \gamma_2 \|(\Sigma^+)^{\frac{1}{2}}\|$. In this case the SDP from (45) when run on $\hat{\Sigma}^+$ implies that we will obtain a rank-r subspace $\hat{\Pi}$ such that $\|\hat{\Pi}^{\perp}\Pi^*\| \leq O(\frac{\gamma_1}{\gamma_2})$. Using the same analysis as above we can then show that the classifier

$$f_{\hat{\Pi}}(x) = sgn\Big(\|(\hat{\Gamma}^+)^{\frac{1}{2}}\hat{\Pi}(x-\hat{\mu}_2)\|^2 - \|(\hat{\Gamma}^+)^{\frac{1}{2}}\hat{\Pi}(x-\hat{\mu}_1)\|^2\Big),$$

where $\hat{\Gamma} = \hat{\Pi}\hat{\Sigma}\hat{\Pi}$. The learning algorithm is described below. To analyze the robust classifier we mildly strengthen Assumptions I below, to account for the error in the estimate of Π^*.

Assumptions II.

- For a fixed $\varepsilon \in (0,1)$ and a constant $c \geq 1$, there exists a rank-r projection Π^* with $\|\Pi^*\|_{\infty\to 2} \leq \kappa$, $\delta > 0$, and $\alpha, \beta, \gamma_1, \gamma_2 \in (0,1)$, such that

$$\|\Pi^*(\mu_1-\mu_2)\| - \frac{10\gamma_1}{\gamma_2}\|\mu_1-\mu_2\| \geq c\sqrt{\log\big(\frac{1}{\varepsilon}\big)}\sqrt{\|\Sigma\|} + \frac{\kappa\delta\beta}{\alpha'} \tag{115}$$

$$\alpha^2\|\Sigma\| \leq \sigma_r(\Pi^*\Sigma\Pi^*) \leq \beta^2\|\Sigma\| \tag{116}$$

$$\|(I-\Pi^*)(\Sigma^+)^{\frac{1}{2}}\| \leq \gamma_1\|(\Sigma^+)^{\frac{1}{2}}\| \tag{117}$$

$$\sigma_r(\Pi^*(\Sigma^+)^{\frac{1}{2}}\Pi^*) \geq \gamma_2\|(\Sigma^+)^{\frac{1}{2}}\|. \tag{118}$$

Here $\alpha' = \sqrt{\alpha^2 - \frac{8\gamma_1}{\gamma_2}}$.

RobustClassification$((x_1,y_1),\ldots,(x_m,y_m))$
Input: Labeled examples $(x_1,y_1),\ldots,(x_m,y_m)$.

1. Use the first $\frac{m}{2}$ examples to get estimates of $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}$ from standard estimators applied to positive and negative examples separately.
2. Use the algorithm from Corollary 21 with $A = (\hat{\Sigma}^+)^{\frac{1}{2}}$ to get $\hat{\Pi}$.
3. Define $\hat{\Gamma} = \hat{\Pi}\hat{\Sigma}\hat{\Pi}$.
4. Return the classifier

$$f_{\hat{\Pi}}(x) = sgn\Big(\|(\hat{\Gamma}^+)^{\frac{1}{2}}\hat{\Pi}(x-\hat{\mu}_2)\|^2 - \|(\hat{\Gamma}^+)^{\frac{1}{2}}\hat{\Pi}(x-\hat{\mu}_1)\|^2\Big).$$

Figure 11: Adversarially Robust Classification.

A natural illustrative example to keep in mind is the case when the two Gaussians are in the *parallel pancake* orientation (Brubaker and Vempala, 2008), where the variance is small along one direction u e.g., the x-axis (more generally, the variance is small along a robust subspace), and the variance in all the orthogonal directions are very large. If u is sparse (robust), and in addition the vector between the means has a reasonable projection on to u, then our algorithmic techniques can approximate u well with another sparse vector $\hat{u}$. Performing classification after projecting onto $\hat{u}$ will lead to a significantly more robust classifier without hurting the natural accuracy a lot.

Next we state and prove our main theorem regarding the Algorithm in Figure 11.

Theorem 48 *Given $\varepsilon \in (0,1)$, perturbation radius $\delta > 0$, and $m = poly(n, 1/\varepsilon, 1/\|\Sigma^+\|, 1/\|\mu_1 - \mu_2\|)$ labeled examples from the Gaussian data model $\mathcal{M}(\mu_1, \mu_2, \Sigma)$ satisfying Assumptions II, the algorithm in Figure 11 runs in polynomial time and with high probability outputs a classifier $f_{\hat{\Pi}}(x)$ such that $err_{rob}(f_{\hat{\Pi}}) \leq \varepsilon$.*

Proof By taking the empirical mean and covariance of the positive and negative examples with ε' set to be $\varepsilon' = \min\big(\varepsilon/poly(n, \|\Sigma^+\|), (\gamma_1/\gamma_2)\|\mu_1 - \mu_2\|\big)$ we get estimates $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}$ such that

$$\|\mu_1 - \hat{\mu}_1\| \leq \varepsilon', \quad \|\mu_2 - \hat{\mu}_2\| \leq \varepsilon', \quad \text{and } \|\Sigma - \hat{\Sigma}\|_2 \leq \varepsilon'.$$

The above combined with (116) and (117) implies that

$$\sigma_r(\Pi^*(\hat{\Sigma}^+)^{\frac{1}{2}}\Pi^*) \geq (1 - o(1))\gamma_2\|(\Sigma^+)^{\frac{1}{2}}\|$$
$$\|(I - \Pi^*)(\hat{\Sigma}^+)^{\frac{1}{2}}\| \leq (1 + o(1))\gamma_1\|(\Sigma^+)^{\frac{1}{2}}\|.$$

Hence, Corollary 21 implies that step 2 of Algorithm 11 will output $\hat{\Pi}$ such that $\|\hat{\Pi}^{\perp}\Pi^*\| \leq \frac{4\gamma_1}{\gamma_2}$. Next, to establish that $f_{\hat{\Pi}}$ has robust error at most ε, we need to verify that (109), (110) from Assumptions I hold with $\hat{mu}_1, \hat{\mu}_2, \hat{\Sigma}$ and $\hat{\Pi}$. We have

$$\begin{aligned}
\|\hat{\Pi}\hat{\Sigma}\hat{\Pi}\| &\geq \|\Pi^*\hat{\Sigma}\hat{\Pi}\| - \|(\hat{\Pi} - \Pi^*)\hat{\Sigma}\hat{\Pi}\| \\
&\geq \|\Pi^*\hat{\Sigma}\Pi^*\| - \|\Pi^*\hat{\Sigma}(\hat{\Pi} - \Pi^*)\| - 4\frac{\gamma_1}{\gamma_2}\|\Sigma\| \\
&\geq \|\Pi^*\Sigma\Pi^*\| - \|\Pi^*(\hat{\Sigma} - \Sigma)\Pi^*\| - 8\frac{\gamma^2}{\alpha^2}\|\Sigma\| \\
&\geq \alpha^2\|\Sigma\| - \varepsilon'\|\Sigma\| - 8\frac{\gamma_1}{\gamma_2} \\
&\geq (1 - o(1))\Big(\alpha^2 - 8\frac{\gamma_1}{\gamma_2}\Big)\|\Sigma\| \text{ [from (110)]}.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\|\hat{\Pi}(\hat{\mu}_1 - \hat{\mu_2})\| &\geq \|\hat{\Pi}(\mu_1 - \mu_2)\| - 2\varepsilon' \\
&\geq \|\Pi^*(\mu_1 - \mu_2)\| - 2\varepsilon' - \|(\hat{\Pi} - \Pi^*)(\mu_1 - \mu_2)\| \\
&\geq \|\Pi^*(\mu_1 - \mu_2)\| - 2\varepsilon' - 4\frac{\gamma_1}{\gamma_2}\|\mu_1 - \mu_2\| \\
&\geq c\sqrt{\log\big(\frac{1}{\varepsilon}\big)}\sqrt{\|\Sigma\|} + \frac{\kappa\delta}{\alpha'} \text{ [from (115)]}.
\end{aligned}$$

■

To demonstrate the applicability of Theorem 48 we instantiate it for a special case when $\Sigma = I - \theta\Pi^*$ for a constant $\theta < 1$ and Π^* being a rank-r κ-robust projection matrix. In this case the assumptions simplify to get the following corollary that we state for the case of $q = \infty$. We remark that this particular instantiation is only used to demonstrate the flavour of the condition; in this specialized setting one can of course use our knowledge of the form of covariance matrix to just find Π^* (by subtracting off the identity matrix), instead of our algorithm in step 2.

Corollary 49 *Let $s < 1$ be a fixed constant and $\mathcal{M}(\mu_1, \mu_2, \Sigma)$ be the Gaussian data model with $\Sigma = I - \theta\Pi^*$ with $\theta \in (0, s)$ such that*

$$\|\Pi^*(\mu_1 - \mu_2)\| - 10\sqrt{1-\theta}\|\mu_1 - \mu_2\| \geq c\sqrt{\log(\frac{1}{\varepsilon})} + c'\kappa\delta$$
$$\|(I - \Pi^*)(\mu_1 - \mu_2)\|_1 \geq (n^{0.1} + \frac{\kappa}{1-\theta})\|\mu_1 - \mu_2\|,$$

where c' is constant that depends on s. If δ satisfies

$$\frac{\|\mu_1 - \mu_2\|}{n^{0.1}} \leq \delta \leq \frac{\|\Pi^*(\mu_1 - \mu_2)\|}{\kappa},$$

then we have that the robust error of the Bayes optimal classifier satisfies $err_{rob}(f^*) \geq \frac{1}{2}$. On the other hand given $m = poly(n, 1/\varepsilon, 1/\theta)$ labeled examples from the Gaussian data model $\mathcal{M}(\mu_1, \mu_2, \Sigma)$ satisfying Assumptions II, the algorithm in Figure 11 runs in polynomial time and with high probability outputs a classifier $f_{\hat{\Pi}}(x)$ such that $err_{rob}(f_{\hat{\Pi}}) \leq \varepsilon$.

Appendix G. Related Work

Adversarial Robustness. Existing theoretical work on adversarial robustness has almost exclusively focused on supervised learning, and in particular on binary classification. These works include the study of adversarial counterparts of notions such as VC dimension and Rademacher complexity (Cullina et al., 2018; Khim and Loh, 2018; Yin et al., 2018), evidence of computational barriers (Bubeck et al., 2018b,a; Nakkiran, 2019; Degwekar et al., 2019) and statistical barriers towards ensuring both low test error and low adversarial test error (Tsipras et al., 2018), and computationally efficient algorithms for adversarially robust learning of restricted classes such as degree-1 and degree-2 polynomial threshold functions (Awasthi et al., 2019). Furthermore, recent works also provide evidence that adversarially robust supervised learning requires more training data than its non-robust counterpart (Schmidt et al., 2018; Montasser et al., 2019; Min et al., 2020).

The closest to our work is the result of Garg et al. (2018) that studies a particular formulation of adversarially robust features. The authors consider computing, given i.i.d. samples from a distribution, a map f such that, with high probability over a new example x drawn from the same distribution, points close to x get a nearby mapping (in ℓ_2 distance) under f. While similar in motivation to our work, the results in Garg et al. (2018) do not aim to minimize the projection error and simply require the projection f to be mean

zero and variance one to avoid trivial solutions. Furthermore, the authors look at a specific type of spectral embedding given by the top eigenvectors of the Laplacian of an appropriate graph constructed on the training data. The bounds presented for this embedding depend on the eigenvalue gap present in the Laplacian matrix. Finally, it is not clear how to efficiently use the embedding on new test points, as it involves recomputing the Laplacian by incorporating the new point into the training set.

Low Rank Approximations. There is a large body of work in randomized numerical linear algebra on methods such as column subset selection and CUR decompositions (Kannan and Vempala, 2017; Boutsidis et al., 2009; Deshpande and Rademacher, 2010; Boutsidis et al., 2014; Drineas et al., 2008; Boutsidis and Woodruff, 2017; Song et al., 2017) that aim to approximate a given matrix via a low dimensional subspace spanned by a small number of rows/columns of the matrix. However these algorithms do not necessarily yield robust representations; in particular the subspace that is spanned may not be robust in our sense ($q \to 2$ operator norm).

Sparse PCA. The problem of sparse PCA has been studied both in average-case and worst-case settings.

Average-case setting: In the high dimensional regime where the number of samples is much less than the dimensionality, several works have pointed out inconsistent behavior of PCA (Paul, 2007; Nadler et al., 2008; Johnstone and Lu, 2009). As a result this led to the study of the sparse PCA problem where it is assumed that the leading eigenvector is sparse. This problem is typically studied under an average case model known as the *spiked covariance model* (Johnstone et al., 2001)[10]. In this model the data is assumed to be generated from a Gaussian with covariance matrix $I + \theta vv^\top$, where the leading eigenvector v is assumed to be a sparse vector and θ is a parameter characterizing the signal strength. There have been several works that study minimax rates of estimating the leading eigenvector (and eigenspaces) under the spiked covariance model and for various notions of sparsity (Amini and Wainwright, 2009; Ma et al., 2013; Cai et al., 2013; Shen et al., 2013; Vu and Lei, 2012, 2013). More distantly related works include linearly transformed spiked models (Dobriban et al., 2020), where the focus is on deriving the Bayes optimal robust classifiers and recovering the unobserved signals of interest under noisy linear transforms.

Worst-case setting: There has also been work on the worst-case version of the problem in the special case when the rank $r = 1$, but for the maximization variant of the sparse PCA objective (Chan et al., 2016). For the maximization objective, the ℓ_0 and ℓ_1 versions (for capturing sparsity) are within a factor of 2 from each other (see Section 10.3.3 of Vershynin (2018)). Even when $r = 1$ the best known polynomial time algorithm gives a $O(n^{1/3})$ factor approximation in the worst-case (for both the ℓ_1 and ℓ_0 versions) (Chan et al., 2016). Moreover no constant factor approximation is possible assuming the SSE conjecture (Chan et al., 2016). Appendix H.3 shows how this also immediately implies computational hardness of our minimization version (3). The recent work of (Chowdhury et al., 2020) also studies the maximization version of the ℓ_0-sparse PCA problem and presents bicriteria algorithms for the sparsity and the quality of the returned solution. However, none of those results translate to multiplicative factor approximation algorithms for the minimization variant of

10. Such "spiked" models (signal plus noise) have a long history in statistics (Anderson, 2003) (first edition in 1962), viewed as certain types of *factor* models.

the problem that we study. More importantly, these works are restricted to the rank $r = 1$ setting, while we study the more general version of the problem.

Robustness to Corruptions in the Training Data. There is large body of work, spanning both the theoretical computer science and the statistical communities, that formulates and studies robustness to training data corruptions in the context of both supervised and unsupervised learning (Valiant, 1985; Kearns and Li, 1993; Huber, 2011; Diakonikolas and Kane, 2019). However they do not study the notion of adversarial perturbations to the data, to the best of our knowledge.

There is also a large body of work on Robust Optimization (Ben-Tal et al., 2009), where the input is uncertain and is assumed to belong to a structured *uncertainty* set. In robust optimization one looks for a single solution that is simultaneously good for all inputs in the uncertainty set, leading to a max-min formulation of the problem. In our model of corruption, we are interested in instance wise guarantees - for every input A and its corruption $\tilde{A}$, the algorithm is required to output a solution that is good for A (the solution is not required to be simultaneously good for all possible corruptions $\tilde{A}$). Moreover the resilience of a solution for A to corruptions implies structural properties that can be leveraged algorithmically. Moreover we are not aware of any results related to PCA in this context.

Robust variants of PCA. The problem of robust PCA has received significant attention in recent years (De La Torre and Black, 2003; Candès et al., 2011; Chandrasekaran et al., 2011). Here one assumes that a given corrupted matrix $\tilde{A}$ is a sum of two matrices, the true matrix A that is low-rank and a sparse corruption matrix S with sparsity pattern being essentially random. The corruptions, although sparse, can be unbounded in magnitude. This necessitates an incoherence type assumption that the "mass" or the principal components of A is spread out – recovery of A is impossible under unbounded sparse corruptions when the signal is localized or sparse. On the other hand, the corruptions may not be sparse in our case. In particular, *every* data point (in fact every entry of A) could be corrupted up to some specified magnitude δ. Here as our results show (particularly Theorem 2), localization (or sparsity) of the signal is crucial in tolerating adversarial perturbations in the training data (e.g., a spread out signal can be completely overwhelmed by the corruption in each entry of A). The very recent works of Awasthi et al. (2020a); d'Orsi et al. (2020) study a well-studied average-case model for sparse PCA under the same notion of training-time corruptions that we study in this work. The work of Awasthi et al. (2020a) builds on the current paper and characterizes the recovery error in terms of the $q \to 2$ operator norm, while the authors of d'Orsi et al. (2020) focus on $r = 1$ and characterize some computational vs statistical tradeoffs for this average-case model. On the other hand, our results apply to higher rank settings and for worst case data.

Huber's Contamination Model. In statistics, Huber's ε-contamination model (Huber, 2011) is the most widely studied. In this model the dataset is assumed to be generated i.i.d. from a mixture namely, $(1 - \varepsilon)P + \varepsilon Q$. Here P is the true distribution and is assumed to be well behaved, for example the Gaussian distribution, and Q is an arbitrary distribution modeling the noise. The study of this model has led to insightful results for a variety of problems. Recently, there have been many exciting developments in designing robust estimators of mean and covariance that are also computationally efficient (Diakonikolas

et al., 2019; Lai et al., 2016; Charikar et al., 2017; Diakonikolas and Kane, 2019). We would like to point out that in these works (and several other recent works), the model of corruption is different than ours. In particular, rather than assuming that the data contains a few outliers (Huber's model), in our model an adversary can potentially corrupt *every* data point up to magnitude δ (measured in ℓ_q norm for $q \geq 2$).

Clustering. From the computational point of view, the work of Dasgupta (1999) formulated the goal of clustering data generated from a mixture of well-separated Gaussians. There is a large body of work on designing efficient algorithms for clustering in this setting, both for Gaussians and more general distributions (Arora and Kannan, 2005; Vempala and Wang, 2004; Achlioptas and McSherry, 2005; Moitra and Valiant, 2010; Belkin and Sinha, 2010). See recent works (Regev and Vijayaraghavan, 2017; Hopkins and Li, 2018; Diakonikolas et al., 2018b; Kothari and Steinhardt, 2018) for a detailed discussion. The work of Kumar and Kannan (2010) abstracted out a common property of datasets (spectral stability as defined in (72)) that captures mixtures of well separated Gaussians, the planted partitioning model, and other well clustered instances. They showed that a single algorithm, namely the popular Lloyd's algorithm, with the right initialization, provably computes optimal solution for such stable instances. The separation factor needed for Lloyd's to work in Kumar and Kannan (2010) was later improved by Awasthi and Sheffet (2012). Vijayaraghavan et al. (2017) study Euclidean k-means clustering on instances that satisfy a notion of additive perturbation stability or resilience, where the optimal solution is stable even when each point is moved by a small amount. Analogously, in our problem the $\infty \to 2$ norm and sparsity captures the stability of the solution to small perturbations in ℓ_∞ norm. However the perturbations in Dutta et al. are measured in ℓ_2 norm, and the problem flavor and algorithms are very different.

Building on robust algorithms for mean estimation, there have also been works to perform robust clustering of well separated instances under Huber's contamination model and its variants (Brubaker, 2009; Diakonikolas et al., 2018b; Kothari and Steinhardt, 2018; Kothari and Steurer, 2018; Hopkins and Li, 2018). There have also been works in analyzing the EM algorithm for Gaussian mixtures (see e.g., (Balakrishnan et al., 2017)). While motivated by the study of the phenomenon of robustness, the above results do not provide guarantees in our model of corruption. As in the case of mean estimation, these results are designed to be robust to a small number of outliers (e.g., a small constant fraction) in the training set. In our corruption model on the other hand, every data point could be potentially corrupted up to magnitude δ (measured in ℓ_q norm for $q \geq 2$).

Appendix H. Auxillary and additional claims

H.1. Counterexamples

Claim 50 *The matrix norms $\|\cdot\|_q$ (entry-wise ℓ_q norm) and $\|\cdot\|_{\infty\to\infty}$, $\|\cdot\|_{1\to 1}$ are not monotone.*

Proof Let $v = \frac{1}{\sqrt{n}}\mathbf{1}$, where $\mathbf{1} = (1, 1, \ldots, 1)$. Consider the matrix $M = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Note that $\|v\|_2 = 1$ and $M \succeq 0$. Clearly $\|I\|_1 = \sum_{i,j} |I_{i,j}| = n$. On the other hand, when

$q \in [1,2)$,

$$\|M\|_q^q = \sum_{i,j} M_{i,j}^q = \sum_{i=1}^{n}(1-v(i)^2)^q + \sum_{i\neq j}|v(i)|^q|v(j)|^q = \|v\|_q^{2q} + n(1-\tfrac{1}{n})^q - \frac{n}{n^q} = (n-1)n^{1-q} + n(1-\tfrac{1}{n})^q > n.$$

Note that this particular instance is symmetric, and each column contributes equally; hence $\|M\|_{1\to q}^q = \frac{1}{n}\|M\|_q^q$, and similarly for I. Hence, the same counterexample also works for $1 \to q$ and $\infty \to q^*$ operator norms where $q \in [1,2)$. ■

Claim 51 *The matrix norms $\|\cdot\|_{\infty\to 2}$ and $\|\cdot\|_{2\to 1}$ are not monotone.*

Proof We consider a similar pair of matrices as the above claim:

$$A = \mathrm{diag}(u) \text{ where } u = (\underbrace{2,\ldots,2}_{n/3},\underbrace{1,\ldots,1}_{2n/3}) \text{ and } M = A - \frac{1}{n}\mathbf{1}\mathbf{1}^\top.$$

It will be easier to reason about $\|M\|_{\infty\to 2}$ and $\|A\|_{\infty\to 2}$. Recall $\|A\|_{\infty\to 2} = \max_{y:\|y\|_\infty\le 1}\|Ay\|_2$; since A is a diagonal matrix, the maximum value is $\tau := \sqrt{4\cdot(1/3)+2/3}\cdot\sqrt{n}$, and it is attained by every vector in $\{\pm 1\}^n$. To establish the claim, we now show that for a specific vector $y \in \{\pm 1\}^n$, $\|My\|_2 > \|Ay\|_2$.

$$\text{Consider } y = (\underbrace{1,\ldots,1}_{n/3},\underbrace{-1,\ldots,-1}_{2n/3})$$

$$\langle \mathbf{1}, y\rangle = -\tfrac{1}{3}n, \text{ and } My = (\underbrace{2+\tfrac{1}{3},\ldots,2+\tfrac{1}{3}}_{n/3},\underbrace{-1+\tfrac{1}{3},\ldots,-1+\tfrac{1}{3}}_{2n/3})$$

$$\text{Hence } \|My\|_2 = \sqrt{(\tfrac{1}{3})n\cdot(\tfrac{7}{3})^2 + (\tfrac{2}{3})n\cdot(\tfrac{2}{3})^2}$$

$$= \sqrt{\tau^2 + n(\frac{1}{9})} > \tau = \|Ay\|_2,$$

as required. Hence $\|A\|_{\infty\to 2} < \|M\|_{\infty\to 2}$, which violates the monotonicity property. ■

H.2. What do robust projection matrices look like?

Our robustness parameter $\|\Pi\|_{\infty\to 2}$ ($\|\Pi\|_{q\to 2}$ for general $q \ge 2$) generalizes analytic notions of sparsity for the subspace associated with the orthogonal projector Π (see Lemma 7). For the purposes of this discussion let us restrict our attention to $q = \infty$. As mentioned earlier, for a $r = 1$-dimensional subspace this exactly corresponds to the ℓ_1 sparsity of the unit vector v in that subspace. 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. We remark that for higher-dimensional subspaces, there are several other notions of sparsity that have been explored (Vu and Lei, 2013; Wang et al., 2014), typically measured 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, the quantity $\|\Pi\|_{q\to 2}$ is a basis-independent quantity that only depends on 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 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 $\|V\|_1$ take very different values for V_1 and V_3 that differ by a $\sqrt{r}$ factor. Our robustness parameter $\|\Pi\|_{\infty\to 2} \approx \sqrt{k}$ for all of them; 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 $\sqrt{k}$. Finally we remark that as mentioned in Figure 1 (and shown in Awasthi et al. (2020b)), many natural data sets have such robust low-dimensional projections with small error.

H.3. Computational Lower Bound

The main result in this section is to show that it is NP-hard to solve Program (13) in Section B exactly for $q = \infty$ under the small set expansion(SSE) hypothesis.

Conjecture 52 (SSE hypothesis (Raghavendra and Steurer, 2010)) *For any $\eta > 0$, there is $\delta > 0$ such that it is NP-hard to distinguish between the following two cases given a graph $G = (v, E)$:*

- *Yes: Some subset $S \subseteq V$ with $|S| = \delta n$ has expansion $\frac{|E(S,V\setminus S|}{|S|} \le \eta$*
- *No: Any set $S \subseteq V$ with $|S| \le 2\delta n$ has $\frac{|E(S,V\setminus S|}{|S|} \ge 1 - \eta$*

Based on SSE hypothesis, we state the hardness of solving Program (13) as follows.

Theorem 53 *It is SSE-hard to solve problem (13) given $q = \infty, r = 1$.*

We note the objective in problem (13) is the same as $\|A\|_F^2 - \max_\Pi \langle AA^\top, \Pi\rangle$ and use the hardness of $\max_\Pi \langle AA^\top, \Pi\rangle$ to finish the proof.

Theorem 54 (Theorem 4 in Chan et al. (2016)) *It is SSE-hard to solve the following program given k and a matrix $A \in \mathbb{R}^{n\times n}$ with any constant approximation ratio:* $\max_{\|x\|_2=1, \|x\|_0\le k} \|Ax\|_2^2$.

Next we state the relation between the ℓ_0-sparse and ℓ_1-sparse programs.

Theorem 55 ((Vershynin, 2018)) *Given any matrix $A \in \mathbb{R}^{n\times n}$ and any $k \ge 1$, let $OPT_{\ell_0} = \max_{\|x\|_2=1, \|x\|_0\le k} \|Ax\|_2^2$ and $OPT_{\ell_1} = \max_{\|x\|_2=1, \|x\|_1\le \sqrt{k}} \|Ax\|_2^2$. Then we have $OPT_{\ell_0} \le OPT_{\ell_1} \le 2 \cdot OPT_{\ell_0}$.*

Finally we finish the proof of Theorem 53.

Proof of Theorem 53 For contradiction, suppose there is an algorithm that solves problem (13) for $q = \infty$ and any κ. Given a matrix A and k, since

$$\max_{\Pi:\|\Pi\|_{\infty\to 2}\le\kappa} \langle AA^\top, \Pi\rangle = \max_{\|x\|_2=1, \|x\|_1\le\sqrt{k}} \langle AA^\top, xx^\top\rangle$$

for rank 1 projection matrices, the algorithm for problem (13) also solves $\max_{\|x\|_2=1, \|x\|_1\le\sqrt{k}} \|Ax\|_2^2$ by the reformulation (5). Because of Theorem 55, this gives a 0.5 approximation of $\max_{\|x\|_2=1, \|x\|_0\le k} \|Ax\|_2^2$, which refutes the SSE hypothesis based on Theorem 54. ■

Notice that the above proof only establishes computational hardness for exact minimization of (13) under the small set expansion conjecture. It would be interesting to establish hardness of approximation results for this problem.

H.4. Proof of Lemma 30

Proof By assumption we know that $\|A - C\| \le \sigma\sqrt{m}$. This implies that

$$\begin{aligned}\|A - \Pi^* A\| &\le \|A - C\| + \|C - \Pi^* A\| \\ &= \|A - C\| + \|\Pi^*(C - A)\| \\ &\le 2\|A - C\| \le 2\sigma\sqrt{m}.\end{aligned}$$

Since we set $\tau = 2\sigma\sqrt{m}$, from the guarantee of Theorem 25, we know that if the algorithm outputs Bad Input , the data must be poisoned, i.e., $\|A - \tilde{A}\| > 2\sigma\sqrt{m}$. Next suppose that the algorithm outputs a projection matrix Π. Setting $\hat{\mu} := \textsc{Mean}(\Pi\tilde{A})$, and $\mathbf{1}$ to be

the all-ones vector $(1, 1, \ldots, 1)^\top$ we have that

$$\begin{aligned}
\|\hat{\mu} - \mu\|_2 &= \frac{1}{m}\Big\|\sum_{j=1}^{m}(A_j - \Pi\tilde{A}_j)\Big\|_2 \\
&\leq \frac{1}{m}\Big\|\sum_{j=1}^{m}(A_j - \Pi A_j)\Big\|_2 + \frac{1}{m}\Big\|\sum_{j=1}^{m}(\Pi A_j - \Pi\tilde{A}_j)\Big\|_2 \\
&\leq \frac{1}{m}\|\mathbf{1}^\top(A - \Pi A)\|_2 + \frac{1}{m}\sum_{i=1}^{m}\|\Pi(A_j - \tilde{A}_j)\|_2 \\
&\leq \frac{1}{\sqrt{m}}\|A - \Pi A\| + c_q\kappa\delta. \qquad (119)
\end{aligned}$$

Next we make a crucial observation that if Π is good for $\tilde{A}$ then it is also good for A and hence $\|A - \Pi A\|$ is small. This is formally established in Lemma 56. Applying the lemma on A and $\tilde{A}$ with $\Pi_1 = \Pi^*$, $\Pi_2 = \Pi$, $\kappa_1 = \kappa$, $\kappa_2 = c_q\kappa$, and $\varepsilon = 4\sigma\sqrt{m}/\|A\|$ we get that

$$\|A - \Pi A\| \leq (\varepsilon + \sqrt{\varepsilon})\|A\| + 8\frac{\kappa\delta\sqrt{m}}{\sqrt{\varepsilon}}.$$

Substituting into (119) we get that

$$\begin{aligned}
\|\hat{\mu} - \mu\|_2 &\leq (\varepsilon + \sqrt{\varepsilon})\frac{\|A\|}{\sqrt{m}} + 8\frac{\kappa\delta}{\sqrt{\varepsilon}} + c_q\kappa\delta \\
&\leq (\varepsilon + \sqrt{\varepsilon})\frac{4\sigma}{\varepsilon} + 8\frac{\kappa\delta}{\sqrt{\varepsilon}} + c_q\kappa\delta \text{ (by writing } \|A\| \text{ in terms of } \varepsilon) \\
&\leq O(c_q)(\sigma + \kappa\delta)\Big(1 + \sqrt{\frac{1}{\varepsilon}}\Big). \qquad (120)
\end{aligned}$$

Next, notice that by triangle inequality,

$$\begin{aligned}
\|A\| &\leq \|A - C\| + \|C\| \\
&\leq (\sigma + \|\mu\|)\sqrt{m}.
\end{aligned}$$

Hence we get that

$$\sqrt{\frac{1}{\varepsilon}} = \sqrt{\frac{\|A\|}{4\sigma\sqrt{m}}} \leq \frac{1}{2}\Big(1 + \frac{\|\mu\|}{\sigma}\Big)^{\frac{1}{2}}.$$

Substituting into (120) we get that

$$\begin{aligned}
\|\hat{\mu} - \mu\|_2 &\leq O(c_q)(\sigma + \kappa\delta)\Big(1 + \Big(1 + \frac{\|\mu\|}{\sigma}\Big)^{\frac{1}{2}}\Big) \qquad (121) \\
&\leq O(c_q)(1 + \frac{\kappa\delta}{\sigma})\max\Big(\sigma, \sqrt{\sigma\|\mu\|}\Big). \qquad (122)
\end{aligned}$$

From the above we get the relative error guarantee of

$$\frac{\|\hat{\mu} - \mu\|_2}{\|\mu\|} \leq O(c_q)(1 + \frac{\kappa\delta}{\sigma})\max\Big(\frac{\sigma}{\|\mu\|}, \sqrt{\frac{\sigma}{\|\mu\|}}\Big). \qquad (123)$$

■

Note: We would like to point out that for robust mean estimation, our analysis also shows that in step 2 of the algorithm above, we can replace $\textsc{Mean}(\Pi\tilde{A})$ with $\textsc{Mean}(\tilde{A})$. This is because if the algorithm did not output Bad Input then $\|\tilde{A}-\Pi\tilde{A}\|_2/\sqrt{m} \le 2\sigma$ and hence mean of $\tilde{A}$ and that of $\Pi\tilde{A}$ will be close. However, in this case, the subspace spanned by the output vector, i.e., $\textsc{Mean}(A)$ might not be robust and hence susceptible to test-time perturbations.

Lemma 56 *Fix $q \ge 2, \delta > 0, \kappa \ge 1$. Let A and $\tilde{A}$ be two $n \times m$ matrices, each representing m data points in n dimensions such that for every $j \in [m]$, columns A_j and $\tilde{A}_j$ are close, i.e., $\|A_j - \tilde{A}_j\|_q \le \delta$. Furthermore, assume that there exist projection matrices, $\Pi_1 = vv^\top$ and $\Pi_2 = uu^\top$ such that $\|\Pi_1\|_{q\to 2} \le \kappa_1$ and $\|\Pi_2\|_{q\to 2} \le \kappa_2$ and that $\|A - \Pi_1 A\| \le \varepsilon_1\|A\|$ and $\|\tilde{A} - \Pi_2\tilde{A}\| \le \varepsilon_2\|A\|$. Then, letting $\varepsilon = \varepsilon_1 + \varepsilon_2$ and $\kappa = \kappa_1 + \kappa_2$, it also holds that*

$$\|A - \Pi_2 A\| \le O(\varepsilon + \sqrt{\varepsilon})\|A\| + \frac{8\kappa\delta\sqrt{m}}{\sqrt{\varepsilon}} \tag{124}$$

$$\|\tilde{A} - \Pi_1\tilde{A}\| \le O(\varepsilon + \sqrt{\varepsilon})\|A\| + \frac{8\kappa\delta\sqrt{m}}{\sqrt{\varepsilon}} \tag{125}$$

Proof We will show the desired bound on $\|A-\Pi_2 A\|$ and by symmetry the same bound will also apply to $\|\tilde{A}-\Pi_1\tilde{A}\|$. Notice that both Π_1 and Π_2 are projections onto one dimensional subspaces and a bound on $\|\|_{q\to 2}$ norm of the projection matrices implies that $\|v\|_{q^*} \le \kappa_1$ and $\|u\|_{q^*} \le \kappa_2$, where q^* is such that $1/q + 1/q^* = 1$. Next, let Π be the projection matrix onto the subspace spanned by v and u. By triangle inequality we have that

$$\begin{aligned}
\|A - \Pi_2 A\| &\le \|A - \Pi A\| + \|\Pi A - \Pi_2 A\| \\
&\le \|A - \Pi A\| + \|\Pi A - \Pi_2\tilde{A}\| + \|\Pi_2\tilde{A} - \Pi_2 A\| \\
&\le \|A - \Pi A\| + \|\Pi A - \Pi\tilde{A}\| + \|\Pi\tilde{A} - \Pi_2\tilde{A}\| + \|\Pi_2\tilde{A} - \Pi_2 A\|.
\end{aligned} \tag{126}$$

Recall the standard fact that if P_1 and P_2 are projection matrices on to subspaces S_1 and S_2 such that $S_1 \subseteq S_2$, then for any matrix B, $\|P_1 B\| \le \|P_2 B\|$. Using this we to get that

$$\|A - \Pi A\| \le \|A - \Pi_1 A\| \le \varepsilon_1\|A\| \tag{127}$$

and,

$$\begin{aligned}
\|\Pi_2\tilde{A} - \Pi\tilde{A}\| &\le \|\Pi_2\tilde{A} - \tilde{A}\| + \|\tilde{A} - \Pi\tilde{A}\| \\
&\le 2\|\Pi_2\tilde{A} - \tilde{A}\| \le 2\varepsilon_2\|A\|.
\end{aligned} \tag{128}$$

From the closeness of A and $\tilde{A}$ and the robustness of Π_2 we also know that

$$\|\Pi_2\tilde{A} - \Pi_2 A\| \le \|\Pi_2\|_{q\to 2}\delta\sqrt{m} \le \kappa_2\delta\sqrt{m}. \tag{129}$$

Substituting (127), (128), and (129) into (126) we get that

$$\|A - \Pi_2 A\| \le \varepsilon_1\|A\| + 2\varepsilon_2\|A\| + \kappa_2\delta\sqrt{m} + \|\Pi A - \Pi\tilde{A}\|. \tag{130}$$

Note that if $\|\Pi A - \Pi\tilde{A}\| \leq \kappa\delta\sqrt{m}/\sqrt{\varepsilon}$, then we have the desired bound on $\|A - \Pi_2 A\|$. We now look at the case when $\|\Pi A - \Pi\tilde{A}\| > \kappa\delta\sqrt{m}/\sqrt{\varepsilon}$. Notice that Π is the union of robust subspaces and $A - \tilde{A}$ has columns bounded in q norm. Hence, the only way $\|\Pi A - \Pi\tilde{A}\|$ can be very large is if the $\|\|_{q\to 2}$ norm of the projection matrix of a union of two subspaces (Π) is much larger than the $\|\|_{q\to 2}$ norm of the projection matrices of individual subspaces (Π_1 and Π_2). For this to happen the two subspaces must be very close to each other and then we can bound $\|A - \Pi_2 A\|$ in a different way. Formally, we have that

$$\begin{aligned}
\|\Pi A - \Pi\tilde{A}\| &= \max_{z:\|z\|=1} \|\Pi(A - \tilde{A}) \cdot z\| \\
&= \max_{z:\|z\|=1} \|\sum_{j=1}^{m} z_j \Pi(A_j - \tilde{A}_j)\| \\
&\leq \max_{z:\|z\|=1} \sum_{j=1}^{m} |z_j| \|\Pi(A_j - \tilde{A}_j)\| \\
&\leq \|\Pi\|_{q\to 2}\delta\sqrt{m}. \qquad (131)
\end{aligned}$$

Next we establish an upper bound on $\|\Pi\|_{q\to 2}$ in terms of the distance between subspaces $\Pi_1 = vv^\top$ and $\Pi_2 = uu^\top$. Suppose $\|u - v\| = \gamma$ and that $u \cdot v \geq 0$ (otherwise we work with $-u$). We also know that $\|v\|_{q^*} \leq \kappa_1$ and $\|u\|_{q^*} \leq \kappa_2$. Now, $\|\Pi\|_{q\to 2}$ is the maximum q^* norm of any unit vector in the span of v and u. We can write any such vector z as

$$z = \alpha_1 v + \alpha_2 v^\perp$$

where $\alpha_1^2 + \alpha_2^2 = 1$ and $v^\perp = \frac{u-(u\cdot v)v}{\|u-(u\cdot v)v\|}$. Next we have that

$$\begin{aligned}
\|u - (u \cdot v)v\|^2 &= 1 - (u \cdot v)^2 \\
&\geq 1 - u \cdot v = \frac{\gamma^2}{2}.
\end{aligned}$$

Hence we get that for any z in the span of v and u

$$\begin{aligned}
\|z\|_{q^*} &\leq \|v\|^{q^*} + \|v^\perp\|_{q^*} \\
&\leq \kappa_1 + \frac{\sqrt{2}}{\gamma}\Big(\|u\|_{q^*} + \|v\|_{q^*}\Big) \\
&\leq \frac{2\sqrt{2}}{\gamma}\kappa.
\end{aligned}$$

The above also establishes that

$$\|\Pi\|_{q\to 2} \leq 2\frac{\sqrt{2}}{\gamma}\kappa.$$

Substituting into (131) we get that

$$\|\Pi A - \Pi\tilde{A}\| \leq \frac{2\sqrt{2}}{\gamma}\kappa\delta\sqrt{m}. \qquad (132)$$

Hence, if $\|\Pi A - \Pi\tilde{A}\| > \kappa\delta\sqrt{m}/\sqrt{\varepsilon}$ we must have that

$$\|v - u\| = \gamma \leq 2\sqrt{2}\sqrt{\varepsilon}. \tag{133}$$

In this case we can bound $\|A - \Pi_2 A\|$ as

$$\begin{aligned}
\|A - \Pi_2 A\| &\leq \|A - \Pi_1 A\| + \|(\Pi_1 - \Pi_2)A\| \\
&\leq \varepsilon_1\|A\| + \|\Pi_1 - \Pi_2\|\|A\| \\
&\leq \varepsilon\|A\| + \|vv^\top - uu^\top\|\|A\| \\
&= \varepsilon\|A\| + \|\frac{1}{2}(v+u)(v-u)^\top + \frac{1}{2}(v-u)(v+u)^\top\|\|A\| \\
&\leq \varepsilon\|A\| + 2\|v-u\|\|A\| \text{ (by triangle inequality and the fact that } \|v+u\| \leq 2) \\
&\leq \varepsilon\|A\| + 2\gamma\|A\| \\
&\leq \varepsilon\|A\| + 4\sqrt{2}\sqrt{\varepsilon}\|A\|.
\end{aligned}$$

■

Tightness of the Guarantee in Theorem 30. We close out this section by showing that the dependence on $\sqrt{\sigma\|\mu\|}$ in our bound on mean estimation is necessary even information theoretically. In what follows A will be an $n \times m$ matrix with $\mu = \textsc{Mean}(A)$ such that $\Pi^* = \mu\mu^\top/\|\mu\|^2$ has small norm, i.e., $\|\Pi^*\|_{\infty\to 2} = \kappa$. Furthermore, let $C = \mu\mathbf{1}^\top$ and define $\sigma = \|A - C\|/\sqrt{m}$. We will prove the following.

Theorem 57 *Fix $q = \infty$. Let $\mathcal{M}$ be the set of $n \times m$ matrices A with mean μ that satisfies $\|\mu\| \in [1, 2]$, variance σ^2 around the mean that satisfies $\sigma \in (0, 1/6]$, and the subspace spanned by μ being κ-robust. Call a perturbation $\tilde{A}$ of $A \in \mathcal{M}$ of be valid if $\|A - \tilde{A}\|_\infty \leq \delta$. Then, any algorithm that takes as input a valid perturbation $\tilde{A}$ of a matrix $A \in \mathcal{M}$ and either certifies that the data is poisoned, i.e., $\|A - \tilde{A}\| > 8\sigma\sqrt{m}$ or outputs an estimate $\hat{\mu}$ of μ must incur an error of*

$$\|\mu - \hat{\mu}\| \geq \Omega\Big((1 + \frac{\kappa\delta}{\sigma})\max(\sigma, \sqrt{\sigma\|\mu\|})\Big),$$

where $\mu = \textsc{Mean}(A)$.

Proof We will establish the lower bound by constructing two matrices A and $\tilde{A}$, both of which lie in the set $\mathcal{M}$, satisfy $\|A - \tilde{A}\| = \delta$ for $\kappa\delta = O(\sigma)$, but have means separated by $\Omega(\max(\sigma, \sqrt{\sigma\mu_{\max}}))$, where $\mu_{\max}$ is the maximum ℓ_2 norm among $\textsc{Mean}(A)$ and $\textsc{Mean}(\tilde{A})$. In this case, given either A or $\tilde{A}$ as input, any provably robust certification procedure cannot output Bad Input and must output an estimate $\hat{\mu}$, thereby making $\Omega(\max(\sigma, \sqrt{\sigma\mu_{\max}}))$ error on either A or $\tilde{A}$. We next describe our construction.

For a k to be determined later, let $\mu_1 = (1/\sqrt{k}, 1/\sqrt{k}, \ldots, 1/\sqrt{k}, 0, 0, \ldots, 0)$. Hence, μ_1 is a unit length sparse vector with $\|\mu_1\|_1 = \sqrt{k}$. We define the set of m points in A by generating i.i.d. points of the form $\mu_1 + g$, where g is a mean zero Gaussian with variance 0 in the first k coordinates and variance σ^2 in the other coordinates. Then it is a standard fact that with high probability $\|A - \mu_1\mathbf{1}^\top\| \leq 2\sigma\sqrt{m}$ and that Mean(A) will be $\sigma\sqrt{d/m} = o(\sigma)$-close to μ_1. Next we define the set of points in $\tilde{A}$ to be $\tilde{A}_j = A_j + \delta sgn(\mu_1)$, where $sgn(\mu_1)$

is a ± 1 vector representing the sign of the corresponding coordinate of μ_1. Here we can arbitrarily set $sgn(0)$ to be $+1$. It is easy to see that $\textsc{Mean}(\tilde{A})$ will be $o(\sigma)$-close to $\mu_2 = \mu_1 + \delta sgn(\mu_1)$, and that $\|\tilde{A} - \mu_2 \mathbf{1}^\top\| \leq 2\sigma\sqrt{m}$. Next notice that

$$\|\mu_2\|^2 = 1 + \delta^2 n + \delta\sqrt{k}$$
$$\|\mu_2\|_1 = \sqrt{k} + \delta\sqrt{k}.$$

By setting $\delta\sqrt{k} = 3\sigma$ and $\delta n = \sqrt{k}$ we ensure that $\|\mu_2\| \in [1, 2]$ and $\|\mu_2\|_1 \leq 2\sqrt{k}$. Hence we get that for the matrix $\Pi = \mu_2\mu_2^\top/\|\mu_2\|^2$, $\|\Pi\|_{\infty\to 2} \leq 2\sqrt{k}$. Hence, both A and $\tilde{A}$ lie in the set $\mathcal{M}$ with sparsity bound $\kappa = 2\sqrt{k}$. Furthermore, the fact that $\delta\sqrt{k} = 3\sigma$, ensures that $\kappa\delta \leq 6\sigma$. Finally, notice that the difference between two means is

$$\|\mu_1 - \mu_2\| = \delta\sqrt{n} = \sqrt{\delta}k^{1/4} = \sqrt{3\sigma} = \Omega\Big((1 + \frac{\kappa\delta}{\sigma})\max(\sigma, \sqrt{\sigma\mu_{\max}})\Big).$$

■

We end this section by showing that via an (inefficient) algorithm one can get the same guarantee for mean estimation as in Theorem 30 without the need for certification.

Theorem 58 (Information Theoretic Upper Bound) *Let A be an $n \times m$ matrix representing m data points in n dimensions and let μ be the mean of the data points in the matrix A with C representing the $n \times m$ matrix with each column being μ. Let $\Pi^* = \mu\mu^\top/\|\mu\|^2$ be the one dimensional subspace denoting the projection onto μ and assume that $\|\Pi^*\|_{q\to 2} \leq \kappa$, for some $q \geq 2$. Let $\tilde{A}$ be the given input such that for every column $j \in [m]$ we have $\|A_j - \tilde{A}_j\|_q \leq \delta$. Furthermore, let $\sigma^2 > 0$ be a given upper bound on the variance of the data around the mean, i.e., $\|A - C\| \leq \sigma\sqrt{m}$. Then there is an (inefficient) exponential time algorithm that takes $\tilde{A}$ as input and outputs an estimate $\hat{\mu}$ of the true mean μ such that*

$$\|\hat{\mu} - \mu\|_2 \leq O(c_q)(1 + \frac{\kappa\delta}{\sigma})\max\Big(\sigma, \sqrt{\sigma\|\mu\|}\Big),$$

where c_q is a constant that depends on q. In particular, the above implies a relative error guarantee of

$$\frac{\|\hat{\mu} - \mu\|_2}{\|\mu\|} \leq O(c_q)(1 + \frac{\kappa\delta}{\sigma})\max\Big(\frac{\sigma}{\|\mu\|}, \sqrt{\frac{\sigma}{\|\mu\|}}\Big).$$

Proof In order to establish the theorem above we first optimize (70) exactly. Let A' be the matrix and Π be the projection that achieve the minimum of (70). Then we have that

$$\|A' - \Pi A'\| \leq \|A - \Pi^* A\|$$

Furthermore, we also have that

$$\begin{aligned}\|A - \Pi^* A\| &\leq \|A - C\| + \|C - \Pi^* A\| \\ &\leq 2\|A - C\| \\ &\leq 2\sigma\sqrt{m}.\end{aligned}$$

Hence both A and A' have good projections onto rank-1 subspaces. Plugging into Lemma 56 we get that

$$\|A - \Pi A\| \leq O(\varepsilon + \sqrt{\varepsilon})\|A\| + \frac{8\kappa\delta\sqrt{m}}{\sqrt{\varepsilon}},$$

where $\varepsilon = 2\sigma\sqrt{m}/\|A\|$. Hence, letting $\hat{\mu} = \textsc{Mean}(\Pi A')$ we get from (119) that

$$\begin{aligned}\|\mu - \hat{\mu}\| &\leq \frac{1}{\sqrt{m}}\|A - \Pi A\| + c_q\kappa\delta \\ &\leq O(\varepsilon + \sqrt{\varepsilon})\frac{\|A\|}{\sqrt{m}} + \frac{8\kappa\delta\sqrt{m}}{\sqrt{\varepsilon}} + c_q\kappa\delta.\end{aligned}$$

The rest of the argument proceeds exactly as in the Proof of Theorem 30 by writing $\|A\|$ in terms of ε, as done in (120). ■