Sparse PCA: Algorithms, Adversarial Perturbations and Certificates

Tommaso d'Orsi[*]
Comp. Science Department
ETH Zürich
Zürich, Switzerland
tommaso.dorsi@inf.ethz.ch

Pravesh K. Kothari[†]
Comp. Science Department
Carnegie Mellon University
Pittsburgh, US
praveshk@cs.cmu.edu

Gleb Novikov
Comp. Science Department
ETH Zürich
Zürich, Switzerland
gleb.novikov@inf.ethz.ch

David Steurer[§]
Comp. Science Department
ETH Zürich
Zürich, Switzerland
david.steurer@inf.ethz.ch

***Abstract*—We study efficient algorithms for Sparse PCA in standard statistical models (spiked covariance in its Wishart form). Our goal is to achieve optimal recovery guarantees while being resilient to small perturbations. Despite a long history of prior works, including explicit studies of perturbation resilience, the best known algorithmic guarantees for Sparse PCA are fragile and break down under small adversarial perturbations.**

We observe a basic connection between perturbation resilience and *certifying algorithms* that are based on certificates of upper bounds on sparse eigenvalues of random matrices. In contrast to other techniques, such certifying algorithms, including the brute-force maximum likelihood estimator, are automatically robust against small adversarial perturbation.

We use this connection to obtain the first polynomial-time algorithms for this problem that are resilient against additive adversarial perturbations by obtaining new efficient certificates for upper bounds on sparse eigenvalues of random matrices. Our algorithms are based either on basic semidefinite programming or on its low-degree sum-of-squares strengthening depending on the parameter regimes. Their guarantees either match or approach the best known guarantees of *fragile* algorithms in terms of sparsity of the unknown vector, number of samples and the ambient dimension.

To complement our algorithmic results, we prove rigorous lower bounds matching the gap between fragile and robust polynomial-time algorithms in a natural computational model based on low-degree polynomials (closely related to the pseudo-calibration technique for sum-of-squares lower bounds) that is known to capture the best known guarantees for related statistical estimation problems. The combination of these results provides formal evidence of an inherent price to pay to achieve robustness.

Beyond these issues of perturbation resilience, our analysis also leads to new algorithms for the fragile setting, whose guarantees improve over best previous results in some parameter regimes (e.g. if the sample size is polynomially smaller than the dimension).

***Keywords*-Sum-of-Squares; Sparse PCA; Low-degree Polynomials; Adversarial Perturbations; Eigenvalues of Random Matrices; Robustness.**

I. INTRODUCTION

Sparse principal component analysis (sparse PCA) is a fundamental primitive in high-dimensional statistics. Given a collection of vectors $y_1, \ldots, y_n \in \mathbb{R}^d$, we seek a "structured" direction $v_0 \in \mathbb{R}^d$ with $\|v_0\| = 1$ maximally correlated with the vectors, commonly measured by the empirical variance of $\{\langle y_i, v_0\rangle\}_{i\in[n]}$. The structure we impose on v_0 is sparsity, that is, an upper bound on the number of its non-zero entries.

Spiked covariance model: A widely studied statistical model for sparse PCA is the *spiked covariance model* (also called *Wishart model*). Here, $y_1, \ldots, y_n$ are independent draws from the distribution $N(0, \mathrm{Id}_d + \beta \cdot v_0 v_0^T)$ for an unknown k-sparse unit vector $v_0 \in \mathbb{R}^d$. (For simplicity, we will assume that the sparsity parameter k is known.) The goal is to compute an estimate $\hat{v}$ for v_0 with correlation[1] bounded away from 0 so that $\|\hat{v}\| = 1$ and $\langle \hat{v}, v_0\rangle^2 \geqslant c$ for an absolute constant $c > 0$. (Here, we square the inner product because v_0 is identifiable only up to sign.)

In order to simplify our discussion, we hide multiplicative factors logarithmic in d using the notation $\tilde{O}(\cdot)$. Similarly, we hide absolute constant multiplicative factors using the standard notations $\lesssim$, $O(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$.

If we ignore computational efficiency, we can achieve optimal statistical guarantees for sparse PCA in the spiked covariance model by the following kind of exhaustive search: among all k-by-k principal submatrices of the empirical covariance matrix of the vectors $\{y_i\}_{i\in[n]}$, find one with maximum eigenvalue and output a corresponding eigenvector (e.g., [1]–[3]). This procedure achieves constant correlation with high probability as long as $n \geqslant \tilde{O}(k/\min\{\beta, \beta^2\})$. However, the running-time is exponential in k. When the number of samples n is significantly smaller than the ambient dimension d as well as the sparsity parameter k, an alternative approach is to find a unit vector u such that $u^\top Y$ is close to a k-sparse vector. This procedure also works for $n \geqslant \tilde{O}(k/\min\{\beta, \beta^2\})$ and the running time is exponential in n.

[*] Supported by Steurer's ERC Consolidator Grant.

[†] Part of this work done while at Princeton University and the Institute for Advanced Study.

[§] Supported by an ERC Consolidator Grant.

[1] Instead of asking for the correlation to be bounded away from 0, we could also ask for it to approach 1. Alternatively, we could ask to recover the support of v_0. At the granularity of our discussion here, these measures of success are equivalent in most regards.

2575-8454/20/$31.00 ©2020 IEEE
DOI 10.1109/FOCS46700.2020.00058

The spiked covariance model exhibit a sharp transition in the top eigenvalue for $n \gtrsim \frac{d}{\beta^2}$ (called *BPP transition* [4] in reference to the authors' names). In this regime, called *strong-signal regime*, the following spectral algorithm matches the optimal statistical guarantees of exhaustive search: compute the top right singular vector of Y and restrict it to the k largest entries [5]. We refer to this algorithm as *SVD with thresholding*.

Whenever $n \lesssim \frac{d}{\beta^2}$, principal component analysis of $\{y_i\}_{i\in[n]}$ cannot be used to recover v_0. One of the best known polynomial-time algorithms for this regime (called *weak-signal-regime*) is *diagonal thresholding* [6]: restrict the empirical covariance matrix to the principal submatrix that contains the k largest diagonal entries and output the top eigenvector of this submatrix. This algorithm succeeds with high probability whenever $n \gtrsim \frac{k^2}{\beta^2}\log\frac{d}{k}$ — almost quadratically worse than exhaustive search.[2] Similar guarantees were shown to be achievable in polynomial time through a semidefinite relaxation [1], [8] (which we refer to as the *basic SDP*) . A large and diverse body of work [1], [2], [5], [7], [9], [10] has been dedicated to the question of understanding if this quadratic gap between the sample sizes required for computationally efficient and inefficient methods is inherent or if better polynomial-time algorithms exist for this problem. Hardness results addressing this question take two forms: either reductions from conjecturally hard problems, such as *planted clique* [2] or concrete lower bounds against restricted classes of algorithm such as the sum-of-squares [11], [12] or low-degree polynomials [7].

While these results provide evidence that a quadratic gap between polynomial-time algorithms and exhaustive search is inherent in the weak signal regime, it turns out that a logarithmic improvement over diagonal thresholding is possible (for a broad parameter range): in the regime $k \leqslant \sqrt{d}/2$, a more sophisticated algorithm called *covariance thresholding* [5], [13] succeeds for $n \gtrsim \max\left\{\frac{k^2}{\beta^2}\log\frac{d}{k^2}, k^2\right\}$. This turns into an asymptotic improvement over diagonal thresholding in the settings $d^{1-o(1)} \leqslant k^2 \leqslant o(d)$, but requires the constraint $n \geqslant k^2$. For example, if $\beta = 1$ and $k^2 = \varepsilon d$ for some small enough $\varepsilon > 0$, covariance thresholding works with $n \gtrsim k^2\log(1/\varepsilon)$, while SVD requires $n \gtrsim k^2/\varepsilon$ and diagonal thresholding requires $n \gtrsim k^2 \log d$.

Adversarial entry-wise perturbations: In a seminal work, Huber [14] asked how the guarantees of estimators—designed to work under the assumption of observing Gaussian noise—would change if the data were roughly normal, but not exactly so, thus broadening the circumstances under which the performance of an estimator should be judged. This is especially relevant if we consider that in many real world problems, data may be preprocessed, or the precision of an individual input may be limited. For example, digital images may use few bits to encode a pixel and discard all residual information. For these reasons, it is not desirable for an estimator to drastically change its response as the input changes between Y and $Y + E$ for a small perturbation matrix E. In this sense, the robustness of an estimator is an important aspect for understanding its performances in real-world environments [15], [16].

It turns out that the algorithmic landscape for sparse PCA changes drastically in the presence of adversarial perturbations, where an adversary may change each entry of the input vectors $y_1, \ldots, y_n$ by a small amount. On the one hand, exhaustive search and the basic SDP continue to give the same guarantees as in the vanilla single-spike model. On the other hand, all aforementioned thresholding algorithms are highly sensitive to small adversarial perturbations.

Concretely, in the strong signal regime $\beta \lesssim d/n$, it is possible to adversarially perturb the vectors $y_1, \ldots, y_n$ by at most $\tilde{O}(1/\sqrt{n})$ per entry such that SVD with thresholding achieves only vanishing correlation. Indeed an adversarial perturbation with this effect can be viewed as a whitening transformation and corresponds to a natural generative process for $y_1, \ldots, y_n$, where the vectors are chosen randomly from an n-dimensional subspace containing an approximately sparse vector . We also show that adversarial perturbations of this magnitude can fool diagonal thresholding and covariance thresholding.

Sparse eigenvalues certificates: It is remarkable to notice the stark contrast that appears when instead adversarial perturbations are used against the basic SDP[3], indeed it is easy to show that the algorithm succeeds whenever adversarial perturbations are bounded (in absolute value) by $\sqrt{\frac{\beta}{k}} \cdot \min\{\sqrt{\beta}, 1\}$. If, for example, we assume $\beta \geqslant 1$ and consider the regime in which diagonal thresholding works, that is $\beta \geqslant \tilde{O}(k/\sqrt{n})$, this bound means the algorithm can afford perturbations bounded by $O(1/n^{1/4})$. This is even more remarkable when one notices that for perturbations larger than $\tilde{O}(1/n^{1/4})$ an adversary could plant a matrix with k-sparse norm greater than βn, thus fooling even the exhaustive search algorithm (moreover, this adversary can completely remove the signal from Y).

Considering these observations, it is only natural to ask what is the reason that makes some algorithms robust[4] to corruptions while others turn out to be highly susceptible to small perturbations in the samples. This lead us to the central questions of this paper:

[2] We remark that [7] provides an algorithm that interpolates between Diagonal Thresholding and brute force search. Concretely, given any natural number $t \leqslant n/\log d$, the algorithm recover the sparse vector in time $d^{O(t)}$ if $\beta \gtrsim \frac{k}{\sqrt{tn}}\sqrt{\log d}$. Whenever our discussion will revolve around polynomial time algorithms, we will simply talk about Diagonal Thresholding.

[3] We remark that a certain informal notion of robustness to entry-wise perturbations of the basic SDP program was already argued in [8]. Additionally, in [2] the authors observed that the algorithm is robust to small perturbations of the empirical covariance matrix. We allow here more general perturbations.

[4] In this paper we will interchangeably use the terms robust and resilient.

> *Is there some inherent property that makes an algorithm resilient to adversarial perturbations?*

In the context of Sparse PCA, we answer this question showing how algorithms that come with *certificates* of sparse quadratic forms[5] are intrinsically better in the sense that small perturbations – which by virtue of being small cannot significantly change the sparse eigenvalues of the instance – cannot be used to fool them. In contrast, fragile algorithms – which do not produce such certificates – may be fooled by adversarial perturbations into outputting an estimation uncorrelated with the sparse vector v_0.

We remark that the insight obtained in this analysis also led us to new improvements in the single spiked covariance model.

Certification and the cost of resilience: The robustness of semidefinite programs had already been noted in the literature. For the stochastic block model, efficient spectral algorithms (see [17]) are known to recover the partitions up to the (conjectured) computational threshold.[6] However, few adversarial edge deletions and additions can fool such estimators. On the other hand, algorithms based on semidefinite programming were shown to be resilient to adversarial perturbations [18]–[23], albeit far from the Kesten-Stigum thresold in general settings.[7] The underlying question of this line of work is whether the additional property of resilience comes ”for free”.

In the context of this paper, with the idea of certification mechanisms being a sufficient algorithmic property for adversarial resilience, it becomes relevant to look into the limitations of certification algorithms as well. For the *Sherrington-Kirkpatric* problem [24] – the problem of maximizing the quadratic form $x^{\mathsf{T}}Wx$ where $x \in \left\{\pm 1/\sqrt{n}\right\}^n$ and W is a symmetric random matrix with iid Gaussian entries above the diagonal – [25] showed (modulo a reasonable conjecture) that for any $\varepsilon > 0$ there exists a polynomial-time optimization algorithm returning a value ε-close to the optimum. Conversely, [26] proved that no low-degree polynomial can obtain an ε-close certificate for the problem. Thus suggesting that certification may be a inherently harder task than optimization.

For sparse PCA in the strong signal regime, we observe a strikingly steep *statistical price to pay for robustness*, in the form of a lower bound on the guarantees of low-degree polynomials. That is a *fundamental separation between the power of fragile and resilient algorithms*.

A. Results

So far, we have generically said that an algorithm is ”robust” if it recovers the planted signal even in the presence of malicious noise. However, several issues arise if one tries to make this vague definition more concrete. At first, one could say that robust algorithms achieve comparable guarantees both in the presence and the absence of adversarial corruptions. Yet, in general, this interpretation makes little sense. Malicious perturbations may remove part of the signal, making the guarantees of the fragile settings statistically impossible to achieve or –as we will see for the sparse PCA in certain regimes– they might make the goal of achieving such guarantees *computationally much harder*, thus at the very least forcing us to spend a significantly higher amount of time to obtain the same aforementioned guarantees.

For this reason, in many settings it will make sense to say that an algorithm is resilient if it recovers the sparse signal in the presence of adversarially chosen perturbations *even though* its guarantees may not fair well when compared to those achievable in the fragile settings.

The second fundamental aspect concerns the desirable degree of robustness that an algorithm should possess. Indeed, any reasonable algorithm can likely tolerate sufficiently small adversarial perturbations. Therefore, it is important to quantify the magnitude of the perturbations we ask algorithms to tolerate. Here, we also expect this magnitude to decrease monotonically with the signal strength β. A natural concrete way to formalize this idea is the following: *the algorithm should be expected to obtain correlation bounded away from zero, as long as v_0 remains the principal sparse component*. That is, as long as the vector maximizing the k-sparse norm of Y is correlated with v_0, then the algorithm should be able to output an estimator correlated with v_0.

Concretely, these observations lead us to the following problem formulation.

Problem I.1 (Robust sparse PCA)**.** Given a matrix of the form

$$Y = W + \sqrt{\beta}u_0{v_0}^{\mathsf{T}} + E, \quad \text{where} \tag{I.1}$$

- $v_0 \in \mathbb{R}^d$ is a unit k-sparse vector,
- $u_0 \sim N(0, \mathrm{Id}_n)$ is a standard Gaussian vector,
- $W \sim N(0,1)^{n\times d}$ is a Gaussian matrix and W, u_0, v_0 are distributionally independent,
- $E \in \mathbb{R}^{n\times d}$ is an arbitrary perturbation matrix satisfying[8]

$$\|E\|_\infty \lesssim \sqrt{\beta/k} \cdot \min\{\sqrt{\beta}, 1\}\,. \tag{I.2}$$

Return a unit vector $\hat{v}$ having non-vanishing correlation with v_0.

[5] For a matrix $M \in \mathbb{R}^{d\times d}$ we study the values of the quadratic form $\|Mv\|^2$ at k-sparse vectors v. We define the k-sparse norm of M as $\max_{\|v\|=1, v\ k\text{-sparse}} \|Mv\|$. We sometimes refer to the k-sparse unit vector v that maximizes $\|Mv\|$ as a sparse eigenvector, and to the corresponding value as a sparse eigenvalue.

[6] Called the Kesten-Stigum threshold.

[7] Another qualitative difference between the semidefinite programs studied in the paper above and other families of algorithms is the resilience to monotonic perturbations (see [18], [20]).

[8] In non-robust settings, we simply enforce the constraint $\|E\|_\infty = 0$.

To get an intuition why bound Eq. (I.2) is canonical, observe that for $\beta \geqslant \Omega(1)$ adversarial perturbations of magnitude $\tilde{O}(\sqrt{\beta/k})$ could remove all information about v_0 (see Section II-A). With this formalization of the problem we can now unambiguously define robust algorithms. Specifically, we say that an algorithm is $(n, d, k, \beta, \delta, p)$*–perturbation resilient* if, for parameters (n, d, k, β), with probability at least p it outputs a unit vector $\hat{v}$ such that $1 - \langle \hat{v}, v_0 \rangle^2 \leqslant \delta$.

Resilient algorithms in the strong signal regime: With the above discussion in mind, one may ask whether the same guarantees known for the single spike covariance model may also be achieved in the presence of adversarial perturbations. In the strong signal regime $\beta \gtrsim \sqrt{d/n}$, this amounts to finding a robust and efficient algorithm that achieves the same guarantees as SVD with thresholding. As we will see however, this is most likely impossible. That is, we will provide compelling evidence that *resilient algorithms cannot match the guarantees of fragile algorithms in the strong signal regime.*

Since for $\sqrt{d/n} \lesssim \beta \lesssim d/n$ adversarial perturbations of the order $\tilde{O}(1/\sqrt{n})$ can change the top eigenvalue of the covariance matrix, PCA arguments cannot be used to obtain resilient algorithms. Thus intuitively, this suggests that different kinds of certificates are needed.

We provide a Sum-of-Squares algorithm that recovers in time $d^{O(t)}$ the sparse vector whenever $n \gtrsim \frac{k}{\beta} \cdot t \left(\frac{d}{k}\right)^{1/t}$ and $d^{1/t} \geqslant \tilde{\Omega}(n)$. The key contribution is indeed an efficient algorithm to certify upper bounds on random quadratic forms. For subgaussian[9] low-rank quadratic forms, these upper bounds approach information-theoretically optimal bounds.

Concretely, for an n-by-d matrix W with i.i.d. Gaussian entries, with high probability the degree-t sum-of-squares algorithm (with running time $d^{O(t)}$) certifies an upper bound of $O(k \cdot (k/d)^{-1/t} \cdot t)$ on the quadratic form $Q(x) = \|Wx\|^2$ over all k-sparse unit vectors x if $d^{1/t} \geqslant \tilde{\Omega}(n)$. With these certificates, a robust algorithm for Sparse PCA follows then as a specific corollary.

It is important to notice how this result for sparse PCA is interesting regardless of its resilience properties. As t approaches $\log(d/k)$, the algorithm approaches the information theoretic optimal bound $O(\frac{k}{\beta} \cdot \log(d/k))$. For example, consider the case $n = 2^{\Theta\left(\sqrt{\log d}\right)}$. If also $\frac{d}{k} = 2^{\Theta\left(\sqrt{\log d}\right)}$, the Sum of Squares algorithm works in time $d^{O\left(\sqrt{\log d}\right)} = n^{O\left(\log^2 n\right)}$ with information theoretically optimal guarantees, while exhaustive search takes time exponential in n.

The specific algorithmic result is shown in the following theorem.

Theorem I.2 (Perturbation Resilient Algorithm in the Strong Signal Regime)**.** *Given an n-by-d matrix Y of the form,*

$$Y = \sqrt{\beta} \cdot u_0 {v_0}^{\mathsf{T}} + W + E\,,$$

for $\beta > 0$, a unit k-sparse vector $v_0 \in \mathbb{R}^d$, a Gaussian matrix $W \sim N(0,1)^{n \times d}$, a vector $u_0 \in \mathbb{R}^n$ independent of W with $\|u_0\|^2 = \Theta(n)$, and a matrix $E \in \mathbb{R}^{n \times d}$ satisfying $\|E\|_\infty \lesssim \sqrt{\beta/k} \cdot \min\{\sqrt{\beta}, 1\}$.

For $t \in \mathbb{N}$ suppose that $d \gtrsim n^t \log^t(n) t^t$ and

$$\beta \gtrsim \frac{k}{n} \cdot t \cdot \left(\tfrac{d}{k}\right)^{1/t}\,.$$

Then, there exists an algorithm that computes in time $d^{O(t)}$ a unit vector $\hat{v} \in \mathbb{R}^d$ such that

$$1 - \langle \hat{v}, v_0 \rangle^2 \leqslant 0.01$$

with probability at least 0.99.

In any case, the fundamental limitation of the above algorithm is the requirement $d^{1/t} \geqslant \tilde{\Omega}(n)$. This constraint makes it impossible to match the guarantees of SVD+ thresholding in most regimes, but a priori it remains unclear why better robust algorithms could not be designed. To provide formal evidence that without the requirement $d^{1/t} \geqslant \tilde{\Omega}(n)$ achieving the kind of guarantees of Theorem I.2 may be computationally intractable, we make use of a remarkably simple method (sometimes called analysis of the *low degree likelihood ratio*), developed in a recent line of work on the the sum of squares hierarchy [12], [27]–[29]. That is, we show that in the restricted computational model of low-degree polynomials, there is no efficient algorithm that can improve over the Sum-of-Squares algorithm. This hardness results suggests a *fundamental separation between fragile and resilient algorithms, in other words, an inherent cost to pay in exchange for perturbation-resilience.*

Concretely, we construct $n \times d$ matrices of the form $Y = \sqrt{\beta} u_0 {v_0}^{\mathsf{T}} + W + E$ where E is a perturbation matrix with entries bounded by $\tilde{O}\left(1/\sqrt{n}\right)$ such that, whenever d is significantly smaller than n^t, multilinear polynomials of degree at most $n^{0.001}$ cannot distinguish these Y's from $n \times d$ Gaussian matrices (in the sense that w.h.p. every such polynomial takes roughly the same values under both distributions). These ideas are formalized in the theorem below.

Theorem I.3 (Lower Bound for Resilient Algorithms in the Strong Signal Regime, Informal)**.** *Let t be a constant and let $d \leqslant n^{0.99t-1}$. Suppose that*

$$\beta \leqslant O\left(\frac{k}{n} \cdot t \cdot (d/k)^{1/t}\right).$$

[9] Formally we require a stronger property, we need matrices to be *certifiably subgaussian.*

and[10] $\beta n/k \leqslant n^{0.49}$*. Then, there exists a distribution* μ *over* $n \times d$ *matrices* Y *of the form* $Y = \sqrt{\beta} u_0 v_0^T + W + E$ *where* $\|E\|_\infty \leqslant \tilde{O}\left(1/\sqrt{n}\right)$*, with the following properties:*

- μ *is indistinguishable from the Gaussian distribution* $N(0,1)^{d \times n}$ *with respect to all multilinear polynomials of degree at most* $n^{0.001}$ *,*
- *the jointly-distributed random variables* W*,* u_0*,* v_0 *are independent,*
- *the marginal distribution of* v_0 *is supported on unit vectors with entries in* $\left\{-1/\sqrt{k}, 0, 1/\sqrt{k}\right\}$*,*
- *the marginal distribution of* u_0 *is uniform over* $\{-1,1\}^n$*,*
- *the marginal distribution of* W *is* $N(0,1)^{n \times d}$*.*

Informally speaking, Theorem I.3 conveys the following message. Any resilient algorithm for Sparse PCA can also distinguish the distribution μ over n-by-d matrices Y from the Gaussian distribution $N(0,1)^{n\times d}$. Therefore, if an estimator returned by this algorithm can be approximated by low-degree polynomials, then this algorithm cannot certify upper bounds of sparse eigenvalues of Gaussian matrices that are sharp enough to significantly improve the guarantees of Theorem I.2.

Sparse principal component analysis is intimately related to the problem of learning Gaussian mixtures. Indeed, for a vector v_0 with entries in $\{\pm 1/\sqrt{k}, 0\}$, sparse PCA can be rephrased as the problem of learning a non-uniform mixture M of three subgaussian distributions, one centered at zero, one centered at $\sqrt{\beta/k} \cdot u_0$ and the last at $-\sqrt{\beta/k} \cdot u_0$. As we will see, this is true even for the distribution μ used in Theorem I.3 Thus, from this perspective the result also provides interesting insight on the complexity of this problem. The theorem suggests that to distinguish between M and a standard Gaussian $W \sim N(0,1)^{n\times d}$, an algorithm would either need $d \gtrsim n^t$ samples or should not be computable by polynomials of degree at most $n^{0.001}$.

Resilient algorithms in the weak signal regime: Having cleared the picture for efficient algorithms in the strong signal regime, we may focus our attention to the weak signal settings $\beta \lesssim \sqrt{d/n}$. Surprisingly, in these settings adversarial perturbations do not change the computational landscape of the problem. As a matter of fact, a robust algorithm was already known. In fragile settings, the basic SDP program was proved (e.g. see [3]) to have the same guarantees as diagonal thresholding. But as the algorithm can certify the upper bounds $\|Mx\|^2 \leqslant k \cdot \|M\|_\infty^2$ and $\|Wx\|^2 \leqslant n + Ck\sqrt{n \log d}$ over k-sparse unit vectors $x \in \mathbb{R}^d$ and matrices $M \in \mathbb{R}^{n\times d}$, $W \sim N(0,1)^{n\times d}$ (where $C > 0$ is some absolute constant), it is therefore resilient to adversarial corruptions. We improve this latter upper bound showing that the algorithm can also certify the inequality $\|Wx\|^2 \leqslant n + Ck\sqrt{n}\log(d/\min\{k^2, n\})$, thus matching the guarantees of covariance thresholding and leading us to the following result.

Theorem I.4 (Perturbation Resilient Algorithm in the Weak Signal Regime)**.** *Given an* n*-by-*d *matrix* Y *of the form,*

$$Y = \sqrt{\beta} \cdot u_0 {v_0}^\mathsf{T} + W + E\,,$$

for $\beta > 0$*, a unit* k*-sparse vector* $v_0 \in \mathbb{R}^d$*, a Gaussian matrix* $W \sim N(0,1)^{n\times d}$*, a vector* $u_0 \in \mathbb{R}^n$ *independent of* W *with* $\|u_0\|^2 = \Theta(n)$*, and a matrix* $E \in \mathbb{R}^{n\times d}$ *satisfying* $\|E\|_\infty \lesssim \sqrt{\beta/k} \cdot \min\{\sqrt{\beta}, 1\}$*.*

Suppose that

$$\beta \gtrsim \min\left\{\frac{k}{\sqrt{n}}\sqrt{\log\left(2 + \frac{d}{k^2} + \frac{d}{n}\right)}, \frac{d}{n} + \sqrt{\frac{d}{n}}\right\}.$$

Then, there exists an algorithm that uses the basic SDP program for sparse PCA, and computes in polynomial time a unit vector $\hat{v} \in \mathbb{R}^d$ *such that*

$$1 - \langle \hat{v}, v_0\rangle^2 \leqslant 0.01$$

with probability at least 0.99*.*

Theorem I.4 says that among polynomial time algorithms, in the weak signal regime or whenever $\beta < 1$, the basic SDP achieves the best known guarantees. Furthermore, in contrast to thresholding and PCA algorithms, it works even in the presence of adversarial corruptions.

High degree certificates in the weak signal regime: A consequential observation of the previous paragraphs is that, perhaps, the Sum-of-Squares algorithm of larger degree can improve over the guarantees of the basic SDP even in the weak signal regime. Indeed in many settings, these guarantees can be improved observing that the (degree t) Sum-of-Squares algorithm can certify upper bounds of the form $\|Wx\|^2 \leqslant n + k\sqrt{(n/t)\log d}$ in time $d^{O(t)}$. Hence offering a smooth trade-off between sample complexity and running time.

Theorem I.5 (Perturbation Resilient Algorithm via Limited Exhaustive Search)**.** *Given an* n*-by-*d *matrix* Y *of the form,*

$$Y = \sqrt{\beta} \cdot u_0 {v_0}^\mathsf{T} + W + E\,,$$

for $\beta > 0$*, a unit* k*-sparse vector* $v_0 \in \mathbb{R}^d$*, a Gaussian matrix* $W \sim N(0,1)^{n\times d}$*, a vector* $u_0 \in \mathbb{R}^n$ *independent of* W *with* $\|u_0\|^2 = \Theta(n)$ *and a matrix* $E \in \mathbb{R}^{n\times d}$ *satisfying* $\|E\|_\infty \lesssim \sqrt{\beta/k} \cdot \min\{\sqrt{\beta}, 1\}$*.*

Suppose that for some positive integer $t \leqslant \frac{1}{\ln d}\min\{d, n\}$*,*

$$\beta \gtrsim \frac{k}{\sqrt{nt}}\sqrt{\log d}\,.$$

[10] This constraint is used to ensure that inequalities of the form $\beta \gtrsim \frac{k}{\sqrt{n \cdot D}}$ for any $D \leqslant n^{0.001}$ are never satisfied. Informally speaking, we restrict our statement to the settings where algorithms with guarantees similar to diagonal thresholding do not work.

Then, there exists an algorithm that computes in time $n^{O(1)}d^{O(t)}$ *a unit vector* $\hat{v} \in \mathbb{R}^d$ *such that*

$$1 - \langle \hat{v}, v_0 \rangle^2 \leqslant 0.01$$

with probability 0.99.

Whenever $k^2 \leqslant d^{1-\Omega(1)}$, Theorem I.5 provides better guarantees than Theorem I.4 (with worse running time).

It is also important to compare this result with the bound of Theorem I.2. For some t, we can determine the parameter regimes when one theorem provides better guarantees then the other for running time $d^{O(t)}$. Assume that $((t+1)n \log n)^{t+1} \gtrsim d \gtrsim (tn \log n)^t$. Then there exist constants $0 < C < C'$ such that:

- If $k^2 \leqslant d \cdot (Ct)^t$, we get $t \cdot \left(\frac{d}{k}\right)^{1/t} > \sqrt{\frac{n}{t} \log d}$, so in this case the guarantees in Theorem I.5 are better.
- If $k^2 \geqslant d\,(n \log n)^2 \cdot (C't)^t$, we get $t \cdot \left(\frac{d}{k}\right)^{1/t} < \sqrt{\frac{n}{t} \log d}$, so in this case the guarantees in Theorem I.2 are better.

Informally speaking, these conditions show that the guarantees in Theorem I.2 are better when the vector is only mildly sparse, $k^2 \gg d$, and the number of samples is very small.

Theorem I.5, along with Theorem I.4 and Theorem I.2 provides also a nice consequence, namely it allows us to state that f*or the problem of Sparse PCA, the Sum-of-Squares algorithm achieves the best known guarantees among perturbation resilient polynomial time algorithms. Furthermore, under the restrict computational model of low-degree polynomials, these guarantees are nearly optimal.*

1) Sharp bounds for the Wishart model: In the regime where $k \leqslant \sqrt{d}$, covariance thresholding succeeds for $\beta \gtrsim \frac{k}{\sqrt{n}}\sqrt{\log \frac{d}{k^2}}$. This turns into an asymptotic improvement over diagonal thresholding in the settings $d^{1-o(1)} \leqslant k^2 \leqslant o(d)$ but requires a constraint on the sample complexity of the form $n \geqslant k^2$, for which there is no evidence in the known lower bounds. This picture raises the following questions: can we obtain guarantees of the form $\beta \gtrsim \frac{k}{\sqrt{n}}\sqrt{\log \frac{d}{k^2}}$ even for $n \leqslant k^2$? And furthermore, can we improve over this logarithmic factor?

Studying low-degree polynomials we improve over this incomplete picture providing a new algorithm which succeed in recovering the sparse vector in polynomial time whenever $\beta \gtrsim \frac{k}{\sqrt{n}}\sqrt{\log \frac{d}{k^2}}$ and $n \gtrsim d^{1/\log\left(\frac{d}{k^2}\right)} + \log^5 d$. Thus obtaining an asymptotic improvement over diagonal thresholding in a significantly large set of parameters.

Concretely, the algorithm improves over the state-of-the-art whenever $d^{1/\log \frac{d}{k^2}} + \log^5 d \lesssim n \lesssim d^{1-\Omega(1)}$. In other words, the algorithm requires much fewer samples than covariance thresholding. This result is captured by the theorem below.

Theorem I.6 (Polynomials based Algorithm for the Strong Signal Regime)**.** *Given an* n*-by-*d *matrix* Y *of the form,*

$$Y = \sqrt{\beta} \cdot u_0 v_0^T + W\,,$$

for a unit vector $v_0 \in \mathbb{R}^d$ *with entries in* $\{\pm 1/\sqrt{k}, 0\}$*, a vector* u_0 *with i.i.d. entries satisfying* $\mathbb{E}\, u_i = 0$*,* $\mathbb{E}\, {u_i}^2 = 1$*,* $\mathbb{E}\, u_i^4 \leqslant O(1)$ *and a matrix* $W \in \mathbb{R}^{n \times d}$ *with i.i.d. entries satisfying* $\mathbb{E}\left[W_{ij}\right] = 0$*,* $\mathbb{E}\left[W_{ij}^2\right] = 1$*, such that* W *and* u_0 *are independent; suppose that* $n \gtrsim \log^5 d$*,* $d^{1-o(1)} \leqslant k^2 \leqslant d/2$*, and*

$$\beta \gtrsim \frac{k}{\sqrt{n}}\sqrt{\log\left(\frac{d}{k^2}\right) + \frac{\log d}{\log n}}\,.$$

Then, there exists a probabilistic algorithm that computes in polynomial time a unit vector $\hat{v} \in \mathbb{R}^d$ *such that*

$$1 - \langle \hat{v}, v_0 \rangle^2 \leqslant 0.01$$

with probability at least 0.99.

Along with Theorem I.6, we provide a fine-grained lower bound that in many settings matches the known algorithmic guarantees for the single spiked model. Some relevant lower bounds were already known. In [2] the authors used a reduction to the planted clique problem to provide evidence that in the weak signal regime[11] efficient algorithms cannot recover the sparse vector if $\beta \ll \frac{k}{\sqrt{n}}$. In [7] similar lower bound was obtained: in the weak signal regime low-degree polynomials cannot succeed if $\beta \lesssim \frac{k}{\sqrt{n}}$. This lower bounds fall short of matching the guarantees of diagonal thresholding by a logarithmic factor. Here, we show that whenever $k^2 \leqslant d^{1-\Omega(1)}$, polynomials of degree $O(\log d)$ cannot recover the sparse vector for $\beta \lesssim \min\left\{\sqrt{\frac{d}{n}}, \frac{k}{\sqrt{n}}\sqrt{\log d}\right\}$. In particular, we provide strong evidence that in the weak signal regime, in the settings where our polynomials based algorithm does not improve over the state-of-the-art, the known efficient algorithms (diagonal thresholding, basic SDP) are optimal up to constant factors.

Theorem I.7 (Lower Bound for Standard Sparse PCA, Informal)**.** *There exists a distribution* μ_k *over* k*-sparse* d*-dimensional unit vectors such that if* Y *is an* n*-by-*d *matrix of the form*

$$Y = \sqrt{\beta} \cdot u_0 {v_0}^{\mathsf{T}} + W\,,$$

for a vector v_0 *sampled from* μ_k*, a Gaussian matrix* $W \sim N(0,1)^{n \times d}$ *and a Gaussian vector* $u_0 \sim N(0, \mathrm{Id}_n)$ *such that* v_0, u_0, W *are distributionally independent, then the distribution of* Y *is indistinguishable from the Gaussian distribution* $N(0,1)^{n \times d}$ *with respect to all polynomials of degree* $D \leqslant n/\log^2 n$ *, whenever*

$$\beta \lesssim \min\left\{\sqrt{\frac{d}{n}}, \frac{k}{\sqrt{Dn}}\log\left(2 + \frac{D \cdot d}{k^2}\right)\right\}.$$

[11] Actually the parameter regime they considered is a proper subset of the weak signal regime.

2) Additional Results: Practical Algorithms and Experiments: From a practical perspective, the main issue with the results of Theorem I.2 is the reliance on solving large semidefinite programs, something that is often computationally too expensive to do in practice for the large-scale problems that arise in machine learning. In the same fashion of [30], from the insight of the SoS analysis we develop a *fast* spectral algorithm (which we will call *SVD-t*) with guarantees matching Theorem I.2 for degree $t \leqslant 3$ for some interesting family of adversaries. Our algorithm runs in time $O(nd \log n)$, which for high dimensional settings, can be considerably faster than algorithms that rely on computing the covariance matrix[12]. Furthermore, while not showing robustness of the algorithm (indeed the algorithm cannot certify upper bounds), we prove that SVD-t succeeds under the adversarial perturbations which are enough to prove Theorem I.3. Such adversarial settings are especially interesting since the problem has a nice geometric description in which the objective is to recover an approximately sparse vector planted in a random subspace. We remark that it is not known how to generalize the algorithm for larger t. Finally, we complement this result with experiments on synthetic data which highlights how in many practical settings the algorithm outperforms (*and outruns*) diagonal thresholding. The following theorem presents the guarantees of the algorithm in the spiked covariance model.

Theorem I.8 (Fast Spectral Algorithm for the Strong Signal Regime, Informal)**.** *Given an n-by-d matrix Y of the form,*

$$Y = \sqrt{\beta} u_0 {v_0}^{\mathsf{T}} + W + E\,,$$

for $\beta > 0$, a unit k-sparse vector $v_0 \in \mathbb{R}^d$, a Gaussian matrix $W \sim N(0,1)^{n\times d}$, a Gaussian vector $u_0 \sim N(0, \mathrm{Id}_n)$ such that v_0, u_0, W are distributionally independent, and $E \in \mathbb{R}^{n\times d}$ is a matrix from Theorem I.3 for $t = 3$.[13]

Suppose that $d \gtrsim n^3 \log d \log n$, $k \gtrsim n \log n$ and

$$\beta \gtrsim \frac{k}{\sqrt{n}} \left(\frac{d}{k}\right)^{1/3}.$$

Then there exits an algorithm that computes in time $O(nd \log n)$ a unit vector $\hat{v} \in \mathbb{R}^d$ such that

$$1 - \langle v_0, \hat{v} \rangle \leqslant 0.01$$

with probability at least 0.99.

We conclude our introduction with some notation.

Notation: We say that a unit vector $v \in \mathbb{R}^d$ is *flat* if its entries are in $\left\{\pm\frac{1}{\sqrt{t}}, 0\right\}$ for some t. For a matrix $M \in \mathbb{R}^{n\times d}$, we will denote its entry ij with M_{ij}. Depending on the context we may refer to the i-th row or the i-th column of M with M_i or m_i, we will specify it each time to avoid ambiguity. We call $\|M\|_1 = \sum_{i,j\in[d]} \left|M_{ij}\right|$ the ”absolute norm” of M. For a Gaussian matrix $W \sim N(0,1)^{n\times d}$, we denote with $w_1, \ldots, w_d$ its columns. For a vector $v \in \mathbb{R}^n$, we denote its j-th entry as v_j. We hide absolute constant multiplicative factors using the standard notations $\lesssim$, $O(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$, we hide multiplicative factors logarithmic in d using the notation $\tilde{O}(\cdot)$. For a set $S \subseteq [d] \times [d]$, and a matrix $M \in \mathbb{R}^{d\times d}$, we denote by $M[S]$ the matrix with entries $M[S]_{ij} = M_{ij}$ if $(i,j) \in S$, and $M[S]_{ij} = 0$ otherwise. For a matrix $M \in \mathbb{R}^{d\times d}$ and $\tau \in \mathbb{R}$, we define $\eta_\tau(M) \in \mathbb{R}^{d\times d}$ to be the matrix with entries

$$\eta_\tau(M)_{ij} = \begin{cases} M_{ij} & \text{if } \left|M_{ij}\right| \geqslant \tau \\ 0 & \text{otherwise.} \end{cases}$$

Remark I.9 (Strong and weak signal regimes in robust settings)*.* The attentive reader may have noticed how the notions of strong and weak signal regime should differ in the robust settings. Indeed there is no easy algorithm that looks at the spectrum of Y and begins to work as β approaches $\sqrt{\frac{d}{n}}$. In this sense, in the presence of an adversary the bound $\beta \lesssim \sqrt{\frac{d}{n}}$ looses significance. However we will continue using these terms to orientate ourselves and implicitly describe which are the desirable guarantees an algorithm should possess in a given regime. For this reason, when talking about weak-signal regime, our discussion will implicitly revolve around settings in which $\beta \gtrsim \frac{k}{\sqrt{n}}\sqrt{\log \frac{d}{k^2}}$.

II. Techniques

A. *Perturbation-resilience from Sparse Eigenvalue Certificates*

Here we outline the structure of our Sum-of-Squares algorithm and the basic SDP.

How robust should an algorithm be?: In light of our discussion in Section I-A, we would like efficient algorithms to be as resilient as exhaustive search. In order for such brute-force algorithm to recover the sparse vector v_0, there must be no other sparse vector x far from v_0 such that $\|Yx\| \approx \|Yv\|$. This also means that the adversary should not be able to plant a k-sparse vector z far from v_0 such that $\|Yx\| \gtrsim \|Yv_0\|$. To see what bound to enforce on the adversarial matrix, first observe that if E were the zero matrix then

$$\|Yv_0\| = \left\|Wv_0 + \sqrt{\beta}u_0\right\| \gtrsim \sqrt{n + \beta n}.$$

[12] While computing $Y^{\mathsf{T}}Y$ is necessary for covariance thresholding, the diagonal thresholding algorithm can run in time $O(k^2 + dn)$.

[13] More precisely, to prove Theorem I.3 we consider a specific distribution over matrices E (this distribution depends on v_0, u_0 and W), and here we mean that E is sampled from this distribution.

Now consider the following adversarial matrix, let x be a k-sparse unit vector with entries in $\left\{0, \pm 1/\sqrt{k}\right\}$ and such that the intersection between supp $\{x\}$ and supp $\{v_0\}$ is the empty set. With high probability $\|Wx\| \approx \sqrt{n}$. So let $z = \frac{1}{\|Wx\|} Wx$ and define E as the matrix with entries $E_{ij} = b \cdot z_i \cdot \operatorname{sign}(x_j)$, where $b > 0$ is some parameter that we will choose later. Then

$$\|Yx\| = \|(W+E)\,x\| = \left\|\left(\|Wx\| + b\sqrt{k}\right) z\right\| \approx \sqrt{n} + b\sqrt{k}.$$

Consequently, $\|Yx\| \geqslant \|Yv_0\|$ whenever $\sqrt{n} + b\sqrt{k} \geqslant \sqrt{n + \beta n}$. The inequality is true for $b \gtrsim \sqrt{\frac{\beta n}{k}} \cdot \min\left\{\sqrt{\beta}, 1\right\}$. In other words, the perturbation matrix must satisfy the bound:

$$\|E\|_\infty \leqslant \tilde{\Omega}\left(\sqrt{\frac{\beta}{k}} \cdot \min\left\{\sqrt{\beta}, 1\right\}\right). \qquad \text{(Bound-1)}$$

For a set of parameters d, n, k, β, we call an algorithm *perturbation resilient* if it can successfully recover the sparse vector for any adversarial perturbation satisfying bound Bound-1.

B. Algorithms that Certify Sparse Eigenvalues

For simplicity of the discussion we illustrate the idea of sparse eigenvaluex certificates for the Wigner model: $Y = \gamma v_0 {v_0}^{\mathsf{T}} + W + E$, where $\gamma > 0$, $v_0 \in \mathbb{R}^d$ is a k-sparse unit vector, $W \sim N(0,1)^{d \times d}$ and E is some matrix with small entries. Denote the set of k-sparse unit vectors by S_k. The starting idea is to turn the following intuition into an identifiability proof and then a Sum of Squares program: if $\hat{v}$ is a k-sparse unit vector which maximizes $v^{\mathsf{T}} Y v$ over S_k and γ is large enough, then with high probability $\langle \hat{v}, v_0\rangle^2 \geqslant 0.99$.

Concretely, observe that

$$\begin{aligned} &\text{on one side} && {v_0}^{\mathsf{T}} Y v_0 = \gamma + {v_0}^{\mathsf{T}} W v_0 + {v_0}^{\mathsf{T}} E v_0,\\ &\text{on the other} && \hat{v}^{\mathsf{T}} Y \hat{v} = \gamma \langle \hat{v}, v_0\rangle^2 + \hat{v}^{\mathsf{T}} W \hat{v} + \hat{v}^{\mathsf{T}} E \hat{v}. \end{aligned}$$

Combining the two and rearranging we obtain the inequality

$$\langle \hat{v}, v_0\rangle^2 \geqslant 1 - \frac{1}{\gamma} O\left(\max_{v \in S_k} v^{\mathsf{T}} W v + \max_{v \in S_k} v^{\mathsf{T}} E v\right).$$

Now, this is where certified upper bounds come in to the picture. There is an easy certificate (capture by SoS and the basic SDP) of the fact that for any matrix M, $\max_{v \in S_k} v^{\mathsf{T}} M v \leqslant \|M\|_\infty k$ Using such bound we get

$$\langle \hat{v}, v_0\rangle^2 \geqslant 1 - \frac{1}{\gamma} O\left(\max_{v \in S_k} v^{\mathsf{T}} W v + k \|E\|_\infty\right). \qquad \text{(II.1)}$$

Eq. (II.1) already shows how an algorithm that can certify sparse eigenvalues is perturbation resilient (in the sense of the previous paragraph). Indeed for $\|E\|_\infty = \varepsilon \cdot \gamma / k$, the inequality becomes

$$\langle \hat{v}, v_0\rangle^2 \geqslant 1 - O(\varepsilon) - \frac{1}{\gamma} O\left(\max_{v \in S_k} v^{\mathsf{T}} W v\right). \qquad \text{(II.2)}$$

At this point, the guarantees of the algorithm depend only on the specific certified upper bound on $\max_{v \in S_k} v^{\mathsf{T}} W v$ it can obtain.

For the Wishart Model $Y = \sqrt{\beta} u_0 {v_0}^{\mathsf{T}} + W + E$, the reasoning is essentially the same. However we need to work with $Y^{\mathsf{T}} Y - n\mathrm{Id}$ and carefully bound the cross terms. Similar to the Wigner model, the guarantees of the algorithm depend only on the certified upper bound on $\max_{v \in S_k} v^{\mathsf{T}} \left(W^{\mathsf{T}} W - n\mathrm{Id}\right) v$ it can obtain. For the rest of our preliminary discussion we go back to the Wishart model.

C. New Certificates via basic SDP

For a matrix $M \in \mathbb{R}^{d \times d}$, the basic SDP program[14]

$$\operatorname{argmax}\left\{\langle Y^{\mathsf{T}} Y, X\rangle \;\middle|\; X \succeq 0, \operatorname{Tr} X = 1, \|X\|_1 \leqslant k\right\} \qquad \text{(II.3)}$$

can certify two types of upper bound:

$$\langle M, X\rangle \leqslant \|M\|_\infty \cdot \sum_{i,j \leqslant d} X_{ij}\,, \qquad \text{(II.4)}$$

$$\langle M, X\rangle \leqslant \|M\|\,. \qquad \text{(II.5)}$$

The first follows using $\|X\|_1 \leqslant k$ and the second applying $X \succeq 0, \operatorname{Tr} X = 1$. These are enough to capture standard principal component analysis as well as diagonal and covariance thresholding.

Specifically, Eq. (II.5) can be used to certify the upper bound $\langle W^{\mathsf{T}} W - n\mathrm{Id}, X\rangle \leqslant O\left(d + \sqrt{dn}\right)$ – obtaining the guarantees of PCA – and Eq. (II.4) the bound $\langle W^{\mathsf{T}} W - n\mathrm{Id}, X\rangle \leqslant O\left(k \cdot \sqrt{n \log d}\right)$, as in diagonal thresholding[15]. Now these results were already known, but surprisingly a combination of the two bounds can also be used to show $\langle W^{\mathsf{T}} W - n\mathrm{Id}, X\rangle \leqslant k \cdot \sqrt{n \log(d/k^2)}$. Thus allowing us to match the guarantees of covariance thresholding.

Concretely, using the notation from the introduction,

$$\begin{aligned} \langle W^{\mathsf{T}} W - n\mathrm{Id}, X\rangle = \langle \eta_\tau\left(W^{\mathsf{T}} W - n\mathrm{Id}\right), X\rangle +&\\ \langle W^{\mathsf{T}} W - \eta_t\left(W^{\mathsf{T}} W\right), X\rangle.& \end{aligned}$$

Here $W^{\mathsf{T}} W - \eta_t\left(W^{\mathsf{T}} W\right)$ is a matrix with entries bounded (in absolute value) by τ for which we can plug in Eq. (II.4) and get

$$\langle W^{\mathsf{T}} W - \eta_t\left(W^{\mathsf{T}} W\right), X\rangle \leqslant \tau \cdot k$$

The same argument cannot be used for $\eta_\tau\left(W^{\mathsf{T}} W\right)$, but notice that this matrix is suspiciously close (up to an addition

[14] Recall $\|X\|_1 = \sum_{i,j \in [d]} \left|X_{ij}\right|$ is the "absolute norm".

[15] A more careful analysis can get $k \cdot \sqrt{n} \log(d/k)$, but we ignore it here.

of $n \cdot \mathrm{Id}$) to the thresholded covariance matrix obtained in covariance thresholding. Hence, taking $\tau = \sqrt{n \log(d/k^2)}$ and using Eq. (II.5), we get

$$\langle \eta_\tau \left(W^\mathsf{T} W - n\mathrm{Id} \right), X \rangle \leqslant O\left(k\sqrt{n \log \frac{d}{k^2}} \right),$$

where we get the spectral bound (almost) for free by the analysis in [13].

D. New certificates via higher-level Sum-of-Squares

1) Certificates via Certifiable Subgaussianity: The Sum-of-Squares algorithm can certify more refined bounds on sparse eigenvalues of $W \sim N(0,1)^{n\times d}$. In particular we can exploit Gaussian moments bound $\mathbb{E}\langle W_i, u\rangle^{2t} \leqslant t^t \cdot \|u\|^{2t}$ for all $t \in \mathbb{N}$, $u \in \mathbb{R}^d$.

Concretely let's see how to use such property to obtain an identifiability proof of a bound on the k-sparse norm of W. To this end let v be a k-sparse vector and let $s \in \{0,1\}^d$ be the indicator vector of its support (here we drop the subscript v_0 to ease the notation). Using Cauchy-Schwarz,

$$\begin{aligned}\|Wv\|^4 &= \left(\sum_{i \leqslant d} v_i \langle W_i, Wv\rangle \right)^2 \\ &\leqslant \left(\sum_{i \leqslant d} v_i^2 \right) \left(\sum_{i \leqslant d} s_i^2 \langle W_i, Wv\rangle^2 \right) \\ &\leqslant \left(\sum_{i \leqslant d} s_i^2 \langle W_i, Wv\rangle^2 \right).\end{aligned}$$

Then applying Holder's inequality with $1/p + 1/t = 1$, and using the fact that s is binary with norm k,

$$\begin{aligned}\left(\sum_{i \leqslant d} s_i^2 \langle W_i, Wv\rangle^2 \right) &\leqslant \left(\sum_{i \leqslant d} s_i^{2p} \right)^{1/p} \left(\sum_{i \leqslant d} \langle W_i, Wv\rangle^{2t} \right)^{1/t} \\ &\leqslant \|Wv\|^2 \cdot k^{1-1/t}. \\ &\quad \cdot \left(\sum_{i \leqslant d} \langle W_i, \frac{1}{\|Wv\|} Wv\rangle^{2t} \right)^{1/t}.\end{aligned}$$

This gets us to,

$$\|Wv\|^2 \leqslant k^{1-1/t} \cdot \left(\sum_{i \leqslant d} \langle W_i, \frac{1}{\|Wv\|} Wv\rangle^{2t} \right)^{1/t}. \qquad \text{(II.6)}$$

Now, whenever $d \gtrsim n^t t^t \log^t n$, the t-moment of the column vectors $W_1 \ldots, W_d$ converges with high probability. That is, for any unit vector u,

$$\frac{1}{d} \sum_{i \leqslant d} \langle W_i, u\rangle^{2t} \leqslant O(t^t). \qquad \text{(II.7)}$$

Thus, combining Eq. (II.6) and Eq. (II.7) we can conclude

$$\|Wv\|^2 \lesssim k^{1-1/t} \cdot d^{1/t} \cdot t.$$

The catch is that all the steps taken can be written as polynomial inequalities of degree at most $O(t)$. So we can certify the same bound through the Sum-of-Squares proof system.

2) Certificates via Limited Brute Force: Whenever the sparse vector v_0 is almost flat, that is when for all $i \in \operatorname{supp}\{v_0\}$ we have $|v_{0i}| \in \left[\frac{1}{C\sqrt{k}}, \frac{C}{\sqrt{k}} \right]$, the guarantees of diagonal thresholding can be improved at the cost of increasing its running time (see [7]).

Diagonal thresholding can be viewed as selecting the k vectors of the standard basis $e_1, \ldots, e_d$ maximizing $\|Ye_i\|^2$, and then returning a top eigenvector of the covariance matrix projected onto the span of such vectors. Indeed this formulation has an intuitive generalization, namely instead of looking at 1-sparse vectors, the algorithm could look into t-sparse vectors u with entries in $\left\{ \pm 1/\sqrt{t}, 0 \right\}$, pick the top $\binom{k}{t}$ and use them to recover v_0.

This idea can be translated into a certified upper bound for the sparse eigenvalues of $W \sim N(0,1)^{n\times d}$. Although we will be able to recover general sparse vectors, for the sake of this discussion we assume v_0 is flat.[16] Let's denote the set of t-sparse flat vectors by $\mathcal{N}_t$. Let $v_0 \in \mathbb{R}^d$ be a k-sparse vector and denote with D the uniform distribution over the vectors in $\mathcal{N}_t$ such that $\langle u, v_0\rangle = \sqrt{t/k}$. That is, the set of vectors u such that $\operatorname{supp}\{u\} \subseteq \operatorname{supp}\{v_0\}$ and with sign pattern matching the sign pattern of v restricted to $\operatorname{supp}\{u\}$.

Notice now that for any matrix $M \in \mathbb{R}^{d\times d}$,

$$v_0{}^\mathsf{T} M v_0 = \frac{k}{t} \mathop{\mathbb{E}}_{u,u' \sim D} u^\mathsf{T} M u'.$$

This equality *per se* is not interesting, but for a Gaussian matrix $W \sim N(0,1)^{n\times d}$, with high probability,

$$\max_{u,u' \in \mathcal{N}_t} \left| u^\mathsf{T} \left(W^\mathsf{T} W - n\mathrm{Id} \right) u' \right| \leqslant O\left(\sqrt{nt \log d} \right).$$

Thus, combining the two we get

$$\begin{aligned}v_0{}^\mathsf{T} \left(W^\mathsf{T} W - n\mathrm{Id} \right) v_0 &= \frac{k}{t} \mathop{\mathbb{E}}_{D} u^\mathsf{T} \left(W^\mathsf{T} W - n\mathrm{Id} \right) u' \\ &\leqslant \frac{k}{t} \max_{u,u' \in \mathcal{N}_t} \left| u^\mathsf{T} \left(W^\mathsf{T} W - n\mathrm{Id} \right) u' \right| \\ &\leqslant \frac{k}{\sqrt{t}} \sqrt{n \log d},\end{aligned}$$

which allows us to conclude that $\|Wv_0\|^2 \leqslant n + \frac{k}{\sqrt{t}} \sqrt{n \log d}$. This certificates can be proved using Sum-of-Squares, hence allowing us to improve over the basic SDP by a factor t in the settings $k^2 \leqslant d^{1-\Omega(1)}$.

E. Concrete lower bounds for perturbation-resilient algorithms

Sparse principal component analysis is what we often call a *planted problem*. These are problems that ask to

[16] So the Sum-of-Squares algorithm works in more general settings than the algorithm from [7].

recover some signal hidden by random or adversarial noise. The easiest way one could formulate a planted problem is its *distinguishing* version: where given two distributions, a *null* distribution without structure and a *planted* distribution containing the hidden signal, the objective is to determine with high probability whether a given instance was sampled from one distribution or the other.

A common strategy to provide evidence for *information-computation gap* in a certain planted problem is to prove that powerful classes of efficient algorithms are unable to solve it in the (conjecturally) hard regime. Indeed our goal here will be that of constructing two distributions under which low-degree polynomials take roughly the same values and hence cannot distinguish from which distribution the instance Y was sampled. Since low-degree polynomials cannot tell if Y has indeed the form $W+\sqrt{\beta}u_0{v_0}^{\top}+E$ (and therefore cannot solve the problem), this would mean they cannot be used to improve over the guarantees of Theorem I.2.

Our null distribution ν will be the standard Gaussian $N(0,1)^{n\times d}$. However, the main question is how to design the planted distribution μ. Recall Y takes the form $W+\sqrt{\beta}u_0{v_0}^{\top}+E$. If we set $E=0$, then our planted distribution corresponds to the single spike covariance model. We could get a lower bound for such problem but this would not help us in showing that the guarantees of Theorem I.2 are tight. On the other hand, if for example we choose E with the goal of planting a large eigenvalue, then the problem of distinguishing between ν and μ may become even easier than without adversarial perturbations.

This suggests that we should choose E very carefully, in particular we should design E so that $Y=W+\sqrt{\beta}u_0{v_0}^{\top}+E$ appears – to the eyes of a low-degree polynomial estimator – as a Gaussian distribution. Our approach will be that of constructing E so that the first few moments of μ will be Gaussian. This will lead us to Theorem I.3 through two basic observations: first, given two distributions with same first $2t$ moments, computing those first $2t$ moments won't help distinguishing between the two distributions. Second, for a Gaussian distribution $N(0,\mathrm{Id}_n)$, at least n^t samples are required in order for the $2t$-th moment of the empirical distribution to converge to $\mathbb{E}\left[w^{\otimes 2t}\right]$.

Concretely, we consider the following model: we choose iid gaussian vectors $z_1,\ldots,z_{n-1}\sim N(0,1)^d$, and a random vector $z_0\in\mathbb{R}^d$ with iid symmetric (about zero) coordinates that satisfies the following properties:

1) z_0 has approximately k large coordinates (larger than $\lambda\approx\sqrt{\beta n/k}$ by absolute value).
2) For any coordinate of z_0 its first $2t-2$ moments coincide with moments of $N(0,1)$, and its higher r-moments (for even r) are close to $\frac{k}{d}\lambda^r$.

Then we obtain the matrix $Y\in\mathbb{R}^{n\times d}$ applying a random rotation $R\in\mathbb{R}^{n\times n}$ to the $n\times d$ matrix with rows $z_0^{\top},z_1^{\top},\ldots,z_{n-1}^{\top}$. It is not difficult to see that indeed such Y can be written as $Y=W+\sqrt{\beta}u_0{v_0}^{\top}+E$, as in the model of Problem I.1.

Now, assume for simplicity that t is constant and denote the distribution of Y described above by μ and the standard Gaussian distribution $N(0,1)^{n\times d}$ by ν. An immediate consequence of our construction is that for any polynomial p of degree at most $2t-2$, $\mathbb{E}_{Y\sim\mu}\left[p(Y)\right]=\mathbb{E}_{Y\sim\nu}\left[p(Y)\right]$. Furthermore, in order to reliably tell the difference between $\mathbb{E}_{\mu}\left[p'(Y)\right]$ and $\mathbb{E}_{\nu}\left[p'(W)\right]$ for a polynomial of even degree $r\geqslant 2t$ (say up to $r=n^{0.001}$), we will need a precise estimate of such r-th moments and *hence* at least $n^{r/2\geqslant t}$ samples. This effect is then shown by proving that for multilinear polynomials $p(Y)$ of degree $D\leqslant n^{0.001}$, if $d\leqslant n^{0.99t-1}$ and $\beta n/k\leqslant n^{0.49}$, then the low-degree analogue of χ^2-divergence $\max\limits_{p(Y)\text{ of degree }\leqslant D}\frac{\left(\mathbb{E}_{\nu}\,p(Y)-\mathbb{E}_{\mu}\,p(Y)\right)^2}{\mathbb{V}_{\nu}\,P(Y)}$ is close to zero. Note that for technical reasons our analysis is restricted to the multilinear polynomials. As shown in [27]–[29] this restricted model of computation captures the best known algorithms for many planted problems.

F. Beyond limitations of CT via low-degree polynomials

An important aspect of the computation of lower bounds for low-degree polynomials is that they may provide valuable insight on how to construct an optimal algorithm. Indeed low-degree polynomials capture many spectral properties of linear operators; for example, the largest singular value of a d-dimensional linear operator with a spectral gap can be approximated by $\lesssim\log d$ degree polynomial in its entries.

We discuss here how they can be used to improve over the guarantees of Covariance Thresholding

Why Covariance Thresholding doesn't work with small sample size: In order to improve over Covariance Thresholding, the first question we need to understand is whether the algorithm could actually work in a larger set of parameters than the one currently known. The answer is no. Recall that for $k^2\leqslant d/2$ and $n\leqslant d$ Covariance Thresholding (with an appropriate choice of thresholding parameter τ) works if $\beta\gtrsim\frac{k}{\sqrt{n}}\sqrt{\log\frac{d}{k^2}+\log\frac{d}{n}}$, and so for $n\geqslant k^2$ and $d^{1-o(1)}\leqslant k^2\leqslant o(d)$ this is asymptotically better than the guarantees of SVD, SVD+Thresholding and Diagonal Thresholding.

It is not difficult to see that Covariance Thresholding with $\tau\geqslant\Omega(\sqrt{n\log d})$ cannot have better guarantees than Diagonal Thresholding. So we consider $\tau\leqslant o(\sqrt{n\log d})$.

Notice that $d^{1-o(1)}\leqslant k^2\leqslant o(d)$ and $n\geqslant k^2$ imply $n>d^{1-o(1)}$. The assumption $n>d^{1-o(1)}$ is crucial for Covariance Thresholding. To show this, it is enough to prove that for some unit $x\in\mathbb{R}^d$, $x^{\top}\eta_\tau\left(Y^{\top}Y-n\mathrm{Id}\right)x>d^{1-o(1)}$. Indeed, as on the other hand $\left|{v_0}^{\top}\eta_\tau\left(Y^{\top}Y-n\mathrm{Id}\right)v_0\right|\approx\beta n$, this would mean that for $\beta\ll d^{1-o(1)}/n$ the top eigenvectors of $\eta_\tau\left(Y^{\top}Y-n\mathrm{Id}\right)$ are uncorrelated with v_0. Additionally, since $\sqrt{\frac{d}{n}} \ll \left(\frac{d^{1-o(1)}}{n}\right)$ in these settings SVD+Thresholding has significantly better guarantees.

An x satisfying our inequality is easy to find, for example any row $W_1, \dots, W_n \in \mathbb{R}^d$ satisfies $W_i^\top \eta_\tau (Y^\top Y - n\mathrm{Id}) W_i > d^{1-o(1)} \|W_i\|^2$ with high probability .

Hence Covariance Thresholding doesn't provide better guarantees than SVD or Diagonal Thresholding if $n \leqslant d^{1-\Omega(1)}$ (for example, if $n = d^{0.99}$).

1) Polynomials based algorithm: Theorem I.7 shows that if $k^2 \leqslant d^{1-\Omega(1)}$ and $\beta \leqslant o\left(\frac{k}{\sqrt{n}}\sqrt{\log d}\right)$, it is unlikely that polynomial time algorithms can solve the problem. So to get an asymptotic improvement over Diagonal Thresholding we need $k^2 \geqslant d^{1-o(1)}$.

However, notice there is no condition $n \geqslant d^{1-o(1)}$ in our lower bound. This suggests that there might be an algorithm that is asymptotically better than SVD and Diagonal Thresholding for small n, for example $n = d^{0.99}$ or $n = d^{0.01}$. Indeed, we show that there exists a polynomial time algorithm that can recover the sparse vector v_0 with entries in $\{0, \pm 1/\sqrt{k}\}$ as long as $d^{1-o(1)} \leqslant k^2 \leqslant d/2$, $\beta \gtrsim \frac{k}{\sqrt{n}}\sqrt{\log \frac{d}{k^2} + \frac{\log d}{\log n}}$ and $n \gtrsim \log^5 d$. In particular, if $d^{1-o(1)} \leqslant k^2 \leqslant o(d)$ and $d^{0.01} \leqslant n \leqslant d^{0.99}$, this algorithm has asymptotically better guarantees than Diagonal Thresholding, SVD, SVD+Thresholding, and Covariance Thresholding.

Our algorithm is based on the approach introduced in [28] for commutinity detection in stochastic block model. An informal description of the algorithm is as follows: we compute some symmetric matrix $P(Y) \in \mathbb{R}^{d \times d}$ whose entries are polynomials $P_{jj'}(Y)$ in the entries of Y of degree $O(\log d)$. The algorithm outputs a top eigenvector of this matrix, which we prove to be highly correlated with v_0. Notice that since the degrees of involved polynomials are $O(\log d)$, simple evaluation takes time $(nd)^{O(\log d)}$. However, we can compute a very good approximation to the values of these polynomials in time $(nd)^{O(1)}$ using a *color coding* technique (this part of the algorithm uses internal randomness).

More precisely, for $j, j' \in [d]$ we compute multilinear polynomials $P_{jj'}(Y)$ of degree $O(\log d)$ such that for every $j \neq j'$, $\mathbb{E} P_{jj'}(Y) = v_0(j) v_0(j')$, and for every $j \in [d]$, $P_{jj}(Y) = 0$. Then we show that variance of $P_{jj'}(Y)$ is small so that $\mathbb{E}\|P(Y) - v_0 {v_0}^\top\|_{\mathrm{F}}^2 < o(1)$. This implies that with probability $1 - o(1)$, $\|P(Y) - v_0 {v_0}^\top\|_{\mathrm{F}}^2 < o(1)$, so the top eigenvector of $P(Y)$ is highly correlated with either v_0 or $-v_0$.

To bound the variance, we represent each monomial as a bipartite multigraph $G = (R, C, E)$, with bipartition $R \subset [n]$ which corresponds to rows of Y and $C \subset [d]$ which correspond to columns of Y. Since the variance is a sum of monomials, we compute the contribution of each monomial and bound the number of corresponding multigraphs. Finally, we show that there exists a polynomial such that in the parameter regime $d^{1-o(1)} \leqslant k^2 \leqslant d/2$, $\beta \gtrsim \frac{k}{\sqrt{n}}\sqrt{\log \frac{d}{k^2} + \frac{\log d}{\log n}}$ and $n \gtrsim \log^5 d$, there is no group of monomials with large contribution in the variance, so we can conclude that this polynomial is a good estimator.

After showing that there are good polynomial estimators of degree $O(\log d)$, we approximately compute them using color coding. All monomials of the polynomials $P_{jj'}$ that we consider have the same structure (in the sence that the graphs corresponding to these monomials are isomorphic). Each of them has the same number r of vertices which correspond to rows and the same number c of vertices which correspond to columns. We show that for each coloring of $[n]$ in r color and each coloring of $[d]$ in c colors, we can in time $(nd)^{O(1)}$ compute the sum of monomials of $P_{jj'}$ colored exactly in colores from $[r]$ and $[c]$. If we average these values over large enough set of random colorings (of size $(nd)^{O(1)}$), we get a value very close to $P_{jj'}(Y)$.

One important advantage of this polynomial-based algorithm is that we only need the following assumptions on W: that the entries of W are i.i.d., $\mathbb{E}\, W_{ij} = 0$ and $\mathbb{E}\, W_{ij}^2 = 1$.[17] All previously known algorithms require bounds on entries or the spectral norm of $W^\top W$ (or related matrices, e.g. thresholded $W^\top W$), so they require χ^2 tail bounds.

Acknowledgment

The authors would like to thank the anonymous reviewers whose suggestions greatly improved the exposition of the paper.

References

[1] A. A. Amini and M. J. Wainwright, "High-dimensional analysis of semidefinite relaxations for sparse principal components," *Ann. Statist.*, vol. 37, no. 5B, pp. 2877–2921, 10 2009. [Online]. Available: https://doi.org/10.1214/08-AOS664 1, 2

[2] Q. Berthet and P. Rigollet, "Computational lower bounds for sparse PCA," *CoRR*, vol. abs/1304.0828, 2013. 1, 2, 6

[3] ——, "Optimal detection of sparse principal components in high dimension," *Ann. Statist.*, vol. 41, no. 4, pp. 1780–1815, 08 2013. [Online]. Available: https://doi.org/10.1214/13-AOS1127 1, 5

[4] J. Baik, G. Ben Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *Ann. Probab.*, vol. 33, no. 5, pp. 1643–1697, 09 2005. [Online]. Available: https://doi.org/10.1214/009117905000000233 2

[5] R. Krauthgamer, B. Nadler, D. Vilenchik *et al.*, "Do semidefinite relaxations solve sparse pca up to the information limit?" *The Annals of Statistics*, vol. 43, no. 3, pp. 1300–1322, 2015. 2

[17] Indeed, prior work [31] observed that polynomial-based algorithms require only first and second moment conditions on the noise entries for a broad range of matrix and tensor estimation problems.

[6] I. M. Johnstone and A. Y. Lu, “On consistency and sparsity for principal components analysis in high dimensions,” *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 682–693, 2009, pMID: 20617121. [Online]. Available: https://doi.org/10.1198/jasa.2009.0121 2

[7] Y. Ding, D. Kunisky, A. S. Wein, and A. S. Bandeira, “Subexponential-time algorithms for sparse pca,” 2019. 2, 6, 9

[8] A. d’Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. Lanckriet, “A direct formulation for sparse pca using semidefinite programming,” in *Advances in neural information processing systems*, 2005, pp. 41–48. 2

[9] T. T. Cai, Z. Ma, and Y. Wu, “Sparse pca: Optimal rates and adaptive estimation,” *Ann. Statist.*, vol. 41, no. 6, pp. 3074–3110, 12 2013. [Online]. Available: https://doi.org/10.1214/13-AOS1178 2

[10] S. B. Hopkins, P. K. Kothari, A. Potechin, P. Raghavendra, T. Schramm, and D. Steurer, “The power of sum-of-squares for detecting hidden structures,” *CoRR*, vol. abs/1710.05017, 2017. 2

[11] T. Ma and A. Wigderson, “Sum-of-squares lower bounds for sparse PCA,” in *NIPS*, 2015, pp. 1612–1620. 2

[12] S. B. Hopkins, P. K. Kothari, A. Potechin, P. Raghavendra, T. Schramm, and D. Steurer, “The power of sum-of-squares for detecting hidden structures,” in *FOCS*. IEEE Computer Society, 2017, pp. 720–731. 2, 4

[13] Y. Deshpande and A. Montanari, “Sparse PCA via covariance thresholding,” in *NIPS*, 2014, pp. 334–342. 2, 9

[14] P. Huber, *Robust Statistics*, ser. Wiley Series in Probability and Statistics. Wiley, 1981. [Online]. Available: https://books.google.it/books?id=hVbhlwEACAAJ 2

[15] S. Morgenthaler, “A survey of robust statistics,” *Statistical Methods and Applications*, vol. 15, no. 3, pp. 271–293, 2007. [Online]. Available: https://doi.org/10.1007/s10260-006-0034-4 2

[16] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, “Robustness of classifiers: from adversarial to random noise,” in *NIPS*, 2016. 2

[17] E. Abbe and C. Sandon, “Achieving the ks threshold in the general stochastic block model with linearized acyclic belief propagation,” in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 1334–1342. 3

[18] U. Feige and J. Kilian, “Heuristics for semirandom graph problems,” *J. Comput. Syst. Sci.*, vol. 63, no. 4, pp. 639–671, 2001. 3

[19] O. Guédon and R. Vershynin, “Community detection in sparse networks via grothendieck’s inequality,” *Probability Theory and Related Fields*, vol. 165, 11 2014. 3

[20] A. Moitra, W. Perry, and A. S. Wein, “How robust are reconstruction thresholds for community detection?” in *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, 2016, pp. 828–841. [Online]. Available: https://doi.org/10.1145/2897518.2897573 3

[21] A. Montanari and S. Sen, “Semidefinite programs on sparse random graphs and their application to community detection,” in *STOC*. ACM, 2016, pp. 814–827. 3

[22] K. Makarychev, Y. Makarychev, and A. Vijayaraghavan, “Learning communities in the presence of errors,” in *COLT*, ser. JMLR Workshop and Conference Proceedings, vol. 49. JMLR.org, 2016, pp. 1258–1291. 3

[23] J. Banks, S. Mohanty, and P. Raghavendra, “Local statistics, semidefinite programming, and community detection,” *CoRR*, vol. abs/1911.01960, 2019. [Online]. Available: http://arxiv.org/abs/1911.01960 3

[24] D. Sherrington and S. Kirkpatrick, “Solvable model of a spin-glass,” *Physical review letters*, vol. 35, no. 26, p. 1792, 1975. 3

[25] A. Montanari, “Optimization of the sherrington-kirkpatrick hamiltonian,” in *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2019, pp. 1417–1433. 3

[26] A. S. Bandeira, D. Kunisky, and A. S. Wein, “Computational hardness of certifying bounds on constrained PCA problems,” in *11th Innovations in Theoretical Computer Science Conference, ITCS 2020, January 12-14, 2020, Seattle, Washington, USA*, 2020, pp. 78:1–78:29. [Online]. Available: https://doi.org/10.4230/LIPIcs.ITCS.2020.78 3

[27] B. Barak, S. B. Hopkins, J. A. Kelner, P. Kothari, A. Moitra, and A. Potechin, “A nearly tight sum-of-squares lower bound for the planted clique problem,” in *FOCS*. IEEE Computer Society, 2016, pp. 428–437. 4, 10

[28] S. B. Hopkins and D. Steurer, “Efficient bayesian estimation from few samples: Community detection and related problems,” in *FOCS*. IEEE Computer Society, 2017, pp. 379–390. 4, 10, 11

[29] S. B. K. Hopkins, “Statistical inference and the sum of squares method,” 2018. 4, 10

[30] S. B. Hopkins, T. Schramm, J. Shi, and D. Steurer, “Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors,” in *STOC*. ACM, 2016, pp. 178–191. 7

[31] J. Ding, S. B. Hopkins, and D. Steurer, “Estimating Rank-One Spikes from Heavy-Tailed Noise via Self-Avoiding Walks,” *arXiv e-prints*, p. arXiv:2008.13735, Aug. 2020. 11