List Decodable Subspace Recovery

Prasad Raghavendra RAGHAVENDRA@BERKELEY.EDU
UC Berkeley EECS

Morris Yau MORRISYAU@BERKELEY.EDU
UC Berkeley EECS

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Learning from data in the presence of outliers is a fundamental problem in statistics. In this work, we study robust statistics in the presence of overwhelming outliers for the fundamental problem of subspace recovery. Given a dataset where an α fraction (less than half) of the data is distributed uniformly in an unknown k dimensional subspace in d dimensions, and with no additional assumptions on the remaining data, the goal is to recover a succinct list of $O(\frac{1}{\alpha})$ subspaces one of which is close to the original subspace. We provide the first polynomial time algorithm for the 'list decodable subspace recovery' problem, and subsume it under a more general framework of list decoding over distributions that are "certifiably resilient" capturing state of the art results for list decodable mean estimation and regression. In a independent and concurrent work Bakshi and Kothari (2020) obtain similar results with the sum of squares method.

Keywords: High Dimensional Statistics, Robustness, Convex Relaxations

1. Introduction

A large hurdle for the deployment of algorithms in high dimensional statistics is their susceptibility to outliers. The central paradigm of statistical inference is finding the parameters of a statistical model given data. A long line of work in the robust statistics literature, models 'real world' data as a distributional perturbation of a parameterized generative model $\mathcal{D}$. Here robust estimators have been designed for decades, see Huber (2011). Under the classic "Huber Contamination Model" data $X_1, ..., X_N$ is drawn i.i.d from a distribution that is a mixture of an inlier distribution $\mathcal{D}$ belonging to a parameterized family, and an outlier distribution A which can be chosen adversarially.

$$X_1, ..., X_N \sim \alpha D + (1-\alpha)A$$

Here α is a constant in $[0, 1]$ corresponding to the fraction of the dataset that is comprised of inliers and is presumed to be known. The goal is to recover the relevant parameters of $\mathcal{D}$ such as mean, covariance, etc. For $\alpha > 1/2$ the inliers overwhelm the outliers, and we are in the setting of classical robust statistics for which a recent flurry of computationally tractable algorithms have been developed, see survey Diakonikolas and Kane (2019).

Less well understood are the settings in which high dimensional statistical inference is possible in the presence of overwhelming outliers. For $\alpha < 1/2$, we are in the setting of overwhelming outliers, where returning a single estimator for relevant parameters of $\mathcal{D}$ is impossible as the outlier distribution A can belong to the same distributional family as D but with wildly different parameters. With the outliers outnumbering the inliers, there is no unique identification of parameters, a problem we refer to as a "failure of identifiability".

However, one could hope to output a short list of estimators of length $O(\frac{1}{\alpha})$, one of which is guaranteed to be close to the true parameters of $\mathcal{D}$. Charikar et al. (2017) introduced this relaxed notion of recovery under the umbrella of "list decodable robust statistics".

A first observation is that list decoding is at least as hard as identifying the parameters of mixture models. With nothing but a planted set of statistical inliers, the outliers can assume any configuration. A remarkably benign configuration is for the outliers to be arranged as independent mixtures. In this manner, the gaussian mixture model is a special case of list decodable mean estimation, the mixtures of linear regressions is a special case of list decodable regression, and likewise subspace clustering is a special case of list decodable subspace recovery. Naturally, any theoretical guarantee for list decodable robust statistics carries directly over to its mixture model counterpart. Although the converse is evidently false, the chief intellectual thrust of list decodable robust statistics is to establish the settings wherein statistical inference in the presence of overwhelming adversarial outliers is computationally no harder than clustering, a remarkable assertion, especially in light of the settings where list decoding is information theoretically impossible (see eg. Karmalkar et al. (2019) Diakonikolas et al. (2018)Kothari and Steinhardt (2017)).

Results. In this paper we build on a series of works for list decoding of mean estimation Charikar et al. (2017) Diakonikolas et al. (2018) Kothari and Steinhardt (2017), regression Karmalkar et al. (2019) Raghavendra and Yau (2020), and tackle the natural problem of subspace recovery. Informally, given a dataset for which an α fraction is drawn uniformly in a k dimensional subspace in d dimensions, denoted U, and with no additional assumptions on the remaining data, our algorithm outputs a succinct list of $O(\frac{1}{\alpha})$ candidate subspaces one of which is close to the true generative U. Our algorithm is computationaly tractable, runs in polynomial time in both d and k. Furthermore, our algorithm is robust to additive noise, well conditioned linear transformations of the underlying inlier distribution, and succeeds even under the substantially more demanding corruption model where the adversary can simulate any $(1-\alpha)$ total variation distance corruption of the data.

Our main algorithmic result is an algorithm for list-decodable subspace recovery.

Theorem 1 *Suppose $\{X_i^*\}_{i\in[N]}$ are drawn from $N(0,_d)$ and let P be a projection to a k-dimensional subspace of $\mathbb{R}^d$. Let $\{X_i | i \in [N]\}$ be generated by setting*

$$X_i = PX_i^* + \gamma_i$$

for some additive noise γ_i satisfying

$$\sum_{i=1}^{N} \|\gamma_i\|^2 = \varepsilon N$$

Let $\{\tilde{X}_i | i \in [N]\}$ be a set of points such that there exists a subset $\mathcal{S} \in [N]$ of size $|\mathcal{S}| \geq \alpha N$ with $\tilde{X}_i = X_i$ for all $i \in \mathcal{S}$. For all $\eta > 0$, given $N > k^{O(1/\eta^4)}$ samples, there is an algorithm running in time $N^{O(1/\eta^4)}$ that computes a list of $O(1/\alpha)$ projection matrices $\{\Pi_1, \ldots, \Pi_\ell\}$ such that

$$\min_j \|P - \Pi_j\|^2 \leq O\left(\frac{\varepsilon}{\eta^2\alpha^5} + \frac{4ck\eta}{\alpha^5}\right)$$

Note that the noise model is perhaps the strongest possible, in that the adversarial corruptions can depend arbitrarily on the samples $\{X_i\}_{i\in[N]}$ and it also includes an additive noise of γ_i. For concreteness, if we consider the case with no additive noise ($\varepsilon = 0$) then the

algorithm recovers a projection Π_j with $\|P - \Pi_j\| \leq O(\sqrt{k})$ in time that is polynomial in k, d. More generally, the Gaussian distribution $N(0, I_d)$ can be replaced by a distribution whose anti-concentration can be efficiently certified by sum-of-squares proofs (see Appendix for a formal definition). Specifically, our results hold for any well conditioned linear transformation of a spherically symmetric distribution with sub-exponential tails (see lemma 9.1 of Raghavendra and Yau (2020)).

Conceptually, we formally state the notion of SoS certifiable resilience, and use it to derive a general algorithm for list-decoding via SoS. While the ideas behind the general algorithm are implicit in Karmalkar et al. (2019), we believe the notion of SoS certifiable resilience gives conceptual clarity and might be useful in further applications of the SoS SDPs. We apply the framework of SoS-certifiable resilience in our presentation of the result for subspace recovery.

Finally we exhibit a lower bound showing that list decodable subspace recovery is impossible even if the inlier distribution is the uniform over the boolean hypercube (see lemma 14).

1.1. Related Work

Subspace Recovery. Here we discuss related work for the problem of subspace recovery, and highlight key similarities and differences with list decoding. The literature on subspace recovery is vast and we do not attempt a full overview–for a survey, see Elhamifar and Vidal (2012). Despite the vast literature, the key takeaway is that existing methods for subspace recovery, to the best of our knowledge, fail in the presence of overwhelming adversarial outliers.

In the worst case setting, Hardt and Moitra (2013) explore robust subspace recovery in a purely deterministic model where inliers are arranged in general position within a subspace and outliers are in general position in the ambient space. They provide an algorithm recovering the planted subspace provided $\alpha \geq \frac{k}{d}$. Their result is essentially optimal as it is Small Set Expansion Hard to recover the subspace if the fraction of outliers is any larger. Although there is no direct comparison with list decodable subspace recovery, the hardness result is solid evidence that subspace recovery in the presence of overwhelming outliers is hard without additional statistical assumptions on the inliers. Just as worst case assumptions are arguably overly pessimistic, average case assumptions are arguably overly optimistic.

In the statistics literature, subspace clustering is a problem where data is distributed in a union of subspaces in high dimensions where the distribution of points within subspaces, and the relative orientation of subspaces are subject to theoretical assumptions. The goal is to cluster points into their respective subspaces. This is in contrast with the goals of list decoding, which is a parameter estimation task.

Statistical approaches model the data according to a mixture of degenerate gaussians. In a sense, this modeling assumption is necessary as subspace clustering is information theoretically impossible even over natural distributional families (see 14).

A representative approach is Sparse Subspace Clustering (SSC) Elhamifar and Vidal (2012) and its robust variant (RSSC) Soltanolkotabi et al. (2013) which uses techniques from sparse and low rank recovery algorithms. RSSC considers subspace clustering in the presence of outliers distributed uniformly on the unit sphere. This stands in contrast with list decodable subspace recovery where the outliers are adversarially introduced. An overview of spectral clustering algorithms can be found in Vidal (2011).

Finally, there are subspace clustering algorithms that either lack provable gaurantees or are computationally intractable. We list a few notable examples. Generalized Principal

Component Analysis Vidal et al. (2012) Ozay et al. (2010) is an algebraic geometric algorithm that treats subspace clustering as a polynomial fitting problem. Although its recovery guarantees and assumptions are minimal, it's fragile to outliers and its runtime is exponential in k. K-Subspaces Tseng (1999) is a generalization of K-means that approaches subspace clustering as a nonconvex optimization. As a consequence, it is sensitive to initalization and fragile to outliers. Other iterative algorithms include Agarwal and Mustafa (2004) Bradley and Mangasarian (2000). Examples of EM style statistical approaches include Mixtures of Probabilistic PCA Tipping and Bishop (1999) and other nonconvex approaches include Agglomerative Lossy Compression Ma et al. (2007). Unfortunately, it is notoriously difficult to prove the convergence of EM and other nonconvex methods to global optima of the likelihood function.

List Decodable Learning, Resilience, and the Sum of Squares. The notion of list decodable learning was introduced by Balcan et al. Balcan et al. (2008) for clustering problems. List learning was extended to small α robust statistics in Charikar et al. (2017). They obtained algorithms for list decodable mean estimation, planted partition problems, subsumed under a general stochastic convex optimization framework. (also see Steinhardt et al. (2016, 2017b)). The same model of *list-decodable learning* has been studied for the case of mean estimation Kothari and Steurer (2017) and Gaussian mixture learning Kothari and Steinhardt (2017); Diakonikolas et al. (2018).

The notion of *resilience* was initialy defined in Steinhardt et al. (2017a) for robust estimation and extended in Zhu et al. (2019). Furthermore, there has been a sequence of works developing the sum of squares method for robust statistics Kothari and Steurer (2017); Kothari and Steinhardt (2017); Kothari et al. (2017); Hopkins and Li (2018).

Robust Subspace Recovery. Furthermore, there is a rich literature on the robust subspace recovery problem. Here, the inliers and outliers are modeled according to a variety of assumptions. Of particular interest are settings where the outliers are modeled adversarially Maunu and Lerman (2019) Maunu et al. (2017). See also smoothed versions of robust subspace recovery Bhaskara et al. (2019). For a survey see Lerman and Maunu (2018).

Independent Work. Bakshi and Kothari (2020) obtain similar results for list decodable subspace recovery using the sum of squares method.

2. List Decoding via SoS

Let $Z_1, \ldots, Z_N$ be samples from a distribution $\mathcal{D}$ over $\mathbb{R}^d$. Often, parameters $\Theta^* \in \mathbb{R}^m$ associated with the distribution $\mathcal{D}$ can be expressed as minima of a cost function associated with each data point. Specifically, let $\Phi(\Theta, Z)$ be a cost function such that the true parameters of the distribution can be expressed as,

$$\Theta^* = \operatorname*{argmin}_{\Theta \in \mathcal{V}} \frac{1}{N} \sum_{i \in [N]} \Phi(Z_i, \Theta) \tag{2.1}$$

where $\mathcal{V} \in \mathbb{R}^m$ is the domain of the parameters.

Since the sum-of-squares proof system can certify facts about low-degree polynomials, we will setup the problem of parameter estimation in this setting. First, we assume that Φ is specified by a polynomial in Θ and Z. Second, we assume that the parameters Θ belong to a semi-algebraic set that is specified by a set of polynomial inequalities,

$$\mathcal{V} = \{q_j(\Theta) \geq 0 | 1 \leq j \leq |\mathcal{V}|\}$$

Notice that equalities $q(\Theta) = 0$ can be expressed using two inequalities $q(\Theta) \geq 0$ and $-q(\Theta) \geq 0$. With this setup, the problem of estimating the parameters Θ^* reduces to solving an optimization problem with polynomial objective and polynomial constraints.

In this work, we will be interested in parameter estimation when an overwhelming fraction of input data is adversarially corrupted. Let $\{\tilde{Z}_1, \ldots, \tilde{Z}_N\}$ be a corrupted data set wherein all but $1 - \alpha$-fraction of the samples are adversarially corrupted. Specifically, for some subset $\mathcal{S} \subset [N]$ with $|\mathcal{S}| \geq \alpha N$, we have that

$$\tilde{Z}_i = \begin{cases} Z_i & \text{if } i \in \mathcal{S} \\ \text{arbitrary} & \text{if } i \notin \mathcal{S} \end{cases}$$

In the list-decodable recovery problem, we are to recover a small list of candidate assignments $\{\Theta_1, \ldots, \Theta_t\}$ for the parameter such that there exists at least one candidate Θ_i close to the true value of the parameter Θ^* on the original data $\{Z_1, \ldots, Z_N\}$.

Parameter estimation in presence of adversarially chosen outliers poses two challenges. First, the algorithm needs to identify which subset $\mathcal{S}$ of αN samples $\tilde{Z}_i$ are uncorrupted. Second, even if the algorithm identifies the subset $\mathcal{S}$ of samples exactly, it is unclear if the surviving uncorrupted samples still yield a good estimate for the parameters.

The problem of identifying the correct subset of samples $\mathcal{S}$ can also be posed as a polynomial optimization problem. The idea is as follows, introduce variables w_i for each sample $\tilde{Z}_i$ to indicate whether the sample is corrupted or not. Each variable w_i takes a boolean value, which can be enforced by the polynomial constraint

$$\text{(Booleanness)} \quad P_{\mathsf{bool}}(w_i) \stackrel{\text{def}}{=} w_i^2 - w_i = 0$$

Furthermore, at least an α-fraction of the samples are uncorrupted yielding the constraint

$$\text{(Sum)} \quad P^{(\alpha)}_{\mathsf{sum}}(\mathbf{w}) \stackrel{\text{def}}{=} \sum_{i \in [N]} w_i \geq \alpha N$$

This formulation underlies all applications of SoS SDPs to robust statisics Hopkins and Li (2018); Kothari and Steurer (2017); Kothari and Steinhardt (2017).

Given the subset $\mathcal{S} \subset \{Z_1, \ldots, Z_N\}$ of uncorrupted samples, a natural estimate of the parameters would be to minimize the total cost for samples within $\mathcal{S}$. Specifically, one can construct an estimate $\Theta_{|\mathcal{S}}$

$$\Theta^*_{|\mathcal{S}} = \operatorname*{argmin}_{\Theta \in \mathcal{V}} \frac{1}{|\mathcal{S}|} \cdot \sum_{Z \in \mathcal{S}} \Phi(Z, \Theta) \tag{2.2}$$

This corresponds to a polynomial constraint of the form,

$$\text{(Cost)} \qquad P^{(\varepsilon)}_{\mathsf{cost}}(\mathbf{w}, \Theta) : \sum_{i \in [N]} w_i \Phi(\tilde{Z}_i, \Theta) \leq \varepsilon N$$

Here we use εN as a generic upper bound, the correct value for the upper bound would depend on the application at hand.

The language of polynomials is very powerful in that a large number of robust parameter estimation problems can be posed as multi-variate constrained polynomial optimization. On the flipside, it is NP-hard to solve these polynomial optimization problems. The Sum-of-Squares SDP hierarchy and associated sum-of-squares proofs provides a family of efficient algorithms to imperfectly reason about such systems of polynomials.

2.1. Sum-of-Squares SDP hierarchy

Pseudoexpectations. The sum-of-squares SDP relaxations for a system of polynomial inequalities $\mathcal{P}$ are a sequence of increasingly stronger SDP relaxations. The degree ℓ SoS SDP relaxation is intended to find the degree ℓ-moments of a "probability distribution" over solutions to the system $\mathcal{P}$. While the SDP relaxation returns a set of candidate "degree ℓ moments" of a distribution, the moments are *pseudo-moments* in that there might exist no distribution over solutions to $\mathcal{P}$ with those moments. It is notationally convenient to state the degree ℓ SoS SDP relaxation in terms of a pseudo-expectation functional $\tilde{\mathbb{E}}$.

Definition 2 *A degree ℓ pseudoexpectation $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leq \ell} \to \mathbb{R}$ satisfying $\mathcal{P}$ is a linear functional over polynomials of degree at most ℓ satisfying*

1. *(Normalization)* $\tilde{\mathbb{E}}[1] = 1$,
2. *(Constraints of $\mathcal{P}$)* $\tilde{\mathbb{E}}[p(x)a^2(x)] \geq 0$ *for all $p \in \mathcal{P}$ and polynomials a with* $\deg(a^2 \cdot p) \leq \ell$,
3. *(Non-negativity on square polynomials)*$\tilde{\mathbb{E}}[q(x)^2] \geq 0$ *whenever* $\deg(q^2) \leq \ell$.

For any fixed $D \in \mathbb{N}$, given a polynomial system,one can efficiently compute a degree D pseudo-expectation in polynomial time.

Fact 3 *(Nesterov (2000), Parrilo (2000), Lasserre (2000/01), Shor (1987)). For any $n, D \in \mathbb{Z}^+$, let $\tilde{\mathbb{E}}_\zeta$ be degree D pseudoexpectation satisfying a polynomial system $\mathcal{P}$. Then the following set has a $n^{O(D)}$-time weak separation oracle (in the sense of Grötschel et al. (1981)):*

$$\{\tilde{\mathop{\mathbb{E}}_{\zeta}}(1, x_1, x_2, ..., x_n)^{\otimes D} \mid \text{degree } D \text{ pseudoexpectations } \tilde{\mathop{\mathbb{E}}_{\zeta}} \text{ satisfying } \mathcal{P}\}$$

Armed with a separation oracle, the ellipsoid algorithm finds a degree D pseudoexpectation in time $n^{O(D)}$, which we call the degree D sum-of-squares algorithm.

Roughly speaking, the degree D-pseudoexpectation functional yields the "degree D moments" of a potential distribution over solutions to the polynomial system. However, there might not exist any probability distribution with these moments. Although, the $\tilde{\mathbb{E}}$ functional does not correspond to an expectation over actual solutions in general, this intuition is useful to keep in mind, and we will appeal to it whenever needed in this overview. To reason about the properties of pseudo-expectations, one harnesses the dual object namely sum-of-squares proofs. We turn our attention to sum-of-squares proofs now.

Sum-of-Squares Proofs. For any nonnegative polynomial $p(x) : \mathbb{R}^d \to \mathbb{R}$, one could hope to prove its nonnegativity by writing $p(x)$ as a sum of squares of polynomials $p(x) = \sum_{i=1}^m q_i(x)^2$ for a collection of polynomials $\{q_i(x)\}_{i=1}^m$. Such a proof would be succinct and easy to verify. Unfortunately, there exist nonnegative polynomials with no sum of squares proof even for $d = 2$. Nevertheless, there is a generous class of nonnegative polynomials that admit a proof of positivity via a proof in the form of a sum of squares. The key insight of the sum of squares algorithm, is that these sum of squares proofs of nonnegativity can be found efficiently provided the degree of the proof is not too large. We begin with a rough overview of sum of squares proofs, their dual object pseudoexpectations, and then present the guarantees of the SoS algorithm.

Definition 4 *(Sum of Squares Proof) Let $\mathcal{A}$ be a collection of polynomial inequalities $\{p_i(x) \geq 0\}_{i=1}^m$. A sum of squares proof that a polynomial $q(x) \geq 0$ for any x satisfying the inequalities in $\mathcal{A}$ takes on the form*

$$\left(1 + \sum_{k \in [m']} b_k^2(x)\right) \cdot q(x) = \sum_{j \in [m'']} s_j^2(x) + \sum_{i \in [m]} a_i^2(x) \cdot p_i(x)\,,$$

where $\{s_j(x)\}_{j\in[m'']}, \{a_i(x)\}_{i\in[m]}, \{b_k(x)\}_{i\in[m']}$ are real polynomials. If such an expression were true, then $q(x) \geq 0$ for any x satisfying $\mathcal{A}$. We call these identities sum of squares proofs, and the degree of the proof is the largest degree of the involved polynomials $\max\{\deg(s_j^2), \deg(a_i^2 p_i)\}_{i,j}$. Naturally, one can capture polynomial equalities in $\mathcal{A}$ with pairs of inequalities. We denote a degree ℓ sum of squares proof of the positivity of $q(x)$ from $\mathcal{A}$ as $\mathcal{A} \left|\frac{x}{\ell}\right. \{q(x) \geq 0\}$ where the superscript over the turnstile denote the formal variable over which the proof is conducted. This is often unambiguous and we drop the superscript unless otherwise specified.

Sum of squares proofs can also be strung together and composed according to the following convenient rules.

Fact 5 *For polynomial systems $\mathcal{A}$ and $\mathcal{B}$, if $\mathcal{A} \left|\frac{x}{d}\right. \{p(x) \geq 0\}$ and $\mathcal{B} \left|\frac{x}{d'}\right. \{q(x) \geq 0\}$ then $\mathcal{A} \cup \mathcal{B} \left|\frac{x}{\max(d,d')}\right. \{p(x) + q(x) \geq 0\}$. Also $\mathcal{A} \cup \mathcal{B} \left|\frac{x}{dd'}\right. \{p(x)q(x) \geq 0\}$*

Sum of squares proofs yield a framework to reason about the properties of pseudo-expectations, that are returned by the SoS SDP hierarchy.

Fact 6 *(Informal Soundness) If $\mathcal{A} \left|\frac{x}{r}\right. \{q(x) \geq 0\}$ and $\tilde{\mathbb{E}}$ is a degree-ℓ pseudoexpectation operator for the polynomial system defined by $\mathcal{A}$, then $\tilde{\mathbb{E}}[q(x)] \geq 0$.*

2.2. Certifiable Resilience

Returning back to our problem of parameter estimation from corrupted data, we need to address the issue that the fragment of uncorrupted data left might be insufficient to faithfully recover the parameter Θ. More precisely, we will need to make an assumption that the estimate $\Theta^*_{|S}$ (in (2.2)) is close to the true estimate Θ^* (in (2.1))).

The notion of *resilience* introduced by Steinhardt et al. (2017a) captures this idea. To exploit the power of sum-of-squares SDP hierarchies, one needs a stronger notion of *certifiable resilience*. A data set is *certifiable resilience* where the dataset is not only resilient, but there is also an efficiently verifiable certificate/proof of its resilience. In particular, we will be define the notion of *Sum-of-Squares certifiable resilience*. The formal definition of certifiable resilience is as follows.

Definition 7 *(SoS certifiable resilience) Fix $\alpha \in (0,1]$ and $\varepsilon, \delta > 0$. A dataset $\{Z_1, \ldots, Z_N\} \in \mathbb{R}^d$ is said to admit a degree D SoS proof of $(\alpha, \varepsilon, \delta)$-resilience if the following polynomial system:*

$$\left\{ \begin{array}{lll} P^{(\alpha)}_{\mathsf{sum}}(w) & \stackrel{\text{def}}{=} \sum_{i\in[N]} w_i - \alpha N & \geq 0 \\ P_{\mathsf{bool}}(w_i) & \stackrel{\text{def}}{=} w_i^2 - w_i & = 0 \quad \forall i \in [N] \\ P^{(\varepsilon)}_{\mathsf{cost}}(\mathbf{w}) & \stackrel{\text{def}}{=} \varepsilon N - \sum_{i\in[n]} w_i \Phi(Z_i, \Theta) & \geq 0 \\ & q(\Theta) & \geq 0 \quad \forall q \in \mathcal{V} \end{array} \right\}$$

can be used to show that $(\sum_i w_i) \cdot \|\Theta - \Theta^*\|^2 \leq \delta N$ *using a sum-of-squares polynomial identity of the form:*

$$\delta N - \left(\sum_i w_i\right) \|\Theta - \Theta^*\|^2 = c(\mathbf{w}, \Theta) \cdot P^{(\varepsilon)}_{\mathsf{cost}}(\mathbf{w}, \Theta) + \sum_{i \in [N]} b_i(\mathbf{w}, \Theta) \cdot P_{\mathsf{bool}}(w_i) + \sum_{q \in \mathcal{V}} A_q(\mathbf{w}, \Theta) \cdot q(\Theta) + \lambda \cdot P^{(\alpha)}_{\mathsf{sum}}(\mathbf{w}) + \lambda_0 \tag{2.3}$$

where $\lambda, \lambda_0 \in \mathbb{R}^+$, $c(\mathbf{w}, \Theta)$ *and* $A_q(\mathbf{w}, \Theta)$ *are sum of squares polynomials and* $b_i(\mathbf{w}, \Theta)$ *are arbitrary polynomials in* $\mathbf{w}, \Theta$. *Furthermore, the degree of all the terms in the equality are at most* D.

Remark 8 *We wish to stress on the important distinction between the SoS certificate* (2.3) *and the standard notion of SoS proofs. In* (2.3)*, the coefficient of* $P^{(\alpha)}_{\mathsf{sum}}$ *is necessarily a real number* λ*, while a general SoS proof would allow the coefficient of* $P^{(\alpha)}_{\mathsf{sum}}$ *to be an arbitrary SoS polynomial. Operationally, if one is constructing the SoS certificate by a proof, this restriction translates to never multiplying* $P^{(\alpha)}_{\mathsf{sum}}$ *constraint with any polynomial.*

2.3. List-decoding

We will now present an SoS based algorithm for list-decoding the parameters Θ under the assumption of certifiable resilience. The essential ingredients of the algorithm are implicit in Karmalkar et al. (2019), but we reformalize the ideas in generality, under the notion of *certifiable resilience*. The notion of certifiable resilience clarifies design of algorithms for list-decodable learning via SoS, and is also useful in presenting our work on subspace recovery.

The general idea behind the algorithm is to solve a sufficiently high-degree SoS SDP relaxation of the polynomial system underlying *certifiable resilience*. This yields a collection of pseudomoments from which we will recover a list of assignments for the parameter Θ, of which one is close to the true value Θ^*.

Frobenius Minimization. This program faces an immediate bottleneck. Even if the pseudo-expectation functional corresponded to a true distribution over solutions to the polynomial system, it is conceivable that the distribution does not include the solution Θ^*. Specifically, the distribution might completely ignore the αN uncorrupted data points, and instead return feasible solutions among the rest. To overcome this issue, we need to find pseudo-expectations that are *comprehensive* in that every valid solution is in their support. This is achieved by finding a pseudo-expectation functional of maximum entropy or minimum Frobenius norm, among all pseudo-expectation functionals that satisfy the constraints. This technique first used in the work of Hopkins and Steurer Hopkins and Steurer (2017), was also used in the two prior works on list-decoding via SoS SDP hierarchy Raghavendra and Yau (2020); Karmalkar et al. (2019). In particular, both these works show that the pseudo-expectation functional that minimizes the Frobenius norm necessarily has good correlation with every possible solution to the polynomial system. Formally, they show the following.

Lemma 9 *(Comprehensive Pseudodistributions are Correlated with Inliers Raghavendra and Yau (2020); Karmalkar et al. (2019)) Let* $\mathcal{P}$ *be a polynomial system in variables* $\mathbf{w} = \{w_i\}_{i \in [N]}$ *and a set*

of indeterminates Θ*, that contains the set of inequalities:*

$$\mathcal{P} = \begin{cases} w_i^2 = w_i & \forall i \in [N] \\ \sum_{j=1}^{N} w_j = \alpha N \end{cases}$$

Let $\tilde{\mathbb{E}}_\zeta : \mathbb{R}[\{w_i\}_{i\in[N]}]^{\leq D} \to \mathbb{R}$ *denote a degree* D *pseudoexpectation that satisfies* $\mathcal{P}$ *and minimizes the norm* $\|\tilde{\mathbb{E}}_\zeta[w]\|$*. If* $\mathbf{w}' = (w'_1, \ldots, w'_N) \in \{0,1\}^N$ *and* Θ' *is a satisfying assignment to* $\mathcal{P}$ *then the* $\tilde{\mathbb{E}}_\zeta$ *has non-negligible support on* $\mathbf{w}'$*,*

$$\tilde{\mathbb{E}}_\zeta \left[\frac{1}{\alpha N} \sum_{i=1}^{N} w_i w'_i \right] \geq \alpha \tag{2.4}$$

Rounding. Assuming we have the moments of a distribution over solutions to the polynomial system, the goal of rounding is to extract each of the solutions to the system. Both the previous works Raghavendra and Yau (2020); Karmalkar et al. (2019) employ the idea of conditioning SoS SDP relaxations towards rounding the SDP solution. Intuitively, the idea is to pick a sample Z_i, and condition the pseudoexpectation on the sample Z_i being an inlier, i.e., condition on the event that $w_i = 1$.

Formally, let $\tilde{\mathbb{E}} : \mathbb{R}[\mathbf{w}, \Theta] \to \mathbb{R}$ denote the pseudo-expectation functional on polynomials of degree at most $D + 1$ in variables $\mathbf{w} := \{w_i\}_{i\in[N]}$ and Θ. Pseudo-expectation functionals (equivalently SoS SDP solutions), can be conditioned on low-degree events. For example, for any $i \in [N]$, the conditioned pseudoexpectation functional $\tilde{\mathbb{E}}[\cdot|w_i = 1]$ is constructed as follows,

$$\tilde{\mathbb{E}}\left[p(\mathbf{w}, \Theta)|w_i = 1\right] \stackrel{\text{def}}{=} \frac{\tilde{\mathbb{E}}[p(\mathbf{w}, \Theta) \cdot w_i]}{\tilde{\mathbb{E}}[w_i]} \text{ for all polynomials } p \in \mathbb{R}[\mathbf{w}, \Theta] \text{ with } \deg(p) \leq D \tag{2.5}$$

For a degree $D + 1$ pseudoexpectation functional $\tilde{\mathbb{E}}$, $\tilde{\mathbb{E}}[\cdot|w_i = 1]$ is a degree D pseudoexpectation functional.

The two works Raghavendra and Yau (2020); Karmalkar et al. (2019) analyze the rounding schemes differently, and we follow the simpler analysis in Karmalkar et al. (2019). We are now ready to formally describe the list-decoding algorithm for certifiably resilient datasets.

Algorithm 1 list Decoding

Result: $\Theta \in \mathbb{R}^m$ such that with probability atleast α, $\|\Theta - \Theta^*\| \leq \delta$

Inputs: Parameters α, ε and a corrupted data set $\mathcal{D} = \{\tilde{Z}_i\}_{i=1}^N$ with αN samples from a $(\alpha, \varepsilon, \delta)$-resilient dataset $\{Z_1, \ldots, Z_N\}$, and it admits a degree D SoS certificate of resilience.

Compute the degree $D+1$ pseudo-expectation functional $\tilde{\mathbb{E}}_\zeta : \mathbb{R}[\mathbf{w}, \Theta] \to \mathbb{R}$ by solving the following SDP.

$$\underset{\text{degree D pseudoexpectations } \tilde{\mathbb{E}}}{\text{minimize}} \quad \sum_{i=1}^N \tilde{\mathbb{E}}[w_i]^2 \tag{2.6}$$

$$\underset{\text{satisfies the polynomial system}}{\text{such that } \tilde{\mathbb{E}}} \quad (w_i^2 - w_i) = 0,\ i \in [N] \tag{2.7}$$

$$\sum_{i=1}^N w_i \geq \alpha N,\ i \in [N] \tag{2.8}$$

$$\sum_{i=1}^N w_i \Phi(\tilde{Z}_i, \Theta) \leq \varepsilon \cdot \left(\sum_{i=1}^n w_i\right) \tag{2.9}$$

$$q(\Theta) \geq 0 \qquad \forall q \in \mathcal{V} \tag{2.10}$$

Rounding: Sample $i \in [N]$ with probability proportional to $\tilde{\mathbb{E}}_\zeta[w_i]$ and return $\frac{\tilde{\mathbb{E}}_\zeta[w_i \Theta]}{\tilde{\mathbb{E}}_\zeta[w_i]}$

We defer the proof of the following theorem to the appendix.

Theorem 10 *Fix $\alpha \in (0, 1]$ and $\varepsilon, \delta > 0$. Let $Z_1, \ldots, Z_N \in \mathbb{R}^d$ be samples that were $(\alpha^2, \varepsilon, \delta)$-resilient, and there is a SoS certificate of resilience of degree D. There is an algorithm $\mathcal{A}$ running in time $O(Nd^{O(D)})$ such that with probability at least $\Omega(\alpha)$, the algorithm outputs Θ such that $\|\Theta - \Theta^*\|^2 \leq O(\delta/\alpha^4)$*

3. Subspace Recovery

In this section, we will setup the list-decodable subspace recovery problem and use the framework of certifiable resilience to devise an algorithm for it.

Let $\mathcal{D}$ be a probability distribution over $\mathbb{R}^d$. For the sake of exposition, we will assume $\mathcal{D} = N(0, \mathrm{Id}_d)$ but the discussion can be generalized to any SoS-anticoncentrated distribution over $\mathbb{R}^d$. Let P denote the projection to a k-dimensional subspace over $\mathbb{R}^d$. The uncorrupted data consists of points that are close to projection of $\mathcal{D}$ to the subspace P. Formally, the uncorrupted data consists of examples X_i of the form, $X_i = PX_i^* + \gamma_i$ where X_i^* is drawn from $\mathcal{D}$ and γ_i is an additive noise. We will assume that the additive noise is bounded variance in that

$$\sum_{i=1}^N \|\gamma_i\|^2 \leq \varepsilon N$$

The input to the algorithm consists of N samples $\{\tilde{X}_i\}_{i \in [N]} \in \mathbb{R}^d$, an α-fraction of which are equal to X_i. Specifically, there exists some subset $\mathcal{S} \in [N]$ such that, $\tilde{X}_i = X_i$ for all

$i \in [N]$. The goal of the list-decoding algorithm is to return a small list of k-dimensional subspaces $\{\Pi_1, \ldots, \Pi_\ell\}$ such that at least one of them is close to P.

The parameter being estimated here is the k-dimensional projection matrix Π. The set of k-dimensional projections can be specified by the following set of polynomial equalities in $d \times k$ matrix of . indeterminates U

$$UU^T = \Pi$$
$$U^T U = \mathrm{Id}_k$$

A natural estimator for the subspace P, if the data were completely uncorrupted would be

$$\Pi^* = \operatorname*{argmin}_{\Pi} \sum_{i \in [N]} \|X_i - \Pi X_i\|^2$$

corresponding to the cost function $\Phi(X_i, \Pi) = \|X_i - \Pi X_i\|^2$. Thus the polynomial system associated with subspace recovery is given by,

$$\left\{ \begin{array}{lll} P^{(\alpha)}_{\mathsf{sum}}(\mathbf{w}) & \stackrel{\text{def}}{=} \sum_{i \in [N]} w_i - \alpha N & \geq 0 \\ P_{\mathsf{bool}}(w_i) & \stackrel{\text{def}}{=} w_i^2 - w_i & = 0 \quad \forall i \in [N] \\ P^{(\varepsilon)}_{\mathsf{cost}}(\mathbf{w}) & \stackrel{\text{def}}{=} \varepsilon N - \sum_{i \in [N]} w_i \|X_i - \Pi X_i\|^2 & \geq 0 \\ \Pi = UU^T & & \\ U^T U = \mathrm{Id}_k & & \end{array} \right\} \tag{3.1}$$

Through the framework of the previous section, devising an algorithm for subspace recovery (proving Theorem 1) reduces to showing that the data set is certifiably resilient. We will sketch the proof of certifiable resilience in this section, and defer the proof of Theorem 1 to the Appendix.

Theorem 11 *(SoS certifiable resilience for subspace recovery) For all $\alpha \in (0,1), \eta \in (0, 1/2)$ and $\varepsilon > 0$, suppose $\mathcal{D}$ is a $(c, D(\eta))$-SoS anticoncentrated distribution over $\mathbb{R}^d$ then with high probability, the data set $\{X_1, \ldots, X_N\}$ can be certified to be $(\alpha, \varepsilon, \delta)$-resilient by a degree $D(\eta) + 4$ SoS certificate for $\delta = \left(\frac{4\varepsilon}{\eta^2 \alpha} + \frac{4ck\eta}{\alpha}\right)$.*

Proof For a sample $X_i = PX_i^* + \gamma_i$, we have $X_i - \Pi X_i = (PX_i^* - \Pi P X_i^*) + (\gamma_i - \Pi \gamma_i)$. We can rewrite it as,

$$PX_i^* - \Pi P X_i^* = (X_i - \Pi X_i) - (\gamma_i - \Pi \gamma_i)$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$ we get that,

$$\|PX_i^* - \Pi P X_i^*\|^2 = 2\|X_i - \Pi X_i\|^2 + 2\|\gamma_i - \Pi \gamma_i\|^2 \leq 2\|X_i - \Pi X_i\|^2 + 2\|\gamma_i\|^2$$

where the last inequality uses the fact that $\Pi^2 = \Pi$. Hence we get that,

$$\sum_{i \in [N]} w_i \|PX_i^* - \Pi P X_i^*\|^2 \leq 2 \sum_{i \in [N]} w_i \|X_i - \Pi X_i\|^2 + w_i \|\gamma_i\|^2 \leq 2\varepsilon N + 2\varepsilon N. \tag{3.2}$$

where the second inequality uses $P^{(\varepsilon)}_{\mathsf{cost}}(\mathbf{w}) \geq 0$, $w_i \leq 1$ for all i and $\sum_i \|\gamma_i\|^2 \leq \varepsilon N$. Fix an orthonormal basis $e_1, \ldots, e_d$ such that $\mathrm{Span}\{e_1, \ldots, e_k\} = P$.

$$\|(P - \Pi P)X_i^*\|^2 = \sum_{\ell=1}^{d} \langle (P - P\Pi)e_\ell, X_i^* \rangle^2 \geq \sum_{j=1}^{k} \langle (P - P\Pi)e_\ell, X_i^* \rangle^2$$

Using (3.2) we get that,

$$\sum_{i \in [N]} w_i \sum_{\ell=1}^{k} \langle (P - P\Pi)e_\ell, X_i^* \rangle^2 \leq 4\varepsilon N \tag{3.3}$$

Suppose $v_j := (P - P\Pi)e_j$, then its norm $\|v_j\|^2 = \|(P - P\Pi)e_j\|^2 \leq 2\|Pe_j\|^2 + 2\|P\Pi e_j\|^2 \leq 4$. Since X_i^* is drawn from a $(c, D(\eta))$-anticoncentrated distribution, there is a degree $D(\eta)$ SoS derivation for

$$\sum_{j \in [N]} w_i \langle v_j, X_i^* \rangle^2 \geq \eta^2 \left(\sum_{i \in [N]} w_i \right) \|v_j\|^2 - 4c\eta^3 N$$

Summing the above inequality over all $j = 1 \ldots k$ we get that,

$$\sum_{\ell=1}^{k} \sum_{i \in [N]} w_i \langle (P - P\Pi)e_\ell, X_i^* \rangle^2 \geq \eta^2 \left(\sum_{i \in [N]} w_i \right) \left(\sum_{\ell} \|(P - P\Pi)e_\ell\|^2 \right) - 4ck\eta^3 N$$

In conjunction with (3.3), this implies that,

$$\left(\sum_{i \in [N]} w_i \right) \left(\sum_{\ell} \|(P - P\Pi)e_\ell\|^2 \right) \leq \frac{1}{\eta^2} \sum_{\ell=1}^{k} \sum_{i \in [N]} w_i \langle (P - P\Pi)e_\ell, X_i^* \rangle^2 + 4ck\eta N \leq \left(\frac{4\varepsilon}{\eta^2} + 4ck\eta \right) N$$

Note that $\sum_{\ell=1}^{k} \|(P - P\Pi)e_\ell\|^2 = \|P - P\Pi P\|_F^2$. Using the constraint $P^{(\alpha)}_{\mathsf{sum}}(\mathbf{w})$, we can rewrite the above equation as,

$$\left(\sum_{i \in [N]} w_i \right) \|P - P\Pi P\|_F^2 \leq \left(\frac{4\varepsilon}{\eta^2 \alpha} + \frac{4ck\eta}{\alpha} \right) \left(\sum_{i \in [N]} w_i \right)$$

By Lemma 12 (see Appendix for proof) this implies that,

$$\left(\sum_{i \in [N]} w_i \right) \|P - \Pi\|_F^2 \leq \left(\frac{4\varepsilon}{\eta^2 \alpha} + \frac{4ck\eta}{\alpha} \right) \left(\sum_{i \in [N]} w_i \right)$$

■

Lemma 12 *For a k-dimensional projection matrix $P \in \mathbb{R}^{d\times d}$ and a $d \times k$ matrix of indeterminates U and $\gamma > 0$,*

$$\left\{ \begin{array}{l} (\sum_i w_i^2) \cdot \|P - P\Pi P\|_F^2 \leq k\gamma(\sum_i w_i^2) \\ \Pi = UU^T \\ U^TU = \mathrm{Id}_k \end{array} \right\} \vdash_4 \left(\sum_i w_i^2\right) \cdot \|P - \Pi\|_F^2 \leq k\gamma \left(\sum_i w_i^2\right)$$

The algorithm produced by the framework in previous section will output a matrix $\Pi_i = \frac{\tilde{\mathbb{E}}[w_i\Pi]}{\tilde{\mathbb{E}}[w_i]}$. The output matrix Π_i satisfies $0 \preceq \Pi_i \preceq I$ and $\mathrm{Tr}[\Pi_i] = k$, but is not necessarily a projection matrix. In order to recover a projection matrix, we will have to round the matrix further into one. The following lemma shows that just picking the projection onto the top k eigenvalues of Π_i yields a projection matrix with only a constant factor loss in the error.

Lemma 13 *Let Π be a matrix satisfying $\mathrm{Tr}(\Pi) = k$ and $\Pi \preceq I$. Let P be a rank k projection. If $\langle \Pi, P\rangle \geq k(1-\varepsilon)$, then $\langle \Pi_k, P\rangle \geq k(1-2\varepsilon)$ where Π_k is the top k eigenspace of Π. Equivalently $\|P - \Pi_k\|_F^2 \leq 4\varepsilon$*

Proof Let $\lambda_1, ..., \lambda_d$ and $v_1, ..., v_d$ be the eigenvalues and eigenvectors of Π. Then $\langle \Pi, P\rangle \geq k(1-\varepsilon)$ imlies $\sum_{j=1}^k \lambda_j \geq k(1-\varepsilon)$ and therefore $\sum_{j=k+1}^d \lambda_j \leq \varepsilon k$. Thus we have the following series of inequalities.

$$k(1-\varepsilon) \leq \langle \Pi, P\rangle = \sum_{j=1}^d \lambda_j \langle P, v_jv_j^T\rangle = \sum_{j=1}^k \lambda_j \langle P, v_jv_j^T\rangle + \sum_{j=k+1}^d \lambda_j \langle P, v_jv_j^T\rangle$$

$$\leq \sum_{j=1}^k \lambda_j \langle P, v_jv_j^T\rangle + \sum_{j=k+1}^d \lambda_j \leq \sum_{j=1}^k \lambda_j \langle P, v_jv_j^T\rangle + \varepsilon k \leq \sum_{j=1}^k \langle P, v_jv_j^T\rangle + \varepsilon k = \langle \Pi_k, P\rangle + \varepsilon k.$$

Rearranging, we obtain, $\langle \Pi_k, P\rangle \geq k(1-2\varepsilon)$.

■

References

Pankaj K. Agarwal and Nabil H. Mustafa. K-means projective clustering. In *Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, page 155–165, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 158113858X. doi: 10.1145/1055558.1055581. URL https://doi.org/10.1145/1055558.1055581.

Ainesh Bakshi and Pravesh Kothari. List-decodable subspace recovery via sum-of-squares, 2020.

Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 671–680. ACM, 2008.

A. Bhaskara, A. Chen, A. Perreault, and A. Vijayaraghavan. Smoothed analysis in unsupervised learning via decoupling. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 582–610, 2019.

P. S. Bradley and O. L. Mangasarian. K-plane clustering. *J. of Global Optimization*, 16(1): 23–32, January 2000. ISSN 0925-5001. doi: 10.1023/A:1008324625522. URL https://doi.org/10.1023/A:1008324625522.

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.

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

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, 2018.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *CoRR*, abs/1203.1005, 2012. URL http://arxiv.org/abs/1203.1005.

M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, Jun 1981. ISSN 1439-6912. doi: 10.1007/BF02579273. URL https://doi.org/10.1007/BF02579273.

Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Conference on Learning Theory*, pages 354–375, 2013.

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

Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 379–390. IEEE, 2017.

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

Sushrut Karmalkar, Adam R. Klivans, and Pravesh K. Kothari. List-decodable linear regression, 2019.

Pravesh K Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. *arXiv preprint arXiv:1711.07465*, 2017.

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

Pravesh K Kothari, Raghu Meka, and Prasad Raghavendra. Approximating rectangles by juntas and weakly-exponential lower bounds for lp relaxations of csps. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 590–603. ACM, 2017.

Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM J. Optim.*, 11(3):796–817, 2000/01. ISSN 1052-6234. doi: 10.1137/S1052623400366802.

G. Lerman and T. Maunu. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8):1380–1410, 2018.

Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1546–1562, Sep. 2007. ISSN 1939-3539. doi: 10.1109/TPAMI.2007.1085.

Tyler Maunu and Gilad Lerman. Robust subspace recovery with adversarial outliers. *CoRR*, abs/1904.03275, 2019. URL http://arxiv.org/abs/1904.03275.

Tyler Maunu, Teng Zhang, and Gilad Lerman. A well-tempered landscape for non-convex robust subspace recovery. *CoRR*, abs/1706.03896, 2017. URL http://arxiv.org/abs/1706.03896.

Yurii Nesterov. *Squared Functional Systems and Optimization Problems*, pages 405–440. Springer US, Boston, MA, 2000. ISBN 978-1-4757-3216-0. doi: 10.1007/978-1-4757-3216-0_17. URL https://doi.org/10.1007/978-1-4757-3216-0_17.

Necmiye Ozay, Mario Sznaier, Constantino Lagoa, and Octavia Camps. Gpca with denoising: A moments-based convex approach. pages 3209–3216, 06 2010. doi: 10.1109/CVPR.2010.5540075.

Pablo A. Parrilo. Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. Technical report, 2000.

Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In *Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '20, page 161–180, USA, 2020. Society for Industrial and Applied Mathematics.

N.Z. Shor. Quadratic optimization problems. *Soviet Journal of Computer and Systems Sciences*, 25, 11 1987.

Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J. Candès. Robust subspace clustering. *CoRR*, abs/1301.2603, 2013. URL http://arxiv.org/abs/1301.2603.

Jacob Steinhardt, Gregory Valiant, and Moses Charikar. Avoiding imposters and delinquents: Adversarial crowdsourcing and peer prediction. In *Advances in Neural Information Processing Systems*, pages 4439–4447, 2016.

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

Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *Advances in neural information processing systems*, pages 3517–3529, 2017b.

Michael E. Tipping and Chris M. Bishop. Probabilistic principal component analysis. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 61(3):611–622, 1999.

P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105, 10 1999. doi: 10.1023/A:1004678431677.

R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, March 2011. ISSN 1558-0792. doi: 10.1109/MSP.2010.939739.

René Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (GPCA). *CoRR*, abs/1202.4002, 2012. URL http://arxiv.org/abs/1202.4002.

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics, 2019.

Appendix A. Hardness of List Decodable Subspace Recovery over Hypercube

We construct a dataset for which list decodable subspace recovery is impossible. The argument is straightforward and follows from a Gilbert-Varshamov style lower bound. Our inliers will be distributed in a k dimensional subspace according to the uniform distribution over the boolean hypercube. This innocuous setup turns out to be impossible to list decode even when the outliers are arranged in a benign mixture model distributed uniformly over the boolean hypercube in $\frac{1}{\alpha}$ orthogonal subspaces. For simplicity of presentation, we will assume that each corner of the hypercube is populated by the same number of points. A key takeaway of our lower bound is that modeling inliers as a standard normal over a planted subspace, in addition to being a popular statistical choice, is in a sense necessary. Distributions that form distinct clumps of points are subject to pathologies of interpolation, a problem that is mitigated for points that are anticoncentrated.

Lemma 14 *Let $\alpha \in [0, 1/2]$ be a fixed constant. For any $d \geq \frac{k}{\alpha}$, let $r_1, ..., r_N \sim \{0,1\}^{\frac{k}{\alpha}}$ be a set of points such that there are an equal number on each corner of the $\frac{k}{\alpha}$ dimensional hypercube. Then let the dataset $X_1, ..., X_N \in \mathbb{R}^d$ be defined such that $X_i := (r_i, 0^{d-k/\alpha})$ for all $i \in [N]$ where each datapoint X_i is formed by padding the end of its corresponding r_i vector with zeros. Then there exists a list $L = \{P_i\}_{i=1}^q$ of projection matrices P_i onto k dimensional subspaces where each $P_i \in L$ contains at least αN points; any pair of projections $P_i, P_j \in L$, satisfies $\|P_i - P_j\|_F \geq \sqrt{2\varepsilon k}$ for a constant $\varepsilon \in [0, 1/2]$; the length of $|L| = q$ is greater than $(\frac{1}{\alpha 2^{H(\varepsilon)}})^k$ where $H(\varepsilon)$ is the binary entropy function. Thus for $\varepsilon < \frac{1}{2}$ there exists a k such that no list decoding algorithm can succeed with a fixed polynomial list length.*

Proof Let $V := \{e_1, ..., e_{k/\alpha}\}$ be the first k/α basis vectors. There exists $\binom{\frac{k}{\alpha}}{k}$ subspaces comprised of k basis vectors from V. We can encode these subspaces as vectors over $\{0,1\}^{\frac{k}{\alpha}}$ where a 1 indicates the presence of a basis vector, and a zero indicates the absence. Proving the theorem then reduces to proving that the maximum size of binary code C of length $\frac{k}{\alpha}$ of hamming weight exactly k with decoding radius εk is lower bounded by $(\frac{1}{\alpha 2^{H(\varepsilon)}})^k$.

The proof follows from classical arguments. Since C is of maximum size, there does not exist a boolean vector c_x of hamming weight k that is further than εk away from its closest codeword $c \in C$. Otherwise, it would be possible to add c_x into C which is a contradiction. Since there are exactly $\binom{\frac{k}{\alpha}}{k}$ vectors of hamming weight exactly k, and the hamming ball of radius εk has exactly $\sum_{i=0}^{\varepsilon k} \binom{k}{i}$ the size of C is lower bounded by

$$|C| \geq \frac{\binom{\frac{k}{\alpha}}{k}}{\sum_{i=0}^{\varepsilon k} \binom{k}{i}}$$

we lower bound the numerator by $(\frac{1}{\alpha})^k$, and upper bound the denominator by the binomial theorem $\sum_{i=0}^{\varepsilon k} \binom{k}{i} \leq 2^{H(\varepsilon)k}$. Thus we have shown $|C| \geq (\frac{1}{\alpha 2^{H(\varepsilon)}})^k$ as desired. ■

Appendix B. SoS-Certifiable Anticoncentration

Here we recall the notion of SoS-certifiable anti-concentration and the relevant results from Raghavendra and Yau (2020).

Definition 15 *Let $D : [0, 1/2] \to \mathbb{N}$. A probability distribution $\mathcal{D}$ over $\mathbb{R}^d$ is said to $(c, D(\eta))$-SoS-anticoncentrated, If for any $0 < \eta < \frac{1}{2}$ there exists $\tau \leq c\eta$ and there exists a constant $k \in \mathbb{N}$ such that for all $N > d^k$, with probability $1 - d^{-k}$, over samples $x_1, \ldots, x_N \sim \mathcal{D}$ the following polynomial system*

$$\mathcal{P} = \begin{cases} w_i^2 = w_i & i \in [N] \\ \|v\|^2 \leq \rho^2 \end{cases}$$

yields a degree $D(\eta)$ SoS proof of the following inequality

$$\mathcal{P} \vdash_{D(\eta)} \left\{ \frac{1}{N} \sum_{i=1}^{N} w_i \langle X_i, v \rangle^2 \geq \eta^2 (\frac{1}{N} \sum_i w_i) \|v\|^2 - \eta^2 \tau \rho^2 \right\}$$

Theorem 16 *(Sufficient conditions for SoS anti-concentration) If the degree $D(\eta)$ empirical moments of $\mathcal{D}$ converge to the corresponding true moments M_t of $\mathcal{D}$, that is for all $t \leq D(\eta)$*

$$\left\| \frac{1}{N} \sum_{i=1}^{N} X_i^{\otimes \frac{t}{2}} (X_i^{\otimes \frac{t}{2}})^T - M_t \right\|_F \leq d^{-k}$$

And if there exists a uni-variate polynomial $I_\eta(z) \in \mathbb{R}[z]$ of degree at most $D(\eta)$ such that

1. *$I_\eta(z) \geq 1 - \frac{z^2}{\eta^2 \rho^2}$ for all $z \in \mathbb{R}$.*
2. *$\mathcal{P} \vdash_{D(\eta)} \left\{ \|v\|^2 \cdot \mathbb{E}_{x \in \mathcal{D}}[I_\eta(\langle v, x \rangle)] \leq c\eta\rho^2 \right\}$.*

Then $\mathcal{D}$ is $(c, D(\eta))$ certifiably anticoncentrated.

Lemma 17 *For every $d \in \mathbb{N}$, the standard Gaussian distribution $\mathcal{N}(0, I_d)$ is $(c, O(\frac{1}{\eta^4}))$-SoS-anticoncentrated. In particular there exists a construction for $c \leq 2\sqrt{e}$*

Appendix C. Sum-of-Squares Toolkit

Here we present some useful inequalities captured by the sum of squares proof system

Useful Inequalities.

Fact 18 *(Cauchy Schwarz) Let $x_1, .., x_n, y_1, ..., y_n$ be indeterminates, than*

$$\vdash_4 \left(\sum_{i \leq n} x_i y_i \right)^2 \leq \left(\sum_{i \leq n} x_i^2 \right) \left(\sum_{i \leq n} y_i^2 \right)$$

Fact 19 *(Triangle Inequality) Let x, y be n-length vectors of indeterminates, then*

$$\vdash_2 \|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$$

Fact 20 *(Pseudoexpectation Cauchy Schwarz). Let $f(x)$ and $g(x)$ be degree at most $\ell \leq \frac{D}{2}$ polynomial in indeterminate x, then*

$$\tilde{\mathbb{E}}[f(x)g(x)]^2 \leq \tilde{\mathbb{E}}[f(x)^2] \, \tilde{\mathbb{E}}[g(x)^2]$$

Fact 21 *(Spectral Bounds) Let $A \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix with λ_{max} and λ_{min} being the largest and smallest eigenvalues of A respectively. Let $\tilde{\mathbb{E}}$ be a pseudoexpectation with degree greater than or equal to 2 over indeterminates $v = (v_1, ..., v_d)$. Then we have*

$$\vdash_2 \langle A, vv^T\rangle \leq \lambda_{max}\|v\|^2$$

and

$$\vdash_2 \langle A, vv^T\rangle \geq \lambda_{min}\|v\|^2$$

Appendix D. Omitted Proofs

D.1. Proof of Theorem 1

Theorem 22 *Suppose $\{X_i^*\}_{i\in[N]}$ are drawn from $N(0, \mathrm{Id}_d)$ and let P be a projection to a k-dimensional subspace of $\mathbb{R}^d$. Let $\{X_i | i \in [N]\}$ be generated by setting*

$$X_i = PX_i^* + \gamma_i$$

for some additive noise γ_i satisfying

$$\sum_{i=1}^{N}\|\gamma_i\|^2 = \varepsilon N$$

Let $\{\tilde{X}_i | i \in [N]\}$ be a set of points such that there exists a subset $\mathcal{S} \in [N]$ of size $|\mathcal{S}| \geq \alpha N$ with $\tilde{X}_i = X_i$ for all $i \in \mathcal{S}$. For all $\eta > 0$, there is an algorithm running in time $d^{O(1/\eta^4)}$ that computes a list of $O(1/\alpha)$ projection matrices $\{\Pi_1, \ldots, \Pi_\ell\}$ such that

$$\min_j \|P - \Pi_j\|^2 \leq O\left(\frac{\varepsilon}{\eta^2\alpha^5} + \frac{4ck\eta}{\alpha^5}\right)$$

Proof The theorem is a consequence of applying the framework from Section 2.2 to the polynomial system (3.1). Specifically, in Theorem 11 we show that for any $(c, D(\eta))$-SoS anticoncentrated distribution, the data set $\{X_1, \ldots, X_N\}$ is SoS-certifiably $(\alpha, \varepsilon, \delta)$-resilient for $\delta = \left(\frac{4\varepsilon}{\eta^2\alpha} + \frac{4ck\eta}{\alpha}\right)$ and degree $D(\eta) + 4$. Using the algorithm in Theorem 10, we can recover a matrix Π such that $\|P - \Pi\|_F^2 = \delta/\alpha^4$. While Π satisfies $\Pi \preceq \mathrm{Id}$ and $\mathrm{Tr}(\Pi) = k$, it is not necessarily a projection matrix. In particular, Π can have eigenvalues that are not 0 or 1. However, we prove in Lemma 13 the matrix Π can be rounded to a projection matrix with only a constant loss in the squared Frobenius norm $\|P - \Pi\|_F^2$. Thus one can recover a subspace Π such that $\|P - \Pi\|_F^2 \leq O\left(\frac{\varepsilon}{\eta^2\alpha^5} + \frac{4ck\eta}{\alpha^5}\right)$. The running time of the algorithm is $d^{O(D(\eta))}$. By Lemma 17, the Gaussian distribution is $(2\sqrt{e}, O\left(\frac{1}{\eta^4}\right))$-SoS anticoncentrated, thus giving a runtime of $d^{O(1/\eta^4)}$.

■

D.2. Proof of Theorem 10

Proof For sake of succinctness, we will denote

$$\tilde{\mathbb{E}}_i \stackrel{\text{def}}{=} \tilde{\mathbb{E}}[\cdot | w_i = 1] \qquad \text{and } \beta_i \stackrel{\text{def}}{=} \tilde{\mathbb{E}}[w_i]\,.$$

Define the pseudoexpectation operator $\mathbb{E}^*$ by modifying $\tilde{\mathbb{E}}$ to make $w_j = 0$ for all $j \notin \mathcal{S}$. In particular, for any $j \notin \mathcal{S}$, pseudoexpectation of all monomials containing w_j is set to 0. Formally, for every monomial we set, It is easy to check that $\mathbb{E}^*[w_i^2] = \mathbb{E}^*[w_i]$, and that $\mathbb{E}^*$ satisfies $\mathbb{E}^*[q(\Theta)] = 0$ for all $q \in \mathcal{V}$. Finally, the cost of the $\mathbb{E}^*$ on true data $\{Z_1, \ldots, Z_N\}$ is not higher than the cost of $\tilde{\mathbb{E}}$. This is because,

$$\mathbb{E}^*\left[\sum_{j\in[N]} w_j \Phi(Z_j, \Theta)\right] = \sum_{j\in\mathcal{S}} \tilde{\mathbb{E}}[w_j \Phi(Z_j, \Theta)] \le \sum_{j\in[N]} \tilde{\mathbb{E}}[w_j \Phi(\tilde{Z}_j, \Theta)] \le \varepsilon N \tag{D.1}$$

where we used the fact that $\tilde{\mathbb{E}}[w_j \Phi(\tilde{Z}_j, \Theta)] \ge 0$ for all $j \in [N]$ and the $P^{(\varepsilon)}_{\text{cost}}$ constraint on the corrupted data,

Notice that the rounding algorithm outputs $\tilde{\mathbb{E}}_i[\Theta]$ with probability proportional to $\beta_i = \tilde{\mathbb{E}}[w_i]$. By Lemma 9, we have that that,

$$\sum_{i\in\mathcal{S}} \beta_i \ge \tilde{\mathbb{E}}_\zeta[\sum_{i\in\mathcal{S}} w_i] \ge \alpha^2 N$$

while $\sum_{i\in[N]} \beta_i \le N$. Therefore with probability at least α^2, the rounding algorithm picks $i \in \mathcal{S}$. Conditioned on picking an element $i \in \mathcal{S}$, the expected distance $\|\Theta - \Theta^*\|^2$ satisfies,

$$\begin{aligned}
&\left(\mathbb{E}_{i\sim\mathcal{S}}[\|\tilde{\mathbb{E}}_i[\Theta] - \Theta^*\|]\right)^2 \\
&= \left(\mathbb{E}_{i\sim\mathcal{S}}[\|\mathbb{E}^*{}_i[\Theta] - \Theta^*\|]\right)^2 && \text{Definition of } \mathbb{E}^* \\
&= \left(\frac{1}{\sum_{i\in\mathcal{S}}\beta_i} \sum_{i\in\mathcal{S}} \beta_i \|\mathbb{E}^*{}_i[\Theta - \Theta^*]\|\right)^2 && \text{Definition of } \mathbb{E}_{i\in\mathcal{S}} \\
&= \left(\frac{1}{\sum_{i\in\mathcal{S}}\beta_i}\right)^2 \cdot \left(\sum_{i\in\mathcal{S}} \|\mathbb{E}^*[w_i(\Theta - \Theta^*)]\|\right)^2 && \text{using } \tilde{\mathbb{E}}_i[\Theta] = \tfrac{\tilde{\mathbb{E}}[w_i\Theta]}{\tilde{\mathbb{E}}[w_i]} \\
&\le \left(\frac{1}{\sum_{i\in\mathcal{S}}\beta_i}\right)^2 \cdot |\mathcal{S}| \cdot \sum_{i\in\mathcal{S}} \|\mathbb{E}^*[w_i(\Theta - \Theta^*)]\|^2 && \text{using Cauchy-Schwartz} \\
&\le \left(\frac{1}{\sum_{i\in\mathcal{S}}\beta_i}\right)^2 \cdot |\mathcal{S}| \cdot \sum_{i\in\mathcal{S}} \mathbb{E}^*\left[\|w_i(\Theta - \Theta^*)\|^2\right] && \text{using pseudo-expectation Cauchy-Schwartz} \\
&= \left(\frac{1}{\sum_{i\in\mathcal{S}}\beta_i}\right)^2 \cdot |\mathcal{S}| \cdot \mathbb{E}^*\left[\left(\sum_{i\in\mathcal{S}} w_i\right) \|\Theta - \Theta^*\|^2\right] && \text{using } w_i^2 - w_i = 0 \\
&= \left(\frac{1}{\sum_{i\in\mathcal{S}}\beta_i}\right)^2 \cdot |\mathcal{S}| \cdot \mathbb{E}^*\left[\left(\sum_{i\in[N]} w_i\right) \|\Theta - \Theta^*\|^2\right] && \text{using } \mathbb{E}^*[w_i] = 0 \text{ for } i \notin \mathcal{S} \\
&\le \tfrac{\delta}{\alpha^4}
\end{aligned}$$

Finally, note that $\mathbb{E}^*$ is a valid degree D pseudo-expectation functional that satisfies the constraints $P_{\text{bool}}(w_i)$ and $P^{(\varepsilon)}_{\text{cost}}$ ((D.1)) on the original data $\{Z_1, \ldots, Z_N\}$. Since the uncorrupted data $\{Z_1, \ldots, Z_N\}$ is $(\alpha, \varepsilon, \delta)$-resilient and admits a degree D SoS certificate, we can conclude that

$$\mathbb{E}^*\left[\left(\sum_{i\in[N]} w_i\right) \|\Theta - \Theta^*\|^2\right] \le \delta N$$

Along with the fact that $\sum_i \beta_i \geq \alpha^2 N$, the above calculation implies that

$$\left(\mathop{\mathbb{E}}_{i\sim\mathcal{S}}[\|\tilde{\mathbb{E}}_i[\Theta] - \Theta^*\|]\right)^2 \leq \frac{\delta}{\alpha^4} \tag{D.2}$$

■

D.3. Proof of Lemma 12

Proof For notational convenience, denote $A \stackrel{\text{def}}{=} \sum_i w_i^2$. Note that

$$2P\Pi P = P^2 + (P\Pi P)^2 - (P - P\Pi P)^2 .$$

Thus

$$\begin{aligned} 2A \cdot \operatorname{Tr}[P\Pi P] &= A\operatorname{Tr}[P^2] + A\operatorname{Tr}[(P\Pi P)^2] - A\operatorname{Tr}[(P - P\Pi P)^2] \\ &= Ak + Ak - A\operatorname{Tr}[(P - P\Pi P)^2] \qquad \text{(Lemma 23)} \\ &= Ak + Ak - Ak\gamma \qquad \text{(Lemma 23)} \\ &= k(2-\gamma)\cdot A \end{aligned}$$

Finally, we have

$$\begin{aligned} A \cdot \|P - \Pi\|_F^2 &= A(\operatorname{Tr}[P^2] + \operatorname{Tr}[\Pi^2] - 2\operatorname{Tr}(P\Pi)) \\ &= A\left(2k - 2\operatorname{Tr}(P\Pi P)\right) \leq k\gamma \cdot A \end{aligned}$$

■

Lemma 23 *Suppose U is a $d \times k$ matrix of indeterminates satisfying $U^TU = \mathrm{Id}_k$. Suppose $\Pi = UU^T$ and $P \in \mathbb{R}^{d\times d}$ is a projection matrix then,*

$$U^TU = \mathrm{Id}_k \;\big|\!\!\frac{}{2}\; \operatorname{Tr}[(P\Pi P)^2] \leq k \tag{D.3}$$

Proof
Observe that for any positive semidefinite matrices $A, B \succeq 0$ and indeterminates U

$$\begin{aligned} \operatorname{Tr}[U^TAUU^TBU] &= \operatorname{Tr}[U^TA^{1/2}A^{1/2}UU^TB^{1/2}B^{1/2}U] \\ &= \operatorname{Tr}[(A^{1/2}UU^TB^{1/2})(A^{1/2}UU^TB^{1/2})^T] \\ &= \|A^{1/2}UU^TB^{1/2}\|_F^2 \geq 0 \end{aligned}$$

Now we can write,

$$\begin{aligned} \operatorname{Tr}[(P\Pi P)^2] &= \operatorname{Tr}[PUU^TPUU^TP] \\ &= \operatorname{Tr}[U^TPUU^TPU] \\ &\leq \operatorname{Tr}[U^TPUU^TPU] + \operatorname{Tr}[U^TPUU^T(\mathrm{Id} - P)U] + \operatorname{Tr}[U^T(\mathrm{Id} - P)UU^TU] \\ &= \operatorname{Tr}[(U^TU)^2] \\ &= \operatorname{Tr}[\mathrm{Id}_k] = k \qquad \text{using } U^TU = \mathrm{Id}_k \end{aligned}$$

■