Higher degree sum-of-squares relaxations robust against oblivious outliers

Conference Paper

Author(s):
d'Orsi, Tommaso; Nasser, Rajai; Novikov, Gleb; Steurer, David

Publication date:
2023

Permanent link:
https://doi.org/10.3929/ethz-b-000587171

Originally published in:
https://doi.org/10.1137/1.9781611977554.ch134

Funding acknowledgement:
815464 - Unified Theory of Efficient Optimization and Estimation (EC)

Higher degree sum-of-squares relaxations robust against oblivious outliers [*]

Tommaso d'Orsi[†] Rajai Nasser[‡] Gleb Novikov[†] David Steurer[†]

November 14, 2022

Abstract

We consider estimation models of the form $\mathbf{Y} = X^* + \mathbf{N}$, where X^* is some m-dimensional structured signal we wish to recover, and $\mathbf{N}$ is symmetrically distributed noise that may be unbounded in all but a small α fraction of the entries. This setting captures problems such as (sparse) linear regression, (sparse) principal component analysis (PCA), and tensor PCA, even in the presence of oblivious outliers and heavy-tailed noise.

We introduce a family of algorithms that under mild assumptions recover the signal X^* in all estimation problems for which there exists a sum-of-squares algorithm that succeeds in recovering the signal X^* when the noise $\mathbf{N}$ is Gaussian. This essentially shows that it is enough to design a sum-of-squares algorithm for an estimation problem with Gaussian additive noise in order to get the algorithm that works with the symmetric noise model.

Our framework extends far beyond previous results on symmetric noise models and is even robust to an ε-fraction of adversarial perturbations. As concrete examples, we investigate two problems for which no efficient algorithms were known to work for heavy-tailed noise: tensor PCA and sparse PCA.

For the former, our algorithm recovers the principal component in polynomial time when the signal-to-noise ratio is at least $\tilde{O}(n^{p/4}/\alpha)$, that matches (up to logarithmic factors) current best known algorithmic guarantees for Gaussian noise. For the latter, our algorithm runs in quasipolynomial time and matches the state-of-the-art guarantees for quasipolynomial time algorithms in the case of Gaussian noise. Using a reduction from the planted clique problem, we provide evidence that the quasipolynomial time is likely to be necessary for sparse PCA with symmetric noise.

In our proofs we use bounds on the covering numbers of sets of pseudo-expectations, which we obtain by certifying in sum-of-squares upper bounds on the Gaussian complexities of sets of solutions. This approach for bounding the covering numbers of sets of pseudo-expectations may be interesting in its own right and may find other application in future works.

[*] This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 815464).

[†] ETH Zürich.

[‡] Google Zürich. This work was done at ETH Zürich.

Contents

1 Introduction

Consider an estimation problem over $\mathbb{R}^m$, in which we observe (a realization of)[1] the random variable $\boldsymbol{Y} = X^* + \boldsymbol{N}$ where $X^* \in \Omega \subseteq \mathbb{R}^m$ is some structured signal we seek to recover and $\boldsymbol{N}$ is some additive noise. This generic primitive captures widely studied models such as compressed sensing, linear regression, principal component analysis, clustering mixture distributions, matrix completion or tensor principal component analysis.

From both a statistical and computational point of view, as one weakens the assumptions on the noise $\boldsymbol{N}$, the task of reconstructing the hidden signal X^* becomes harder. [2] Recent years have seen tremendous advances in the design of efficient algorithms able to recover the planted structure X^*, under weaker and weaker assumptions on the noise (e.g. [KKM18, Hop20, dKNS20, DdNS21, BDJ+22]). In particular, a certain line of work [CLMW11, ZLW+10, TJSO14, BJKK17, SBRJ19, SZF20, PJL20, dNS21, Cd22, dLN+21] has aimed to identify the weakest possible requirements on the signal-to-noise ratio so that it is still possible to efficiently recover the signal X^* with *vanishing* error. In this context an established model[3] is that of assuming the entries of $\boldsymbol{N}$ to be (i) independent, (ii) symmetric about zero and (iii) to have some probability mass in a neighborhood of 0. That is, to satisfy $\mathbb{P}[|\boldsymbol{N}_i| \leqslant 1] \geqslant \alpha$, for $i \in [m]$ with the parameter α possibly vanishingly small.

Remarkably, the framework emerging from these results shows that the Huber loss estimator– when equipped with an appropriate regularizer– offers provably optimal error guarantees among efficient estimators. In particular, it recovers the error convergence rates of classical least squares algorithms in the presence of Gaussian noise.

The general recipe behind these results relies on two main points: first an upper bound on the gradient of the Huber loss function at the true solution X^*,[4] and second a lower bound on the curvature of the loss function (in the form of a local strong convexity bound) within a *structured* neighborhood of X^*. Here the structured neighborhood of X^* is a superset $\bar{\Omega} \supseteq \Omega$. The curvature of the loss function depends on the directions (and the radius) considered and (one expect that) it is sharper in the directions contained in Ω. Thus one can establish stronger statistical guarantees by forcing the minimizer of the loss function to be in a small set of directions close to Ω. The crux of the argument is that the set $\bar{\Omega}$ is controlled by the regularizer: If the chosen regularizer is *norm decomposable*[5] *with respect to a meaningful set* $\bar{\Omega}$, then indeed it will force the minimizer to fall in one of the desired structured directions.

The inherent consequence of this approach is that, in settings where no such decomposable norm regularizer is known –such as for tensor principal component analysis– these estimators cannot provide *any* error guarantees. In this paper, we overcome this limitation and introduce a family of algorithms (based on sum-of-squares) that recover the parameter X^* for a remarkably large set of models. More concretely, our result can be informally read (under certain reasonable conditions) as:

Whenever there exists a degree-ℓ sum-of-squares algorithm that recovers X^* from $\boldsymbol{Y}$ when the

[1] We denote random variables in boldface.

[2] The complexity of the problem may also be affected by the structure of X^*.

[3] Sometimes denoted the *oblivious adversarial model*.

[4] The attentive reader may notice that no such upper bound exists for the least square estimator, under the noise assumptions above. This is evidence confirming the intuition that "least squares estimator are fragile to outliers".

[5] We formally define decomposability in Appendix D.

entries of $\mathbf{N}$ are Gaussian with standard deviation[6] $\sigma = (1 + \|X^*\|_{\max})/\alpha$*, there also exists an algorithm running in time* $m^{O(\ell)}$ *that recovers* X^* *with the same guarantees, even if* $\mathbf{N}$ *only satisfies (i), (ii), (iii).*

In other words, we introduce a framework that allows to directly generalize sum-of-squares algorithms designed to recover the hidden signal in the presence of Gaussian noise, to the significantly more general settings of symmetric noise.

Our result relies on a novel use of the sum-of-squares hierarchy. The core of the argument consists of *bounds on the covering number of sets of pseudo-expectations*, which we obtain via sum-of-squares certificates of the Gaussian complexity of the space of solutions. We then use these small covers to ensure that feasible solutions must fall in one of a few directions close to X^*.

1.1 Results

Our main result is the following meta-theorem for recovering a structured signal from symmetric noise.

Theorem 1.1 (SoS meta-theorem, Informal)**.** *Let* $m, \ell \in \mathbb{N}$. *Let* $\Omega \subset \mathbb{R}^m$ *be a set defined by at most* $m^{O(1)}$ *polynomial constraints of degree*[7] *at most* ℓ. *Suppose that for some* $r > 0$ *and* $\gamma \geqslant r\sqrt{\ln m}$ *the following bounds are certifiable by degree* $O(\ell)$ *sum-of-squares proofs (from the constraints that define* Ω*):*

(1) $\sup_{X\in\Omega}\|X\|_{\max} \leqslant 1$,

(2) $\sup_{X\in\Omega}\|X\|_2 \leqslant r$,

(3) $\mathbb{E}_{\mathbf{W}\sim N(0,\mathrm{Id}_m)} \sup_{X\in\Omega}\langle X, \mathbf{W}\rangle \leqslant \gamma$.

Let $0 < \alpha \leqslant 1$ *and let* $\mathbf{N}$ *be a random* m*-dimensional vector with independent (but not necessarily identically distributed) symmetric about zero*[8] *entries satisfying* $\mathbb{P}[|\mathbf{N}_i| \leqslant 1] \geqslant \alpha$ *for all* $i \in [m]$.

There exists an algorithm running in time $m^{O(\ell)}$ *that on input* $\mathbf{Y} = X^* + \mathbf{N}$ *outputs* $\hat{\mathbf{X}}$ *satisfying*

$$\left\|X^* - \hat{\mathbf{X}}\right\|_2^2 \leqslant O\left(\gamma/\alpha\right)$$

with high probability.

Moreover, for $\varepsilon \lesssim \frac{\gamma^2}{r^2 \cdot m \ln m}$*, the same result holds if an arbitrary (adversarially chosen)* ε*-fraction of entries of* $\mathbf{Y}$ *is replaced by adversarially chosen values.*

It is possible to gain an understanding of the importance of Theorem 1.1 even before applying it to specific problems. First, notice that if $\gamma/\alpha \leqslant o(r^2)$, the error guarantees are non-trivial. In particular this means that the fraction α of entries with bounded noise can be vanishingly small and the algorithm can still reconstruct a meaningful estimate. Second, observe how the error rate crucially depends on the upper bounds we are able to certify on the Gaussian complexity of the space of solutions Ω. By certifying tighter bounds on it one can obtain tighter guarantees on the error of the estimation. This shows the existence of a trade-off between error of the estimate and running time. Finally we remark that the algorithm is robust to an ε-fraction of adversarial corruptions, the magnitude of ε will become clearer when discussing the various applications.

Next we apply Theorem 1.1 to specific problems.

[6] As we will see in the context of sparse PCA problem, it is unlikely that we can relax the condition $\sigma > \|X\|_{\max}$.

[7] These constaints may use up to $m^{O(1)}$ auxiliary variables, and degrees of all polynomials in all variables are at most ℓ.

[8] I.e., $\mathbf{N}_i$ and $-\mathbf{N}_i$ have the same distribution for every $i \in [m]$.

Tensor principal component analysis. We consider the following tensor PCA model (we remark that one may consider further tensor models, in Section 5 we study other versions of Model 1.2 as well).

Model 1.2 (Tensor PCA with asymmetric tensor noise)**.** Let $n, p \in \mathbb{N}$, $n, p \geqslant 2$, and $0 < \alpha \leqslant 1$. We observe (an instance of) $\boldsymbol{Y} = \lambda \cdot v^{\otimes p} + \boldsymbol{N}$, where $\lambda > 0$, $v \in \mathbb{R}^n$ is an unknown unit vector and $\boldsymbol{N}$ is a random order p tensor with independent (but not necessarily identically distributed) symmetric about zero entries such that

$$\mathbb{P}[|\boldsymbol{N}_{i_1\ldots i_p}| \leqslant 1] \geqslant \alpha\,, \quad \text{for all } 1 \leqslant i_1, \ldots, i_p \leqslant n\,.$$

In the significantly more restrictive settings when the noise is standard Gaussian (captured by Model 1.2 by the special case with $\alpha \geqslant \Omega(1)$), this model was studied in [MR14, HSS15]. In these settings, one can recover the hidden vector v in exponential time whenever the signal-to-noise ratio λ is at least $\Omega(\sqrt{n})$, but existing polynomial time algorithms are known to require at least $\lambda \geqslant \Omega(n^{p/4})$. Moreover, evidence of an information-computation gap exists in the literature in the form of lower bounds against different computational models (sum-of-squares lower bounds [HSS15, HKP+17] or low degree polynomial lower bounds [KWB19]), showing that these computational models cannot recover the hidden vector in polynomial time if $\lambda < n^{p/4}/\mathrm{polylog}(n)$.

Less restrictive noise models have been considered more recently. [DHS20] proved that when the noise has zero mean and bounded variance and v is a *random* vector whose entries have small fourth moment, then one can recover it as long as $\lambda \geqslant \Omega(n^{p/4})$. Later, [AY21] showed that if the noise has zero mean and bounded variance, there exists an algorithm that, under mild assumption on the magnitude of the entries of v, can recover v as long as $\lambda \gtrsim n^{p/4} \cdot (\ln n)^{1/4}$.

However, an application of Theorem 1.1 shows that whenever the entries of the noise are symmetric about zero, *no* assumption on the moments is needed to recover the parameter v. The application of Theorem 1.1 only relies on known sum-of-squares certificates for the injective tensor norm of random tensors [HSS15].

Theorem 1.3 (Robust Tensor PCA)**.** *Let $p \geqslant 2$. There exists an absolute constant $C > 1$, and an algorithm running in time $n^{O(p)}$ that, given $\boldsymbol{Y}$ as in Model 1.2, outputs a unit vector $\hat{\boldsymbol{v}} \in \mathbb{R}^n$ satisfying*

$$|\langle v, \hat{\boldsymbol{v}} \rangle| \geqslant 0.99$$

with high probability, whenever

- *If p is even: $\lambda \geqslant \frac{C}{\alpha} \cdot n^{p/4}$ and $\|v\|_{\max} \leqslant \frac{\alpha^{1/p}}{C} \cdot n^{-1/4}$.*
- *If p is odd: $\lambda \geqslant \frac{C(p \ln n)^{1/4}}{\alpha} \cdot n^{p/4}$ and $\|v\|_{\max} \leqslant \frac{\alpha^{1/p}}{C(p \ln n)^{1/4p}} \cdot n^{-1/4}$.*

Moreover, if p is odd, the algorithm recovers the sign of v, that is, $\langle v, \hat{\boldsymbol{v}} \rangle \geqslant 0.99$ with high probability.

Furthermore, for $\varepsilon \leqslant \left(C \cdot p \cdot n^{p/2} \cdot \ln n\right)^{-1}$, the same result holds if an arbitrary (adversarially chosen) ε-fraction[9] of entries of $\boldsymbol{Y}$ is replaced by adversarially chosen values.

[9] For odd p we allow slightly greater fraction of corruptions $\varepsilon \leqslant \left(C \cdot p \cdot n^{p} \cdot \ln n\right)^{-1/2}$.

Let us briefly and informally describe how this result can be obtained from Theorem 1.1. Consider the case when p is odd[10]. Let $b = \frac{\alpha^{1/p}}{C(p\ln n)^{1/4p}} \cdot n^{-1/4}$. We may rescale $\boldsymbol{Y}$ by $1/(\lambda b^p) \leqslant 1$ so that $\|X^*\|_{\max} \leqslant 1$ and the bound $\mathbb{P}[|\boldsymbol{N}_{i_1 \dots i_p}| \leqslant 1] \geqslant \alpha$ still holds for all $1 \leqslant i_1, \dots, i_p \leqslant n$. Note that now $r := \|X^*\|_2 = 1/b^p$. So we trivially have the desired sum-of-squares certificates for (1) and (2) in Theorem 1.1. Most importantly, from [HSS15] we know that for the set Ω of rank-one symmetric tensors of norm r there is a degree $O(p)$ sum-of-squares proof that certifies the bound

$$\mathop{\mathbb{E}}_{\boldsymbol{W}\sim N(0,\mathrm{Id}_{n^p})} \sup_{X\in\Omega} \langle X, \boldsymbol{W}\rangle \leqslant O\big(p \cdot (\ln n) \cdot n^p\big)^{1/4} \cdot r\,.$$

Thus using the value on the right-hand side as γ, we get that $\hat{\boldsymbol{X}}$ that is obtained from Theorem 1.1 satisfies

$$\big\|X^* - \hat{\boldsymbol{X}}\big\|_2^2 \leqslant O\big(\gamma/\alpha\big) \leqslant O\left(\frac{r}{(Cb)^p}\right) = O\left(\frac{1}{C^p}\right) \cdot r^2\,,$$

and hence $\hat{\boldsymbol{X}}$ is highly correlated with X^* and the result follows[11].

Concerning the noise $\boldsymbol{N}$, it is easy to observe that the algorithm works with symmetric heavy tailed noise (e.g., Cauchy noise) and achieves guarantees similar to the best known guarantees for standard Gaussian noise. Moreover, the number of adversarial corruptions that the algorithm allows is nearly optimal: For instance, for constant even p and constant α our bound on the entries allows v to be $O(\sqrt{n})$-sparse. Hence for such v, if the adversary is allowed to make more than $n^{p/2}$ corruptions, the signal can be completely removed and the problem becomes information-theoretically unsolvable. Our theorem guarantees that if the number of corruptions is $o\big(n^{p/2}/\log n\big)$, we can find a vector highly correlated with v in polynomial time.

The dependence of λ on α is also likely to be optimal since we match (up to $\big(\log n\big)^{1/4}$ factor) the current best known guarantees for Gaussian noise with standard deviation $\Theta(1/\alpha)$.

We remark that some bound on the magnitude of the entries is needed[12] even if we do not allow adversarial corruptions. For example, if the vector v is 1-sparse (so it has one large entry), then the unbounded noise removes the information about v with probability $1-\alpha$. Indeed, if the noise entries are sampled from the mixture of the uniform distribution on $[-1, 1]$ with weight α and the Gaussian $N(0, 2^n)$ with weight $1-\alpha$, then with probability $1-\alpha$ the entry that corresponds to the support of v has vanishing small signal-to-noise ratio.

Evidence of the tightness of these requirements can also be found in the observation that, for $p = O(1)$ and arbitrarily small constant $\delta > 0$, it is unlikely that a $n^{1/2-\delta}$-sparse flat v can be recovered in polynomial time from the upper simplex of the input (i.e. the set of entries $\boldsymbol{Y}_{i_1 \dots i_p}$ such that $i_1 < \ldots < i_p$). Indeed the planted clique in random hypergraph problem can be reduced to this question (see Section 6.2). In other words, for certain vectors with $\|v\|_{\max} \leqslant n^{-1/4+\delta/2}$ the problem of recovering v from the upper simplex is likely to be computationally hard. It is not difficult to see that if we can use our SoS-based approach to recover k-sparse flat vectors from $\boldsymbol{Y}$, then we can also add additional sparsity constraints and get an SoS-based algorithm that recovers k-sparse flat vectors from the upper simplex of $\boldsymbol{Y}$ (if $p = O(1)$). This shows that the assumption on $\|v\|_{\max}$ in

[10] The case when p is even is similar.

[11] We also need to perform rounding to obtain the vector from the output tensor. See Appendix B.1 for more details.

[12] In fact, this is a recurring theme for unbounded noise models.

Theorem 1.3 is likely to be inherent, at least for our SoS-based approach. It remains a fascinating open question whether for specific noise distributions (e.g., Cauchy) the bound on $\|v\|_{\max}$ from Theorem 1.3 is tight.

Sparse principal component analysis. We consider the following sparse PCA model with symmetric noise.

Model 1.4 (Sparse PCA, single spike model)**.** Let $n, k \in \mathbb{N}$, $k \leqslant n$ and $0 < \alpha \leqslant 1$. Observe (an instance of) $\boldsymbol{Y} = \lambda \cdot vv^{\mathsf{T}} + \boldsymbol{N}$, where $\lambda > 0$, $v \in \mathbb{R}^n$ is an unknown k-sparse unit vector and $\boldsymbol{N}$ is a random n-by-n matrix with independent (but not necessarily identically distributed) symmetric about zero entries such that

$$\mathbb{P}[|\boldsymbol{N}_{ij}| \leqslant 1] \geqslant \alpha\,, \quad \text{for all } 1 \leqslant i, j \leqslant n\,.$$

When the noise is Gaussian this model is called the *spiked Wigner model* [FP07, JL09, DM16, DKWB19, dKNS20]. For Gaussian noise, when $\lambda > \sqrt{n}$ (this is called the strong signal regime) the leading eigenvector of $\boldsymbol{Y}$ correlates with the signal and thus a simple singular value decomposition provides optimal guarantees. In the weak signal regime –that is when $\lambda < \sqrt{n}$– polynomial time algorithms are known to recover the principal component v whenever $\lambda \gtrsim k\sqrt{\log(n/k^2)}$ [DM16, dKNS20]. In the sparse regime $k < n^{0.5-\delta}$ (for arbitrary constant $\delta > 0$), one can improve over these results in quasipolynomial time. Concretely, there exist algorithms [DKWB19, dKNS20, CdO21] that can recover the signal v in time $n^{O(t)}$ as long as $\lambda \gtrsim k\sqrt{\frac{\log n}{t}}$ for arbitrary $1 \leqslant t \leqslant k$. So for $t = \Theta(\log n)$ these algorithms can recover the signal in time $n^{O(\log n)}$ as long as $\lambda \geqslant k$. In the regime $k < n^{0.5-\delta}$ no $n^{o(\log n)}$ time algorithm is known to recover the signal if $\lambda \leqslant O(k)$, and there exist lower bounds (see [CdO21]) against restricted computational model of low degree polynomials, showing that in this model such algorithms do not exist.

In the context of spare PCA, Theorem 1.1 provides guarantees matching those of known quasipolynomial time algorithms, *but* also works with the heavy tailed noise of Model 1.4 (e.g., standard Cauchy noise):

Theorem 1.5 (Robust Sparse PCA)**.** *There exists an absolute constant $C > 1$ such that if $k \geqslant C \cdot \ln(n)/\alpha^2$, $\lambda \geqslant k$ and $\|v\|_{\max} \leqslant 100/\sqrt{k}$, then there exists an algorithm running in time $n^{O(\log(n)/\alpha^2)}$ that, given $\boldsymbol{Y}$ as in Model 1.4, outputs a unit vector $\hat{\boldsymbol{v}}$ satisfying*

$$|\langle v, \hat{\boldsymbol{v}}\rangle| \geqslant 0.99$$

with high probability.

Moreover, for $\varepsilon \leqslant \frac{\alpha^2 k^2}{Cn^2 \ln n}$, the same result holds if an arbitrary (adversarially chosen) ε-fraction of entries of $\boldsymbol{Y}$ is replaced by adversarially chosen values.

A natural question to ask concerning Theorem 1.5 is whether one could hope to obtain non-trivial guarantees in polynomial time. In Section 6.1 we provide evidence that the quasipolynomial time requirement for the noise model in Model 1.4 might be inherent (and thus the running time of Theorem 5.9 is nearly optimal) via a reduction from the Planted Clique problem. As in the context of tensor PCA, it is an interesting open question whether for specific heavy-tailed distributions (e.g., Cauchy) one can design polynomial time algorithms recovering the signal v (for not very large λ, say, $\lambda = k \operatorname{polylog} n$).

Finally, we remark that the number of adversarial corruptions that the algorithm can handle is nearly optimal: If the adversary that can change $\varepsilon = k^2/n^2$ fraction of the entries then all information about the signal may be removed.

Comparison with other results for symmetric unbounded noise models. Various other estimation problems in the presence of symmetric unbounded noise have been studied, such as linear regression, sparse regression and principal component analysis. We remark that our framework can be used to recover the best previously known results for these models [dNS21, dLN⁺21]. We point out however that compared to these algorithms, Theorem 1.1 provides a slow rate of error convergence. That is, when those algorithms guarantee an error bound $O(\varepsilon)$, Theorem 1.1 provides a bound $O(\sqrt{\varepsilon})$. This phenomenon is a consequence of the decomposability of particular regularizers used in previous works. Our framework does not require a decomposable regularizer and can thus deal with signal sets Ω that may be significantly more challenging than the ℓ_1-ball and nuclear norm ball considered in other works. We provide a more detailed discussion in Appendix D.

2 Techniques

Let $\Omega \subseteq \mathbb{R}^m$ be a set of structured signals we wish to recover (e.g., a sparse rank-1 matrix or a rank-1 tensor). Let $\boldsymbol{N}$ be an m-dimensional random noise with independent, symmetrically distributed entries such that $\min_{i\in[m]} \mathbb{P}\{|\boldsymbol{N}_i| \leqslant 1\} \geqslant \alpha$. Given (a realization of) a random vector $\boldsymbol{Y} = X^* + \boldsymbol{N}$ for some unknown signal $X^* \in \Omega$, our task is to approximately recover the signal X^*.

A common approach for this task is to minimize a loss function $L(X - \boldsymbol{Y})$ over $X \in \Omega$. In the special case of Gaussian noise, this approach recovers the maximum likelihood estimator if we choose the least-squares loss function $L(X - \boldsymbol{Y}) = \|X - \boldsymbol{Y}\|_2^2$. However, a well known weakness of this estimator is that it is extremely susceptible to outliers, thus it cannot be used with noise distributions with diverging moments. In contrast, an estimator that has been observed (both in practice and theory) to be significantly more robust to outliers is the *Huber loss function* $F_h(Z) := \sum_{i\in[m]} f_h(Z_i)$ where f_h is the following *Huber penalty*,

$$f_h(Z_i) := \begin{cases} \frac{1}{2} Z_i^2 & \text{for } |Z_i| \leqslant h\,, \\ h \cdot (|Z_i| - \frac{h}{2}) & \text{otherwise}\,. \end{cases} \tag{2.1}$$

Here, $h > 0$ is a parameter of the estimator to be determined later.

From a computational perspective, the problem is that for many (perhaps most) signal sets Ω one may be interested in, this kind of loss minimization turns out to be NP-hard (regardless of the concrete choice of the loss function). Therefore, we can only expect to solve specific relaxations of this optimization problem.

Previous work [dLN⁺21] considered these kinds of relaxations, but could only obtain meaningful error guarantees for sets Ω that admit convex regularizers with a certain decomposability property. Unfortunately, only few regularizers with this property are known (e.g., the ℓ_1-norm for vectors and the nuclear norm for matrices) and so this limitation turned out to be a fundamental obstacle to the application of this framework to many estimation problems.

Our machinery overcomes this obstacle, extending the approach in [dLN+21] to a significantly broader set of choices for Ω. Concretely, we can consider all choices of Ω such that a natural family of convex relaxations –namely the sum-of-squares hierarchy– succeeds in recovering the signal from Gaussian noise.

Tensor PCA as a running example. In order to illustrate our techniques, we consider the following example. Let $x \in \mathbb{R}^n$ be a unit vector and let $0 < \lambda \leqslant n^{3/2}$. For simplicity of the exposition, we assume here that x has entries from $\left\{\pm 1/\sqrt{n}\right\}$. We would like to recover a tensor $X^* = \lambda x^{\otimes 3}$ from $\boldsymbol{Y} = X^* + \boldsymbol{N}$, and determine how small λ can be so that the recovery of X^* is possible. Notice that in these settings the signal set is $\Omega = \{\lambda \cdot x^{\otimes 3} \mid x \in \left\{\pm 1/\sqrt{n}\right\}^n, \ \|x\| = 1\}$. This set is non-convex –in fact the problem is NP-hard in general– but let us temporarily disregard computational efficiency. Suppose we optimize the Huber loss with parameter $h = 3$ over this set Ω of rank-1 tensors.

Let $\hat{\boldsymbol{X}} \in \Omega$ be a minimizer, and denote $\boldsymbol{\Delta} = X^* - \hat{\boldsymbol{X}}$. A common approach is to apply Taylor's theorem and obtain

$$F_h(\boldsymbol{Y} - X^*) = F_h(\boldsymbol{N}) \geqslant F_h\left(\boldsymbol{Y} - \hat{\boldsymbol{X}}\right) \geqslant F_h(\boldsymbol{N}) + \langle \nabla F_h(\boldsymbol{N}), \boldsymbol{\Delta}\rangle + \frac{1}{2}\boldsymbol{\kappa}(\boldsymbol{\Delta}), \tag{2.2}$$

where $\boldsymbol{\kappa}(\boldsymbol{\Delta})$ is some lower bound on the values $\boldsymbol{\Delta}^\top H(X)\boldsymbol{\Delta}$ for all X from the segment between X^* and $\hat{\boldsymbol{X}}$, where $H(X)$ is the Hessian[13] of the Huber loss at point X. It is not hard to see (see Lemma 3.2) that one can choose

$$\boldsymbol{\kappa}(\boldsymbol{\Delta}) = \sum_{i=1}^{m} \mathbf{1}_{[|\boldsymbol{N}_i| \leqslant 1]} \cdot \mathbf{1}_{[|\boldsymbol{\Delta}_i| \leqslant h-1]} \boldsymbol{\Delta}_i^2 = \sum_{i=1}^{m} \mathbf{1}_{[|\boldsymbol{N}_i| \leqslant 1]} \boldsymbol{\Delta}_i^2 .$$

Now it is clear that if we can show $|\langle \nabla F_h(\boldsymbol{N}), \boldsymbol{\Delta}\rangle| \leqslant \gamma(\boldsymbol{\Delta})$ for some $\gamma(\boldsymbol{\Delta})$ and $\boldsymbol{\kappa}(\boldsymbol{\Delta}) \geqslant 0.9 \cdot \alpha \cdot \|\boldsymbol{\Delta}\|_2^2$, Eq. (2.2) immediately implies the bound

$$\|\boldsymbol{\Delta}\|_2^2 < 3\gamma(\boldsymbol{\Delta})/\alpha . \tag{2.3}$$

That is, the estimator guarantee depends only on an *upper bound on the gradient* and a *lower bound on the curvature* of the space in the direction of $\boldsymbol{\Delta}$.

Let us first obtain the bound

$$\boldsymbol{\kappa}(\boldsymbol{\Delta}) = \sum_{i=1}^{m} \mathbf{1}_{[|\boldsymbol{N}_i| \leqslant 1]} \boldsymbol{\Delta}_i^2 \geqslant 0.9 \cdot \alpha \cdot \|\Delta\|_2^2 .$$

For simplicity assume $\|\boldsymbol{\Delta}\|_2 = \tau$ for some[14] $\tau \geqslant n^{-O(1)}$. A successful strategy here is to derive a lower bound on $\boldsymbol{\kappa}(\Delta)$ for a fixed Δ, and then construct an ε-net over $\Omega' = \{X - X' : X, X' \in \Omega\,, \|X - X'\|_2 = \tau\}$. The idea is that *if* our lower bound holds with sufficiently large probability and *if* the size of the covering is not too large, then we will be able to show the desired curvature in all the possible directions of $\boldsymbol{\Delta}$. Now for fixed Δ, the expected value of $\boldsymbol{\kappa}(\Delta)$ is $\alpha\|\Delta\|_2^2$,

[13] The second derivative of the Huber penalty does not exit at the points $\{\pm h\}$. However, the indicator function I_h of the interval $[-h, h]$ is the second derivative of Huber penalty in L_1 sense, that is $f_h'[b] - f_h'[a] = \int_a^b I_h(t)\,\mathrm{d}t$ for all $a, b \in \mathbb{R}$. So by the Hessian at point X we mean a quadratic form whose matrix in the standard basis is diagonal with (diagonal) entries $I_h(X_i)$.

[14] Estimation error τ cannot be $n^{-\omega(1)}$ in our parameter regime.

and by Hoeffding's inequality, the deviation from the mean is bounded by $O\left(\|\Delta\|_4^2\sqrt{\log(1/\delta)}\right)$ with probability at least $1-\delta$. Since $\|\Delta\|_{\max} \leqslant 2$, $\|\Delta\|_4^2 \leqslant 2\|\Delta\|_2$. Thus we have

$$\boldsymbol{\kappa}(\Delta) \geqslant \alpha\|\Delta\|_2^2 - \|\Delta\|_2 \cdot O\left(\sqrt{\log(1/\delta)}\right) = \tau\left(\alpha\tau - O\left(\sqrt{\log(1/\delta)}\right)\right),$$

which is close to its expectation when $\tau \gtrsim \sqrt{\log 1/\delta}/\alpha$.

We need now to extend this bound to all possible directions of $\boldsymbol{\Delta}$. To this end note that if $\Delta, \Delta' \in \Omega'$ are ε-close to each other for some small enough $\varepsilon = n^{-O(1)}$, then[15]

$$|\boldsymbol{\kappa}(\Delta) - \boldsymbol{\kappa}(\Delta')| \lesssim \alpha\tau^2 .$$

So it remains to show a cover of Ω. Notice that the size of the cover determines in a very strong way the quality of the error guarantees of the estimator. For example one could try to use the ε-net covering the unit ball in $\mathbb{R}^{n^3}$, this does not exploit the structure of Ω and has thus size $(O(1/\varepsilon))^{n^3}$. By the above calculations, with this ε-net we could provide a meaningful lower bound only when $\tau \gtrsim n^{3/2}/\alpha$. In other words, our error estimate would be worse than the trivial estimator outputting the zero tensor! To obtain a tighter covering, recall Ω is a subset of the set of rank one tensors of norm $\lambda \leqslant n^{3/2}$. The size of minimal ε-net in Ω is at most $O\left(n^{O(1)}/\varepsilon\right)^n$ (since the mapping $x \mapsto x^{\otimes 3}$ is $n^{O(1)}$-Lipschitz for $\|x\|_2 \leqslant n^{O(1)}$). Hence the size of the ε-net in Ω' is bounded by $n^{O(n)}$, and by union bound we get

$$\boldsymbol{\kappa}(\boldsymbol{\Delta}) \geqslant \alpha\|\boldsymbol{\Delta}\|_2^2 - \|\boldsymbol{\Delta}\|_2 \cdot O\left(\sqrt{n\log n}\right) = \alpha\tau^2 - O\left(\tau\sqrt{n\log n}\right),$$

with high probability. Hence for $\tau \gtrsim \frac{\sqrt{n\log n}}{\alpha}$ we get the desired bound.

We can now focus on bounding the gradient $|\langle \nabla F_h(\boldsymbol{N}), \boldsymbol{\Delta}\rangle|$. The choice of the Huber loss function makes this very easy: $\nabla F_h(\boldsymbol{N})$ is a random vector with symmetric independent entries bounded by $h = O(1)$ in absolute value, so for fixed Δ, $|\langle \nabla F_h(\boldsymbol{N}), \Delta\rangle|$ is bounded by $O\left(\|\Delta\|_2\sqrt{\log(1/\delta)}\right)$ with probability $1-\delta$. By union bound over the ε-net in Ω', with high probability

$$|\langle \nabla F_h(\boldsymbol{N}), \boldsymbol{\Delta}\rangle| \leqslant \|\boldsymbol{\Delta}\|_2 \cdot O\left(\sqrt{n\log n}\right).$$

Hence by Eq. (2.3), we can conclude that with high probability

$$\|\boldsymbol{\Delta}\|_2 \leqslant O\left(\frac{\sqrt{n\log n}}{\alpha}\right).$$

Therefore, the minimizer $\hat{\boldsymbol{X}}$ of this inefficient estimator is highly correlated with X^* as long as $\lambda \gtrsim \sqrt{n\log n}/\alpha$. This bound is nearly optimal: if $\boldsymbol{N}$ has iid Gaussian entries with standard deviation $\Theta(\alpha)$, it is information-theoretically impossible to recover X^* if $\lambda \leqslant o\left(\sqrt{n}/\alpha\right)$ (see [PWB20]).

[15] Here we assume that $\alpha > 1/n$, otherwise the problem is information theoretically intractable.

Tensor PCA as a running example: efficient estimation. We take now into account the computational complexity of computing the desired estimator. To have a loss function we can minimize efficiently, the idea is to replace the set of rank-1 tensors Ω by some set $\tilde{\Omega} \supset \Omega$ over which we can efficiently optimize. We cannot do this via the framework in [dLN+21] since no appropriate decomposable regularizer is known for high-order tensors. Thus we use instead sum-of-squares relaxations, and take $\tilde{\Omega} = \tilde{\Omega}_t$ to be the set of pseudo-expectations of degree t that satisfy certain constraints. Crucially, in order to apply the argument of the previous paragraph, we need a tight upper bound on the covering number of the set of pseudo-expectations $\tilde{\Omega}_t$.

In the exponential time algorithm described above we had a natural $n^{O(1)}$-Lipschitz mapping from n-dimensional space to n^3-dimensional space, which allowed us to construct such a covering. In the case of pseudo-expectations, we do not have such a mapping, so different techniques are required to to get a bound on the size of ε-net.

We use Sudakov minoration: For every bounded set $A \subset \mathbb{R}^m$, the size of the minimal ε-net of A is bounded by $\exp\left(O\left(\mathcal{G}(A)^2/\varepsilon^2\right)\right)$, where

$$\mathcal{G}(A) = \mathop{\mathbb{E}}_{w \sim N(0,\mathrm{Id}_m)}\left[\sup_{a \in A} \sum_{i=1}^{m} a_i \boldsymbol{w}_i\right].$$

The quantity $\mathcal{G}(A)$ is called the *Gaussian complexity* of the set A. So in order to bound the size of optimal ε-net of the set $\tilde{\Omega}$ of pseudo-expectations it is enough to bound its Gaussian complexity. The good news is that we can bound the Gaussian complexity of the set of pseudo-expectations by certifying in sum-of-squares a bound on the Gaussian complexity of the set Ω of rank-1 tensors! Concretely, Ω can be defined by polynomial constraints with variables $X \in \mathbb{R}^{n^3}$ and auxiliary variables $x \in \mathbb{R}^n$:

$$\mathcal{S}_{X,x} = \left\{X = \lambda x^{\otimes 3}, \quad \|x\|_2^2 = 1, \quad \forall i \in [n], \;\; x_i^2 \leqslant 1/n\right\}.$$

If we can show that with high probability[16] over the tensors $\boldsymbol{W}$ with iid Gaussian entries there exists a degree t sum-of-square proof that these constraints imply

$$\sum_{1 \leqslant i \leqslant j \leqslant k \leqslant n} x_i x_j x_k \boldsymbol{W}_{ijk} \leqslant \gamma_t\,,$$

then we can conclude that $\mathcal{G}\left(\tilde{\Omega}_t\right) \leqslant O\left(\lambda\gamma_t\right)$.

In [HSS15, HKP+17] it was shown that there exists a 4-degree sum-of-squares proof that $\mathcal{S}_{X,x}$ imply the inequality

$$\sum_{1 \leqslant i \leqslant j \leqslant k \leqslant n} x_i x_j x_k \boldsymbol{W}_{ijk} \leqslant O(\ln(n))^{1/4} \cdot n^{3/4}\,.$$

Hence, $\mathcal{G}\left(\tilde{\Omega}_4\right) \leqslant \tilde{O}\left(\lambda n^{3/4}\right)$ as desired.

Note that the analysis of the exponential time algorithm does not work here because the dependence of the size of ε-net on ε in Sudakov's minoration is exponential and not polynomial as in the case of ℓ_2-ball. However, it turns out that via a more careful analysis we can show

$$\|\boldsymbol{\Delta}\|_2^2 \leqslant \tilde{O}\left(\frac{\lambda n^{3/4}}{\alpha}\right).$$

[16] For Gaussian distribution it is not hard to obtain from this a bound on expectation since we have good tail bounds for it.

This bound implies that $\hat{\boldsymbol{X}}$ is highly correlated with X^* as long as $\lambda \gtrsim (\log n)^{1/4} n^{3/4}/\alpha$, which matches (up to a logarithmic factor) the current best known guarantees for polynomial time algorithms when $\boldsymbol{N}$ has i.i.d. Gaussian entries with standard deviation $\Theta(\alpha)$, but also works with significantly more general noise (e.g., Cauchy noise at scale $\Theta(\alpha)$).

Recovery in the presence of adversarial corruptions and oblivious noise. Our framework is robust to additional adversarial corruptions resulting from an adversary corrupting an ε-fraction of the entries of $\boldsymbol{Y}$. In light of our discussion so far, to show this it suffices to check how do the values $|\langle \nabla F_h(\boldsymbol{N}), \boldsymbol{\Delta}\rangle|$ and $\boldsymbol{\kappa}(\boldsymbol{\Delta}) = \sum_{i=1}^{n} \mathbf{1}_{[|N_i|\leqslant 1]} \boldsymbol{\Delta}_i^2$ change in the presence of corruptions. For simplicity, we limit our discussion to the first inefficient estimator introduced in previous paragraphs.

First assume that the adversary corrupts a set of entries of size εn^3 that is random (not adversarially chosen). In this case $|\langle \nabla F_h(\boldsymbol{N}), \boldsymbol{\Delta}\rangle|$ can only be increased by an additive term

$$h \cdot 2\lambda \|v\|_{\max}^3 \cdot n^3 \varepsilon \leqslant O\left(n^{3/2} \lambda \varepsilon\right),$$

since the entries of $\nabla F_h(\boldsymbol{N})$ are bounded by h, and the entries of $\boldsymbol{\Delta}$ are bounded by $2\lambda\|v\|_{\max}^3$. The value $\boldsymbol{\kappa}(\boldsymbol{\Delta})$ also does not change significantly if a small random set of entries is corrupted. Hence in this case, if $n^{3/2}\lambda\varepsilon \leqslant \|\boldsymbol{\Delta}\|_2 \sqrt{n \log n}$, the error does not increase in any significant way. Note that in the regime $\|\boldsymbol{\Delta}\|_2 \geqslant \Omega(\lambda)$ (when we can still have 0.99 correlation with the signal), the number of corruptions εn^3 is allowed to be up to n^2.

In the general case, when the adversary is allowed to choose the corrupted set, we need to use a union bound over all sets of size εn^3 (we use it to bound both $|\langle \nabla F_h(\boldsymbol{N}), \boldsymbol{\Delta}\rangle|$ and $\boldsymbol{\kappa}(\boldsymbol{\Delta})$). Here, the gradient bound becomes

$$|\langle \nabla F_h(\boldsymbol{N}), \boldsymbol{\Delta}\rangle| \leqslant \|\Delta\|_2 \cdot O\left(\sqrt{n \log n} + \sqrt{\varepsilon n^3 \log n}\right).$$

Hence the number of corruptions is only allowed to be at most n. Observe that Theorem 1.3 is robust up to $\tilde{\Omega}(n^{3/2})$ corrupted entries. This is not surprising. The reason is that the algorithm requires signal strength $\lambda = \tilde{\Omega}(n^{3/4})$ compared to $\lambda = \tilde{\Omega}(n^{1/2})$ that is required by the exponential time algorithm.

Sparse PCA. As a second example of the applications of Theorem 1.1 consider the sparse PCA problem: We are given $\boldsymbol{Y} = \lambda \cdot vv^\top + \boldsymbol{N}$, where $v \in \mathbb{R}^n$ is a k-sparse vector, and the goal is to recover v. For simplicity we assume here that v is flat, i.e., that its non-zero entries are in $\{\pm 1/\sqrt{k}\}$.

In order to use our framework, we need to certify in sum-of-squares an upper bound on the Gaussian complexity of the set of sparse vectors. So we need to show that for some (as small as possible) γ, with high probability over matrices $\boldsymbol{W}$ with i.i.d. Gaussian entries there exists a (not very high degree) sum-of-squares proof that some system of constraints $\mathcal{C}$ defining sparse vectors implies

$$\sum_{1\leqslant i,j\leqslant n} x_i x_j \boldsymbol{W}_{ij} \leqslant \gamma,$$

where x are variables that satisfy sparsity constraints of $\mathcal{C}$.

We use the system of constraints $\mathcal{C}_t$ (the subscript $t \in \mathbb{N}$ indicates that the constraints involve degree t polynomials) from [dKNS20] (see Section 5.2 for a precise definition). The authors in [dKNS20] used the program for a different sparse PCA model, but it is possible to adapt their proof and show that with high probability there exist a degree $O(t)$ sum-of-squares proof that C_t implies the inequality

$$\sum_{1\leqslant i,j\leqslant n} x_i x_j \boldsymbol{W}_{ij} \leqslant O\left(k\sqrt{\frac{\log n}{t}}\right).$$

Hence if $\lambda = k$ and $t \gtrsim \log(n)/\alpha^2$, then Theorem 1.1 implies that the Huber loss minimizer has 0.999 correlation with $vv^\top$ (and hence its top eigenvector has correlation 0.99 with v or $-v$).

The running time is $n^{O(t)} = n^{O(\log(n)/\alpha^2)}$, and it is likely to be inherent: For $\alpha = 1$ we can reduce the planted clique problem (with clique size k) to the problem of recovering v from the upper triangle (without the diagonal) of matrix $\boldsymbol{Y}$. The best currently known algorithmic guarantees for sparse PCA are captured by algorithms that can recover v from the upper triangle of the input matrix, hence it is likely that sparse PCA with symmetric noise is at least as hard as the planted clique problem. Finally, we remark that there is a conjecture stating that there is no $n^{o(\log n)}$-time algorithm that can solve the planted clique problem for some values of k (see [MRS21]).

The reduction works as follows. We use the notation $\mathcal{U}(M)$ to denote the upper triangle of matrix M. It is not hard to see that if $\boldsymbol{A}$ is an instance of the planted clique problem (the adjacency matrix of the graph) and J is the matrix with all entries equal to one, then $\mathcal{U}(2\boldsymbol{A} - J)$ is the upper triangle of an instance of the sparse PCA problem with symmetric noise, where $\lambda = k$, $\sqrt{k}\cdot v$ is the 0/1 indicator of the clique, and the noise $\boldsymbol{N}$ is as follows: For the entries $i, j \in \operatorname{supp}(v)$, $\boldsymbol{N}_{ij} = 0$, and for other entries $\boldsymbol{N}_{ij}$ are iid sampled from the uniform distribution on $\{\pm 1\}$.

3 Preliminaries

Notation. We use boldface to denote random variables. We hide absolute constant multiplicative factors using the standard notations $O(\cdot), \Omega(\cdot), \gtrsim, \lesssim$. Similarly, we hide multiplicative logarithmic (in the dimension m of the input) factors using the notation $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$. We use the notation $\|\cdot\|_2$ for the Euclidean norm, $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm, $\|X\|_{\max} = \max_{i\in[m]}|X_i|$. We write log for the logarithm to the base e.

Definition 3.1 (Huber loss function)**.** The Huber loss penalty is defined as:

$$f_h(t) := \begin{cases} \frac{1}{2}t^2 & \text{for } |t| \leqslant h\,, \\ h(|t| - \frac{h}{2}) & \text{otherwise.} \end{cases} \tag{3.1}$$

For a vector $x \in \mathbb{R}^n$ we denote by $F_h(x) := \sum_{i\in[n]} f_h(x_i)$.

The Huber loss satisfies the following inequality.

Lemma 3.2. *Let $h > 0$. For all $t, \delta \in \mathbb{R}$, and all $0 \leqslant \zeta \leqslant h$,*

$$f_h(t+\delta) - f_h(t) - f_h'(t)\cdot\delta \geqslant \frac{\delta^2}{2}\mathbf{1}_{[|t|\leqslant\zeta]}\cdot\mathbf{1}_{[|\delta|\leqslant h-\zeta]}\,. \tag{3.2}$$

Proof. We have two cases:

- If $|t| > \zeta$ or $|\delta| > h - \zeta$, then either $\mathbf{1}_{[|t| \leqslant \zeta]} = 0$ or $\mathbf{1}_{[|\delta| \leqslant h - \zeta]} = 0$. Hence,

$$\mathbf{1}_{[|t| \leqslant \zeta]} \cdot \mathbf{1}_{[|\delta| \leqslant h - \zeta]} = 0\,. \tag{3.3}$$

  For this case, we simply use the convexity of f_h to get

$$\begin{aligned} f_h(t + \delta) - f_h(t) - f_h'(t) \cdot \delta &\geqslant 0 & (3.4)\\ &= \frac{\delta^2}{2} \mathbf{1}_{[|t| \leqslant \zeta]} \cdot \mathbf{1}_{[|\delta| \leqslant h - \zeta]}\,. & (3.5) \end{aligned}$$

- If $|t| \leqslant \zeta$ and $|\delta| \leqslant h - \zeta$, then $|t + \delta| \leqslant h$. In this case, we have $f_h(t + \delta) = \frac{1}{2}(t + \delta)^2$, $f_h(t) = \frac{1}{2}t^2$ and $f_h'(t) = t$. By direct inspection, we get

$$\begin{aligned} f_h(t + \delta) - f_h(t) - f_h'(t) \cdot \delta &= \frac{1}{2}(t + \delta)^2 - \frac{1}{2}t^2 - t\delta = \frac{\delta^2}{2} & (3.6)\\ &= \frac{\delta^2}{2} \mathbf{1}_{[|t| \leqslant \zeta]} \cdot \mathbf{1}_{[|\delta| \leqslant h - \zeta]}\,. & (3.7) \end{aligned}$$

□

3.1 Sum of squares and pseudodistributions

Let $x = (x_1, x_2, \ldots, x_n)$ be a tuple of n indeterminates and let $\mathbb{R}[x]$ be the set of polynomials with real coefficients and indeterminates $x_1, \ldots, x_n$. We say that a polynomial $p \in \mathbb{R}[x]$ is a *sum-of-squares (sos)* if there are polynomials $q_1, \ldots, q_r$ such that $p = q_1^2 + \cdots + q_r^2$.

3.2 Pseudo-distributions

Pseudo-distributions are generalizations of probability distributions. We can represent a discrete (i.e., finitely supported) probability distribution over $\mathbb{R}^n$ by its probability mass function $D \colon \mathbb{R}^n \to \mathbb{R}$ such that $D \geqslant 0$ and $\sum_{x \in \mathrm{supp}(D)} D(x) = 1$. Similarly, we can describe a pseudo-distribution by its mass function. Here, we relax the constraint $D \geqslant 0$ and only require that D passes certain low-degree non-negativity tests.

Concretely, a *level-ℓ pseudo-distribution* is a finitely-supported function $D : \mathbb{R}^n \to \mathbb{R}$ such that $\sum_x D(x) = 1$ and $\sum_x D(x) f(x)^2 \geqslant 0$ for every polynomial f of degree at most $\ell/2$. (Here, the summations are over the support of D.) A straightforward polynomial-interpolation argument shows that every level-∞-pseudo distribution satisfies $D \geqslant 0$ and is thus an actual probability distribution. We define the *pseudo-expectation* of a function f on $\mathbb{R}^d$ with respect to a pseudo-distribution D, denoted $\tilde{\mathbb{E}}_{D(x)} f(x)$, as

$$\tilde{\mathbb{E}}_{D(x)} f(x) = \sum_x D(x) f(x)\ . \tag{3.8}$$

The degree-ℓ moment tensor of a pseudo-distribution D is the tensor $\mathbb{E}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes \ell}$. In particular, the moment tensor has an entry corresponding to the pseudo-expectation of all monomials of degree at most ℓ in x. The set of all degree-ℓ moment tensors of probability distribution is a convex set. Similarly, the set of all degree-ℓ moment tensors of degree d pseudo-distributions is also convex. Key to the algorithmic utility of pseudo-distributions is the fact that while there can be no efficient separation oracle for the convex set of all degree-ℓ moment tensors of an actual probability distribution, there's a separation oracle running in time $n^{O(\ell)}$ for the convex set of the degree-ℓ moment tensors of all level-ℓ pseudodistributions.

Fact 3.3 ([Sho87, Par00, Nes00, Las01]). *For any $n, \ell \in \mathbb{N}$, the following set has a $n^{O(\ell)}$-time weak separation oracle (in the sense of [GLS81]):*

$$\left\{ \tilde{\mathbb{E}}_{D(x)} (1, x_1, x_2, \ldots, x_n)^{\otimes d} \mid \textit{degree-d pseudo-distribution } D \textit{ over } \mathbb{R}^n \right\} . \tag{3.9}$$

This fact, together with the equivalence of weak separation and optimization [GLS81] allows us to efficiently optimize over pseudo-distributions (approximately)—this algorithm is referred to as the sum-of-squares algorithm.

The *level-ℓ sum-of-squares algorithm* optimizes over the space of all level-ℓ pseudo-distributions that satisfy a given set of polynomial constraints—we formally define this next.

Definition 3.4 (Constrained pseudo-distributions)**.** Let D be a level-ℓ pseudo-distribution over $\mathbb{R}^n$. Let $\mathcal{A} = \{f_1 \geqslant 0, f_2 \geqslant 0, \ldots, f_m \geqslant 0\}$ be a system of m polynomial inequality constraints. We say that D *satisfies the system of constraints $\mathcal{A}$ at degree r*, denoted $D \underset{r}{\models} \mathcal{A}$, if for every $S \subseteq [m]$ and every sum-of-squares polynomial h with $\deg h + \sum_{i \in S} \max\{\deg f_i, r\} \leqslant \ell$,

$$\tilde{\mathbb{E}}_D h \cdot \prod_{i \in S} f_i \geqslant 0 \, .$$

We write $D \models \mathcal{A}$ (without specifying the degree) if $D \underset{0}{\models} \mathcal{A}$ holds. Furthermore, we say that $D \underset{r}{\models} \mathcal{A}$ holds *approximately* if the above inequalities are satisfied up to an error of $2^{-n^\ell} \cdot \|h\| \cdot \prod_{i \in S} \|f_i\|$, where $\|\cdot\|$ denotes the Euclidean norm[17] of the cofficients of a polynomial in the monomial basis.

We remark that if D is an actual (discrete) probability distribution, then we have $D \models \mathcal{A}$ if and only if D is supported on solutions to the constraints $\mathcal{A}$.

We say that a system $\mathcal{A}$ of polynomial constraints is *explicitly bounded* if it contains a constraint of the form $\{\|x\|^2 \leqslant M\}$. The following fact is a consequence of Fact 3.3 and [GLS81],

Fact 3.5 (Efficient Optimization over Pseudo-distributions)**.** *There exists an $(n+m)^{O(\ell)}$-time algorithm that, given any explicitly bounded and satisfiable system[18] $\mathcal{A}$ of m polynomial constraints in n variables, outputs a level-ℓ pseudo-distribution that satisfies $\mathcal{A}$ approximately.*

3.3 Sum-of-squares proofs

Let $f_1, f_2, \ldots, f_r$ and g be multivariate polynomials in x. A *sum-of-squares proof* that the constraints $\{f_1 \geqslant 0, \ldots, f_m \geqslant 0\}$ imply the constraint $\{g \geqslant 0\}$ consists of sum-of-squares polynomials $(p_S)_{S \subseteq [m]}$ such that

$$g = \sum_{S \subseteq [m]} p_S \cdot \prod_{i \in S} f_i \, . \tag{3.10}$$

[17] The choice of norm is not important here because the factor 2^{-n^ℓ} swamps the effects of choosing another norm.

[18] Here, we assume that the bitcomplexity of the constraints in $\mathcal{A}$ is $(n+m)^{O(1)}$.

We say that this proof has *degree* ℓ if for every set $S \subseteq [m]$, the polynomial $p_S \prod_{i \in S} f_i$ has degree at most ℓ. If there is a degree ℓ SoS proof that $\{f_i \geqslant 0 \mid i \leqslant r\}$ implies $\{g \geqslant 0\}$, we write:

$$\{f_i \geqslant 0 \mid i \leqslant r\} \vdash_{\ell} \{g \geqslant 0\}\,. \tag{3.11}$$

Sum-of-squares proofs satisfy the following inference rules. For all polynomials $f, g\colon \mathbb{R}^n \to \mathbb{R}$ and for all functions $F\colon \mathbb{R}^n \to \mathbb{R}^m$, $G\colon \mathbb{R}^n \to \mathbb{R}^k$, $H\colon \mathbb{R}^p \to \mathbb{R}^n$ such that each of the coordinates of the outputs are polynomials of the inputs, we have:

$$\frac{\mathcal{A} \vdash_{\ell} \{f \geqslant 0, g \geqslant 0\}}{\mathcal{A} \vdash_{\ell} \{f + g \geqslant 0\}}, \frac{\mathcal{A} \vdash_{\ell} \{f \geqslant 0\}, \mathcal{A} \vdash_{\ell'} \{g \geqslant 0\}}{\mathcal{A} \vdash_{\ell+\ell'} \{f \cdot g \geqslant 0\}}, \tag{addition and multiplication}$$

$$\frac{\mathcal{A} \vdash_{\ell} \mathcal{B}, \mathcal{B} \vdash_{\ell'} C}{\mathcal{A} \vdash_{\ell \cdot \ell'} C}, \tag{transitivity}$$

$$\frac{\{F \geqslant 0\} \vdash_{\ell} \{G \geqslant 0\}}{\{F(H) \geqslant 0\} \vdash_{\ell \cdot \deg(H)} \{G(H) \geqslant 0\}}. \tag{substitution}$$

Low-degree sum-of-squares proofs are sound and complete if we take low-level pseudodistributions as models.

Concretely, sum-of-squares proofs allow us to deduce properties of pseudo-distributions that satisfy some constraints.

Fact 3.6 (Soundness). *If $D \models_r \mathcal{A}$ for a level-ℓ pseudo-distribution D and there exists a sum-of-squares proof $\mathcal{A} \vdash_{r'} \mathcal{B}$, then $D \models_{r \cdot r' + r'} \mathcal{B}$.*

If the pseudo-distribution D satisfies $\mathcal{A}$ only approximately, soundness continues to hold if we require an upper bound on the bit-complexity of the sum-of-squares $\mathcal{A} \vdash_{r'} B$ (number of bits required to write down the proof).

In our applications, the bit complexity of all sum of squares proofs will be $n^{O(\ell)}$ (assuming that all numbers in the input have bit complexity $n^{O(1)}$). This bound suffices in order to argue about pseudo-distributions that satisfy polynomial constraints approximately.

The following fact shows that every property of low-level pseudo-distributions can be derived by low-degree sum-of-squares proofs.

Fact 3.7 (Completeness). *Suppose $d \geqslant r' \geqslant r$ and $\mathcal{A}$ is a collection of polynomial constraints with degree at most r, and $\mathcal{A} \vdash \{\sum_{i=1}^n x_i^2 \leqslant B\}$ for some finite B.*

Let $\{g \geqslant 0\}$ be a polynomial constraint. If every degree-d pseudo-distribution that satisfies $D \models_r \mathcal{A}$ also satisfies $D \models_{r'} \{g \geqslant 0\}$, then for every $\varepsilon > 0$, there is a sum-of-squares proof $\mathcal{A} \vdash_d \{g \geqslant -\varepsilon\}$.

We will repeatedly use the following SoS version of Cauchy-Schwarz inequality and its generalization, HÃűlder's inequality:

Fact 3.8 (Sum-of-Squares Cauchy-Schwarz). *Let $x, y \in \mathbb{R}^d$ be indeterminites. Then,*

$$\vdash_4^{x,y} \left\{ \left(\sum_i x_i y_i \right)^2 \leqslant \left(\sum_i x_i^2 \right) \left(\sum_i y_i^2 \right) \right\}.$$

We will also use the following facts about triangle inequalities and spectral certificates within the SoS proof system.

Lemma 3.9. *There is a degree-2 sum-of-squares proof of the following weak triangle inequality:*

$$2\left(\sum_{i=1}^{n} a_i^2\right) + 2\left(\sum_{i=1}^{n} b_i^2\right) - \left(\sum_{i=1}^{n}(a_i + b_i)^2\right) = \sum_{i=1}^{n}(a_i - b_i)^2 \,. \tag{3.12}$$

Fact 3.10 (Spectral Certificates)**.** *For any $m \times m$ matrix A,*

$$\Big|\frac{u}{2}\ \Big\{\langle u, Au\rangle \leqslant \|A\| \cdot \|u\|_2^2\Big\}\,.$$

We will also use the following results about pseudo-distributions.

Fact 3.11 (Cauchy-Schwarz for Pseudo-distributions)**.** *Let f, g be polynomials of degree at most d in indeterminate $x \in \mathbb{R}^d$. Then, for any degree d pseudo-distribution D, $\tilde{\mathbb{E}}_D[fg] \leqslant \sqrt{\tilde{\mathbb{E}}_D[f^2]}\sqrt{\tilde{\mathbb{E}}_D[g^2]}$.*

4 Meta-theorem

In this section we prove Theorem 1.1.

Theorem 4.1 (Meta-theorem)**.** *Let $\delta, \alpha \in (0,1)$ and $\zeta \geqslant 0$. Let $\tilde{\Omega} \subseteq \mathbb{R}^m$ be a compact convex set. Let $b, r, \gamma \in \mathbb{R}$ be such that*

$$\max_{X\in\tilde{\Omega}} \|X\|_{\max} \leqslant b\,,$$

$$\max_{X\in\tilde{\Omega}} \|X\|_2 \leqslant r\,,$$

and

$$\mathop{\mathbb{E}}_{\boldsymbol{W}\sim N(0,\mathrm{Id})}\left[\sup_{X\in\tilde{\Omega}} \langle X, \boldsymbol{W}\rangle\right] \leqslant \gamma\,.$$

Consider

$$\boldsymbol{Y} = X^* + \boldsymbol{N}\,,$$

where $X^* \in \tilde{\Omega}$ and $\boldsymbol{N}$ is a random m-dimensional vector with independent (but not necessarily identically distributed) symmetric about zero entries satisfying $\mathbb{P}[|\boldsymbol{N}_i| \leqslant \zeta] \geqslant \alpha$.

Let

$$\varepsilon = \frac{\gamma^2}{r^2 m \log m}\,,$$

and let Z be an m-dimensional vector such that at least $(1-\varepsilon)m$ entries of Z coincide with entries of $\boldsymbol{Y}$, and other entries are arbitrary.

Then the minimizer $\hat{X} = \operatorname{argmin}_{X\in\tilde{\Omega}} F_h(Z - X)$ of the Huber loss with parameter $h \geqslant 2b + \zeta$ satisfies

$$\left\|\hat{X} - X^*\right\|_2 \leqslant O\left(\sqrt{\frac{h}{\alpha}\left(\gamma + r\sqrt{\log(1/\delta)}\right)}\right)$$

with probability at least $1-\delta$ over the randomness of $\boldsymbol{N}$.

Note that without loss of generality we can assume that $h \cdot \varepsilon \cdot m \log m \leqslant \gamma$. Indeed, otherwise we would get $h \cdot \gamma > r^2$, and the error bound becomes trivial. Similarly, we can assume that $h \log(1/\delta) \leqslant r\sqrt{\log(1/\delta)}$.

To prove the theorem we need the next two intermediate lemmas.

Lemma 4.2 (Gradient bound)**.** *Consider the settings of Theorem 4.1. Then with probability at least* $1-\delta$ *over the randomness of* $\mathbf{N}$*, for every* $X, X' \in \tilde{\Omega}$*, we have*

$$|\langle \nabla F_h(\mathbf{N} + Z - \boldsymbol{Y}), X - X' \rangle| \leqslant 100h \cdot \left(\gamma + r\sqrt{\log(1/\delta)}\right).$$

Proof. Let C be the set of entries where $\boldsymbol{Y}$ differs from Z. The size of C is at most εm, hence

$$\left|\sum_{i \in C} f_h'(\mathbf{N}_i + Z_i - \boldsymbol{Y}_i) \cdot (X_i - X_i')\right| \leqslant h \cdot 2b \cdot \varepsilon m \leqslant h^2 \cdot \varepsilon m \leqslant h\gamma\,. \tag{4.1}$$

Now consider some fixed (non-random) subset S of entries of size $(1-\varepsilon)m$. By Lemma C.4, the random variable $\sup_{X,X' \in \tilde{\Omega}} \sum_{i \in S} f_h'(\mathbf{N}_i) \cdot \left(X_i - X_i'\right)$ has expectation bounded by $6 \cdot h \cdot \gamma$, and for every $0 < \delta' < 1$, we get that with probability at least $1 - \delta'$,

$$\left|\sum_{i \in S} f_h'(\mathbf{N}_i) \cdot (X_i - X_i')\right| \leqslant h \cdot \left(6\gamma + 10r\sqrt{\log(1/\delta')}\right).$$

Now choose

$$\delta' := \frac{\delta}{\binom{m}{\varepsilon m}} \geqslant \frac{\delta}{\left(\frac{e}{\varepsilon}\right)^{\varepsilon m}}.$$

By taking a union bound over all subsets[19] of size $(1-\varepsilon)m$, we can see that with probability at least $1-\delta$, we have

$$\begin{aligned}\left|\sum_{i \in S} f_h'(\mathbf{N}_i) \cdot (X_i - X_i')\right| &\leqslant h \cdot \left(6\gamma + 10r\sqrt{\log(1/\delta')}\right)\\ &\leqslant h \cdot \left(6\gamma + 10r\sqrt{\log(1/\delta)} + 10r\sqrt{\varepsilon m \log(e/\varepsilon)}\right)\end{aligned}$$

for all subsets of size $(1-\varepsilon)m$. In particular, for $S = [m] \setminus C$, with probability at least $1-\delta$, we have

$$\left|\sum_{i \in [m] \setminus C} f_h'(\mathbf{N}_i) \cdot (X_i - X_i')\right| \leqslant h \cdot \left(6\gamma + 10r\sqrt{\log(1/\delta)} + 10r\sqrt{\varepsilon m \log(e/\varepsilon)}\right). \tag{4.2}$$

Now notice that

$$\begin{aligned} r^2 \varepsilon m \log(e/\varepsilon) &= r^2 \cdot \frac{\gamma^2}{r^2 m \log m} \cdot m\left(1 + \log\frac{r^2 m \log m}{\gamma^2}\right)\\ &= \frac{\gamma^2}{\log m}\left(1 + \log m + \log\log m + \log\frac{r^2}{\gamma^2}\right) \leqslant 4\gamma^2\,. \end{aligned}$$

[19] The number of such subsets is $\binom{m}{(1-\varepsilon)m} = \binom{m}{\varepsilon m}$.

By combining this with (4.1) and (4.2), we get that with probability at least $1-\delta$, we have

$$\begin{aligned}|\langle \nabla F_h(\boldsymbol{N}+Z-\boldsymbol{Y}), X-X'\rangle| &\leqslant h\cdot\left(10\gamma+10r\sqrt{\varepsilon m\log(e/\varepsilon)}+10r\sqrt{\log(1/\delta)}\right)\\ &\leqslant 100h\cdot\left(\gamma+r\sqrt{\log(1/\delta)}\right).\end{aligned}$$

□

Lemma 4.3 (Local strong convexity)**.** *Consider the settings of Theorem 4.1 and let $h \geqslant \zeta + 2b$. Denote $M = \boldsymbol{N} + Z - \boldsymbol{Y} = Z - X^*$. With probability at least $1-\delta$ over the randomness of $\boldsymbol{N}$, for every $X, X' \in \tilde{\Omega}$ satisfying $\|X - X'\| \geqslant 20\sqrt{\frac{b}{\alpha}(4\gamma + b\log(1/\delta))}$, we have*

$$F_h(M + X - X') \geqslant F_h(M) + \langle \nabla F_h(M), X - X'\rangle + \frac{\alpha}{10}\|X - X'\|_2^2 .$$

Proof. Fix some X, X' and let $\Delta = X - X'$. Using Lemma 3.2,

$$\begin{aligned}F_h(M+\Delta) &= \sum_{i\in[m]} f_h(M_i+\Delta_i)\\ &\geqslant \sum_{i\in[m]}\left(f_h(M_i)+f_h'(M_i)\cdot\Delta_i+\frac{\Delta_i^2}{2}\mathbf{1}_{[|M_i|\leqslant\zeta]}\cdot\mathbf{1}_{[|\Delta_i|\leqslant h-\zeta]}\right)\\ &= F_h(M)+\langle\nabla F_h(M),\Delta\rangle+\frac{1}{2}\sum_{i\in[m]}\mathbf{1}_{\{|M_i|\leqslant\zeta\}}\,\Delta_i^2 .\end{aligned}$$

Denote $U = \{i \in [m] : Z_i = \boldsymbol{Y}_i\}$. It suffices to show that with probability at least $1-\delta$, for every $X, X' \in \tilde{\Omega}$ satisfying $\|X - X'\| \geqslant 20\sqrt{\frac{b}{\alpha}(4\gamma + b\log(1/\delta))}$, we have

$$\sum_{i\in U}\mathbf{1}_{\{|\boldsymbol{N}_i|\leqslant\zeta\}}\cdot\Delta_i^2 \geqslant \frac{\alpha}{10}\|\Delta\|_2^2 ,$$

where $\Delta = X - X'$. To this end fix such an $X, X' \in \tilde{\Omega}$. For every $i \in [m]$, define the random variable

$$\boldsymbol{z}_i := \mathbf{1}_{\{|\boldsymbol{N}_i|\leqslant\zeta\}}\cdot\Delta_i^2 .$$

Let S be an arbitrary fixed (non-random) set of size $(1-\varepsilon)m$ and let

$$\boldsymbol{z} = \sum_{i\in S}\boldsymbol{z}_i .$$

We have

$$\mathbb{E}[\boldsymbol{z}] = \sum_{i\in S}\mathbb{E}[\boldsymbol{z}_i] = \sum_{i\in S}\mathbb{P}[|\boldsymbol{N}_i|\leqslant\zeta]\cdot\Delta_i^2 \geqslant \sum_{i\in[m]}\alpha\cdot\Delta_i^2 - 4\varepsilon m\cdot b^2 \geqslant \frac{\alpha}{2}\cdot\|\Delta\|_2^2 ,$$

where the last inequality holds because we assumed (without loss of generality) that $h\cdot\varepsilon\cdot m\log m \leqslant \gamma$, and hence

$$4\varepsilon m\cdot b^2 \leqslant h^2\cdot\varepsilon\cdot m\log m \leqslant h\gamma \leqslant \frac{\alpha}{2}\cdot\|\Delta\|_2^2.$$

On the other hand, since $0 \leqslant z_i \leqslant \Delta_i^2 \leqslant \|\Delta\|_{\max}^2 \leqslant 4b^2$ for all $i \in [m]$, we have

$$\sum_{i \in S} \mathbb{E}[z_i^2] \leqslant \sum_{i \in S} 4b^2 \cdot \mathbb{E}[z_i] = 4b^2 \cdot \mathbb{E}[z]\,.$$

It follows from Bernstein's inequality that for every $t > 0$, we have

$$\mathbb{P}[z - \mathbb{E}[z] \leqslant -t] \leqslant \exp\left(-\frac{t^2/2}{4b^2 \cdot \mathbb{E}[z] + \|\Delta\|_{\max}^2 \cdot t}\right) \leqslant \exp\left(-\frac{t^2/2}{4b^2 \cdot \mathbb{E}[z] + 4b^2 \cdot t}\right).$$

By taking $t = \frac{\mathbb{E}[z]}{2} \geqslant \frac{\alpha}{4} \cdot \|\Delta\|_2^2$, we get

$$\begin{aligned}\mathbb{P}\left[z \leqslant \frac{\alpha}{4}\|\Delta\|_2^2\right] \leqslant \mathbb{P}\left[z \leqslant \frac{\mathbb{E}[z]}{2}\right] \leqslant \mathbb{P}\left[z - \mathbb{E}[z] \leqslant -\frac{\mathbb{E}[z]}{2}\right]& \\ \leqslant \exp\left(-\frac{\mathbb{E}[z]^2/8}{4b^2\,\mathbb{E}[z] + 4b^2\,\mathbb{E}[z]/2}\right) \leqslant \exp\left(-\frac{\mathbb{E}[z]}{64b^2}\right) \leqslant \exp\left(-\frac{\alpha\|\Delta\|_2^2}{128b^2}\right).&\end{aligned}$$

By taking a union bound over all subsets[20] of size $(1-\varepsilon)m$, we get

$$\begin{aligned}\mathbb{P}\left[\sum_{i \in U} \mathbf{1}_{\{|N_i| \leqslant \zeta\}} \cdot \Delta_i^2 \leqslant \frac{\alpha}{4}\|\Delta\|_2^2\right] &\leqslant \binom{m}{(1-\varepsilon)m} \exp\left(-\frac{\alpha\|\Delta\|_2^2}{128b^2}\right) = \binom{m}{\varepsilon m} \exp\left(-\frac{\alpha\|\Delta\|_2^2}{128b^2}\right) \\ &\leqslant \left(\frac{e}{\varepsilon}\right)^{\varepsilon m} \exp\left(-\frac{\alpha\|\Delta\|_2^2}{128b^2}\right) = \exp\left(\varepsilon m \log(e/\varepsilon) - \frac{\alpha\|\Delta\|_2^2}{128b^2}\right) \\ &\leqslant \exp\left(\frac{2\gamma}{b} - \frac{\alpha\|\Delta\|_2^2}{128b^2}\right), \qquad (4.3)\end{aligned}$$

where in the last inequality, we used the fact that we assumed (without loss of generality) that $h \cdot \varepsilon \cdot m \log m \leqslant \gamma$, which means that

$$\begin{aligned}\varepsilon m \log(e/\varepsilon) &= \varepsilon m\left(1 + \log\frac{r^2 m \log m}{\gamma^2}\right) \\ &= \varepsilon m\left(1 + \log m + \log\log m + \log\frac{r^2}{\gamma^2}\right) \leqslant 4\varepsilon m \log m \leqslant \frac{4\gamma}{h} \leqslant \frac{2\gamma}{b}\,.\end{aligned}$$

Now define

$$L = 20\sqrt{\frac{b}{\alpha}\left(4\gamma + b\log(1/\delta)\right)}\,,$$

$$\mathfrak{D} := \left\{X - X' : X, X' \in \tilde{\Omega}\,,\ \|X - X'\|_2 \geqslant L\right\},$$

and

$$\tilde{\varepsilon} = \frac{L\sqrt{\alpha}}{10}\,,$$

[20] The number of such subsets is $\binom{m}{(1-\varepsilon)m} = \binom{m}{\varepsilon m}$.

and let $\mathcal{N}_{\tilde{\varepsilon}}(\mathfrak{D})$ be an $\tilde{\varepsilon}$-net of $\mathfrak{D}$ of minimal size. Using Sudakov's minoration Fact C.5, we have

$$\begin{aligned}\frac{\tilde{\varepsilon}}{2}\sqrt{\log|\mathcal{N}_{\tilde{\varepsilon}}(\mathfrak{D})|} &\leqslant \mathop{\mathbb{E}}_{\mathbf{g}\sim N(0,I_m)}\left[\sup_{\Delta\in\mathfrak{D}}\langle \mathbf{g},\Delta\rangle\right] \leqslant \mathop{\mathbb{E}}_{\mathbf{g}\sim N(0,I_m)}\left[\sup_{X,X'\in\tilde{\Omega}}\langle \mathbf{g}, X-X'\rangle\right] \\ &\leqslant \mathop{\mathbb{E}}_{\mathbf{g}\sim N(0,I_m)}\left[\sup_{X,X'\in\tilde{\Omega}}\langle \mathbf{g}, X\rangle + \sup_{X,X'\in\tilde{\Omega}}\langle \mathbf{g}, -X'\rangle\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{g}\sim N(0,I_m)}\left[\sup_{X\in\tilde{\Omega}}\langle \mathbf{g}, X\rangle\right] + \mathop{\mathbb{E}}_{\mathbf{g}\sim N(0,I_m)}\left[\sup_{X'\in\tilde{\Omega}}\langle \mathbf{g}, X'\rangle\right] = 2\gamma\,.\end{aligned}$$

By taking a union bound over $\mathcal{N}_{\tilde{\varepsilon}}(\mathfrak{D})$ and applying (4.3), it follows that

$$\begin{aligned}&\mathbb{P}\left[\exists\Delta\in\mathcal{N}_{\tilde{\varepsilon}}(\mathfrak{D})\,,\ \sum_{i\in U}\mathbf{1}_{\{|\mathbf{N}_i|\leqslant\zeta\}}\cdot\Delta_i^2\leqslant\frac{\alpha}{4}\|\Delta\|_2^2\right] \\ &\quad\leqslant|\mathcal{N}_{\tilde{\varepsilon}}(\mathfrak{D})|\exp\left(\frac{2\gamma}{b}-\frac{\alpha L^2}{128b^2}\right)\leqslant\exp\left(\frac{16\gamma^2}{\tilde{\varepsilon}^2}+\frac{2\gamma}{b}-\frac{\alpha L^2}{128b^2}\right)\leqslant\exp\left(\frac{1600\gamma^2}{\alpha L^2}+\frac{2\gamma}{b}-\frac{\alpha L^2}{128b^2}\right) \\ &\quad\overset{(*)}{\leqslant}\exp\left(\frac{\alpha L^2}{1600b^2}+\frac{\alpha L^2}{800b^2}-\frac{\alpha L^2}{128b^2}\right)\leqslant\exp\left(-\frac{\alpha L^2}{400b^2}\right)\overset{(\dagger)}{\leqslant}\delta\,,\end{aligned}$$

where $(*)$ follow from the fact that $L\geqslant 20\sqrt{\frac{4\gamma b}{\alpha}}$, which implies that $\frac{1600\gamma^2}{\alpha L^2}\leqslant\frac{\alpha L^2}{1600b^2}$ and $\frac{2\gamma}{b}\leqslant\frac{\alpha L^2}{800b^2}$. $(\dagger)$ follows from the fact that $L\geqslant 20\sqrt{\frac{b^2}{\alpha}\log(1/\delta)}$.

We conclude that with probability at least $1-\delta$, it holds that for every $\Delta\in\mathcal{N}_{\tilde{\varepsilon}}(\mathfrak{D})$, we have

$$\sum_{i\in U}\mathbf{1}_{\{|\mathbf{N}_i|\leqslant\zeta\}}\cdot\Delta_i^2\geqslant\frac{\alpha}{4}\|\Delta\|_2^2\,.$$

Assume that this event happens, and let $\Delta\in\mathfrak{D}$ be arbitrary. We can decompose Δ as

$$\Delta = A + B\,,$$

where $A\in\mathcal{N}_{\tilde{\varepsilon}}(\mathfrak{D})$ and $\|B\|_{\max}\leqslant\tilde{\varepsilon}$. We have

$$\begin{aligned}\sum_{i\in U}\mathbf{1}_{\{|\mathbf{N}_i|\leqslant\zeta\}}\cdot\Delta_i^2 &= \sum_{i\in U}\mathbf{1}_{\{|\mathbf{N}_i|\leqslant\zeta\}}\cdot(A_i+B_i)^2\geqslant\frac{1}{2}\sum_{i\in U}\mathbf{1}_{\{|\mathbf{N}_i|\leqslant\zeta\}}\cdot A_i^2-\sum_{i\in U}\mathbf{1}_{\{|\mathbf{N}_i|\leqslant\zeta\}}\cdot B_i^2 \\ &\geqslant\frac{\alpha}{4}\|A\|_2^2-\|B\|_2^2=\frac{\alpha}{4}\|\Delta-B\|_2^2-\|B\|_2^2\geqslant\frac{\alpha}{4}(\|\Delta\|_2-\|B\|_2)^2-\|B\|_2^2 \\ &\overset{(\ddagger)}{\geqslant}\frac{\alpha}{4}\left(\|\Delta\|_2-\frac{\|\Delta\|_2\sqrt{\alpha}}{10}\right)^2-\left(\frac{\|\Delta\|_2\sqrt{\alpha}}{10}\right)^2\geqslant\frac{\alpha}{10}\|\Delta\|_2^2\,,\end{aligned}$$

where $(\ddagger)$ follows from the fact that $\|B\|_{\max}\leqslant\tilde{\varepsilon}=\dfrac{L\sqrt{\alpha}}{10}\leqslant\dfrac{\|\Delta\|_2\sqrt{\alpha}}{10}$. □

We can now prove the theorem.

Proof of Theorem 4.1. We may assume $\|X^*-\hat{X}\|_2\geqslant 20\sqrt{\frac{b}{\alpha}\left(4\gamma+b\log(1/\delta)\right)}$, since otherwise the statement is trivially true. By definition, for $M=Z-X^*$, we have

$$F_h(Z-\hat{X})\leqslant F_h(Z-X^*)=F_h(M)\,,$$

and by Lemma 4.3, with probability $1-\delta$, we have

$$F_h(Z-\hat{X}) \geqslant F_h(M) - \langle \nabla F_h(M), X^* - \hat{X}\rangle + \frac{\alpha}{10}\left\|X^* - \hat{X}\right\|_2^2 .$$

Combining the two inequalities and rearranging, we get

$$\left\|X^* - \hat{X}\right\|_2^2 \leqslant \frac{10}{\alpha}\left|\langle \nabla F_h(\boldsymbol{N}), X^* - \hat{X}\rangle\right| \leqslant O\left(\frac{h}{\alpha}\left(\gamma + r\sqrt{\log(1/\delta)}\right)\right),$$

and the result follows. □

5 Applications

In this section we apply Theorem 4.1 to various estimation problems.

5.1 Tensor PCA with oblivious outliers

We show here how Theorem 4.1 can be used to recover a rank-1 tensor under symmetric noise. We will need the following fact:

Fact 5.1 ([HSS15]). *Let $p \geqslant 3$ be an odd number, and let $\boldsymbol{W} \in (\mathbb{R}^n)^{\otimes p}$ be a tensor with i.i.d. entries from $N(0,1)$. Then with probability*[21] $1-\delta$ *(over $\boldsymbol{W}$) every pseudo-distribution μ of degree at least $2p-2$ on indeterminates $x = (x_1, \ldots, x_n)$ satisfies*

$$\tilde{\mathbb{E}}_{x\sim\mu}\langle x^{\otimes p}, \boldsymbol{W}\rangle \leqslant C \cdot \left(\left(n^p \cdot p \cdot \ln n\right)^{1/4} + n^{p/4}(\ln(1/\delta))^{1/4} + n^{1/4}(\ln(1/\delta))^{3/4}\right) \cdot \left(\tilde{\mathbb{E}}_{x\sim\mu}\|x\|^{2p-2}\right)^{\frac{p}{2p-2}}$$

for some absolute constant C.

The following corollary is a simple application of Lemma A.4 to Fact 5.1.

Corollary 5.2. *Let $p \geqslant 3$ be an odd number and let $\boldsymbol{W} \in (\mathbb{R}^n)^{\otimes p}$ be a tensor with i.i.d. entries from $N(0,1)$. Then for every $d \geqslant 2p-2$, we have*

$$\mathbb{E}\left[\sup_{X\in\tilde{\Omega}_{n,d}} \langle X, \boldsymbol{W}\rangle\right] \leqslant O\left(\left(n^p \cdot p \cdot \ln n\right)^{1/4}\right),$$

where

$$\tilde{\Omega}_{n,d} = \left\{\tilde{\mathbb{E}}_{x\sim\mu} x^{\otimes p} : \mu \in \mathcal{P}_d\ ,\ \tilde{\mathbb{E}}_{x\sim\mu}\|x\|^2 \leqslant 1\right\},$$

and $\mathcal{P}_d$ is the set of pseudo-distributions over $\mathbb{R}[x] = \mathbb{R}[x_1, \ldots, x_n]$ of degree d.

The following fact is an easy consequence of Fact 3.10 and the bound on the expected value of the spectral norm of Gaussian $n^{p/2} \times n^{p/2}$ matrix:

[21] Note that Fact 5.1 follows from applying Theorem 56 of [HSS15] to p-tensors in the same way as it was applied to 3-tensors in order to prove Theorem 11 and Corollary 12 of [HSS15]. It is worth mentioning that the probability that was reported in [HSS15, Theorem 56] is $1 - O(n^{-100})$. However, a closer look at the proof of Theorem 56 in Page 47 of [HSS15], we can see that the probability at least $1-\delta$ can be easily obtained for arbitrarily small δ.

Fact 5.3. *Let $p \geqslant 2$ be an even number and let $\boldsymbol{W} \in (\mathbb{R}^n)^{\otimes p}$ be a tensor with i.i.d. entries from $N(0,1)$. Then for every $d \geqslant p$, we have*

$$\mathbb{E}\left[\sup_{X \in \tilde{\Omega}_{n,d}} \langle X, \boldsymbol{W} \rangle\right] \leqslant O\left(n^{p/4}\right),$$

where

$$\tilde{\Omega}_{n,d} = \left\{\tilde{\mathbb{E}}_{x\sim\mu} x^{\otimes p} : \mu \in \mathcal{P}_d \ , \ \tilde{\mathbb{E}}_{x\sim\mu}\|x\|^2 \leqslant 1\right\},$$

and $\mathcal{P}_d$ is the set of pseudo-distributions over $\mathbb{R}[x] = \mathbb{R}[x_1, \dots, x_n]$ of degree d.

Proof. Fix $X \in \tilde{\Omega}_{n,d}$ and let $\mu \in \mathcal{P}_d$ be such that $\tilde{\mathbb{E}}_{x\sim\mu}\|x\|^2 \leqslant 1$ and $X = \tilde{\mathbb{E}}_{x\sim\mu} x^{\otimes p}$. Denote by $\boldsymbol{W}'$ the $n^{p/2} \times n^{p/2}$ matrix that is obtained by reshaping $\boldsymbol{W}$. We have:

$$\langle X, \boldsymbol{W} \rangle = \tilde{\mathbb{E}}_{x\sim\mu}\left\langle x^{\otimes p}, \boldsymbol{W} \right\rangle = \tilde{\mathbb{E}}_{x\sim\mu}\left\langle x^{\otimes p/2}, \boldsymbol{W}' x^{\otimes p/2} \right\rangle \overset{(*)}{\leqslant} \|\boldsymbol{W}'\|_2 \tilde{\mathbb{E}}_{x\sim\mu}\|x^{\otimes p/2}\|_2^2 \overset{(\dagger)}{\leqslant} \|\boldsymbol{W}'\|_2\,,$$

$(*)$ follows from Fact 3.10 and $(\dagger)$ follows from the fact that

$$\tilde{\mathbb{E}}_{x\sim\mu}\|x^{\otimes p/2}\|_2^2 = \tilde{\mathbb{E}}_{x\sim\mu}(\|x\|_2^2)^{p/2} \leqslant 1\,.$$

Therefore,

$$\mathbb{E}\left[\sup_{X \in \tilde{\Omega}_{n,d}} \langle X, \boldsymbol{W} \rangle\right] \leqslant \mathbb{E}[\|\boldsymbol{W}'\|_2] \leqslant O\left(n^{p/4}\right).$$

The last inequality follows from the well known bounds on the expected spectral norm of a Gaussian matrix, and can be immediately seen from Fact A.1 and Lemma A.4. □

The following two theorems are about tensor PCA with asymmetric tensor noise.

Theorem 5.4 (Asymmetric Tensor Noise of odd order)**.** *Let $p \geqslant 3$ be an odd number. Let $n \in \mathbb{N}$, $n \geqslant 2$, $\lambda > 0$ and $\alpha \in (0, 1]$. Let $\boldsymbol{T} = \lambda \cdot v^{\otimes p} + \boldsymbol{N}$, where $v \in \mathbb{R}^n$ is a unit vector and $\boldsymbol{N}$ is a random tensor whose entries are independent (but not necessarily identically distributed), symmetric about zero and satisfy $\mathbb{P}\left[\left|\boldsymbol{N}_{i_1\dots i_p}\right| \leqslant 1\right] \geqslant \alpha$ for all $i_1, \dots, i_p \in [n]$.*

There exists an absolute constant $C > 1$ and an algorithm such that if

$$\lambda \geqslant \frac{C}{\alpha} \cdot \left(p \ln n\right)^{1/4} \cdot n^{p/4}$$

and

$$\|v\|_{\max} \leqslant \frac{(\alpha/C)^{1/p}}{n^{1/4}(p \ln n)^{1/(4p)}}\,,$$

then the algorithm on input $\boldsymbol{T}$ runs in time $\left(n^p\right)^{O(1)}$ and outputs a unit vector $\hat{\boldsymbol{v}} \in \mathbb{R}^n$ satisfying

$$\langle v, \hat{\boldsymbol{v}} \rangle \geqslant 0.99$$

with probability at least $1 - 2^{-n}$.

Furthermore, for $\varepsilon \leqslant \left(Cn^p \cdot p \ln n\right)^{-1/2}$, the same result holds if an arbitrary (adversarially chosen) ε-fraction of entries of $\boldsymbol{T}$ is replaced by adversarially chosen values.

Proof. We can apply Theorem 4.1 for input

$$\boldsymbol{Y} = \boldsymbol{T}/\lambda = v^{\otimes p} + \boldsymbol{N}' ,$$

where

$$\boldsymbol{N}' = \frac{1}{\lambda}\boldsymbol{N} .$$

In order to do so, define the set

$$\tilde{\Omega} = \left\{ \tilde{\mathbb{E}}_{x\sim\mu} x^{\otimes p} : \mu \in \mathcal{E} \right\} ,$$

where $\mathcal{E}$ is the set of pseudo-distributions over $\mathbb{R}[x_1, \ldots, x_n]$ of degree $2p$ that satisfy the constraints $\tilde{\mathbb{E}}_{x\sim\mu}\|x\|_2^2 \leqslant 1$ and $\tilde{\mathbb{E}}_{x\sim\mu} x_i^2 \leqslant \frac{1}{\lambda^{2/p}}$ for all $i \in [n]$.

Define $\zeta = \frac{1}{\lambda}$ so that for every $i_1, \ldots, i_p \in [n]$, we have

$$\mathbb{P}\left[\left| \boldsymbol{N}'_{i_1 \ldots i_p} \right| \leqslant \zeta \right] = \mathbb{P}\left[\left| \boldsymbol{N}_{i_1 \ldots i_p} \right| \leqslant 1 \right] \geqslant \alpha .$$

Now notice that

$$\begin{aligned} \max_{X\in\tilde{\Omega}} \|X\|_{\max} &= \max_{\mu\in\mathcal{E}} \|\tilde{\mathbb{E}}_{x\sim\mu} x^{\otimes p}\|_{\max} = \max_{\mu\in\mathcal{E}} \max_{i_1,\ldots,i_p\in[n]} \left| \tilde{\mathbb{E}}_{x\sim\mu} x_{i_1} \ldots x_{i_p} \right| \\ &\leqslant \max_{\mu\in\mathcal{E}} \max_{i_1,\ldots,i_p\in[n]} \sqrt{\tilde{\mathbb{E}}_{x\sim\mu} x_{i_1}^2 \ldots x_{i_p}^2} \overset{(*)}{\leqslant} \sqrt{\left(\frac{1}{\lambda^{2/p}}\right)^p} = \frac{1}{\lambda} , \end{aligned}$$

where $(*)$ follows from the fact that $\tilde{\mathbb{E}}_{x\sim\mu} x_i^2 \leqslant \frac{1}{\lambda^{2/p}}$ for all $\mu \in \mathcal{E}$ and all $i \in [n]$. Furthermore,

$$\max_{X\in\tilde{\Omega}} \|X\|_2 = \max_{\mu\in\mathcal{E}} \|\tilde{\mathbb{E}}_{x\sim\mu} x^{\otimes p}\|_2 \leqslant \max_{\mu\in\mathcal{E}} \sqrt{\tilde{\mathbb{E}}_{x\sim\mu} \|x^{\otimes p}\|_2^2} = \max_{\mu\in\mathcal{E}} \sqrt{\tilde{\mathbb{E}}_{x\sim\mu} \|x\|_2^{2p}} \leqslant 1 ,$$

where the last inequality follows from the fact that $\tilde{\mathbb{E}}_{x\sim\mu}\|x\|_2^2 \leqslant 1$ for all $\mu \in \mathcal{E}$.

Let $b = \frac{1}{\lambda}$ and $r = 1$ so that $\max_{X\in\tilde{\Omega}} \|X\|_{\max} \leqslant b$ and $\max_{X\in\tilde{\Omega}} \|X\|_2 \leqslant r$. By defining $h = \frac{3}{\lambda}$, we can see that

$$h \geqslant \zeta + 2b ,$$

Now from Corollary 5.2 and Fact 5.3, we can see that there is

$$\gamma = \Theta\Big(O\Big(\big(p \ln n\big)^{1/4} \cdot n^{p/4} \Big) \Big)$$

such that the Gaussian complexity of $\tilde{\Omega}$ satisfies

$$\mathbb{E}\left[\sup_{X\in\tilde{\Omega}} \langle X, \boldsymbol{W} \rangle \right] \leqslant \gamma ,$$

where $\boldsymbol{W}$ is a random tensor in $(\mathbb{R}^n)^{\otimes p}$ whose entries are i.i.d. standard Gaussian $N(0, 1)$.

If follows from Theorem 4.1 that with probability at least $1 - 2^{-n}$, the pseudo-distribution $\tilde{\mathbb{E}}$ that minimizes the Huber loss satisfies

$$\left\|v^{\otimes p} - \tilde{\mathbb{E}}x^{\otimes p}\right\|_2 \leqslant O\left(\sqrt{\frac{h}{\alpha}\left(\gamma + r\sqrt{\log(1/2^{-n})}\right)}\right)$$
$$\leqslant O\left(\sqrt{\frac{3}{\lambda\alpha}\left(\left(p\ln n\right)^{1/4} \cdot n^{p/4} + 1\sqrt{n\log 2}\right)}\right) = O\left(\frac{1}{C}\right).$$

By making $C > 0$ arbitrarily large, we can make the above bound on $\left\|v^{\otimes p} - \tilde{\mathbb{E}}x^{\otimes p}\right\|_2$ arbitrarily small.

Now we take $\hat{v} = \tilde{\mathbb{E}}x/\|\tilde{\mathbb{E}}x\|$.

By Lemma B.1, $\hat{v}$ satisfies the desired bound with probability at least $1 - 2^{-n}$. Note that Theorem Theorem 4.1 also implies that we can also afford an ε fraction of arbitrary adversarial changes in the observed tensor as long as

$$\varepsilon \leqslant \frac{\gamma^2}{r^2 n^p \log(n^p)} = \frac{O\left(\left(p\ln n\right)^{1/2} \cdot n^{p/2}\right)}{n^p p \log n} = O\left(\left(n^p \cdot p\ln n\right)^{-1/2}\right).$$

□

Theorem 5.5 (Asymmetric Tensor Noise of even order)**.** *Let $p \geqslant 2$ be an even number. Let $n \in \mathbb{N}$, $n \geqslant 2$, $\lambda > 0$ and $\alpha \in (0, 1]$. Let $\boldsymbol{T} = \lambda \cdot v^{\otimes p} + \boldsymbol{N}$, where $v \in \mathbb{R}^n$ is a unit vector and $\boldsymbol{N}$ is a random tensor whose entries are independent (but not necessarily identically distributed), symmetric about zero and satisfy $\mathbb{P}\left[\left|\boldsymbol{N}_{i_1 \ldots i_p}\right| \leqslant 1\right] \geqslant \alpha$ for all $i_1, \ldots, i_p \in [n]$.*

There exists an absolute constant $C > 1$ and an algorithm such that if

$$\lambda \geqslant \frac{C}{\alpha} \cdot n^{p/4}$$

and

$$\|v\|_{\max} \leqslant \frac{(\alpha/C)^{1/p}}{n^{1/4}},$$

then the algorithm on input $\boldsymbol{T}$ runs in time $(n^p)^{O(1)}$ and outputs a unit vector $\hat{v} \in \mathbb{R}^n$ satisfying

$$|\langle v, \hat{v}\rangle| \geqslant 0.99$$

with probability at least $1 - 2^{-n}$.

Furthermore, for $\varepsilon \leqslant \left(Cn^{p/2} \cdot p\ln n\right)^{-1}$, the same result holds if an arbitrary (adversarially chosen) ε-fraction of entries of $\boldsymbol{T}$ is replaced by adversarially chosen values.

Proof. The proof is very similar to the proof of Theorem 5.4, we only need to use Fact 5.3 to bound the Gaussian complexity. For rounding, we can take $\hat{v}$ to be the top eigenvector of $\tilde{\mathbb{E}}xx^\top$ and by Lemma B.2 $\hat{v}$ satisfies the desired bound with probability at least $1 - 2^{-n}$. □

The next two theorems is about tensor PCA with symmetric tensor noise.

Theorem 5.6 (Symmetric Tensor Noise of odd order)**.** *Let $p \geqslant 3$ be an odd number. Let $n \in \mathbb{N}$, $n \geqslant 2$, $\lambda > 0$ and $\alpha \in (0,1]$. Let $\boldsymbol{T} = \lambda \cdot v^{\otimes p} + \boldsymbol{N}$, where $v \in \mathbb{R}^n$ is a unit vector and $\boldsymbol{N}$ is a random symmetric tensor whose entries $\boldsymbol{N}_{i_1 \ldots i_p}$ with indices $i_1 \leqslant i_2 \leqslant \ldots \leqslant i_p$ are independent (but not necessarily identically distributed), symmetric about zero and satisfy $\mathbb{P}\left[\left|\boldsymbol{N}_{i_1\ldots i_p}\right| \leqslant 1\right] \geqslant \alpha$.*

There exists an absolute constant $C > 1$ and an algorithm such that if

$$\lambda \geqslant \frac{Cp!}{\alpha} \cdot \left(p \ln n\right)^{1/4} \cdot n^{p/4}$$

and

$$\|v\|_{\max} \leqslant \frac{\alpha^{1/p}}{(Cp!)^{1/p} \cdot n^{1/4} \cdot (p \ln n)^{1/4p}}\,,$$

then the algorithm on input $\boldsymbol{T}$ runs in time $(n^p)^{O(1)}$ and outputs a unit vector $\hat{\boldsymbol{v}} \in \mathbb{R}^n$ satisfying

$$\langle v, \hat{\boldsymbol{v}} \rangle \geqslant 0.99$$

with probability at least $1 - 2^{-n}$.

Furthermore, for $\varepsilon \leqslant \left(Cn^p \cdot p \ln n\right)^{-1/2}$, the same result holds if an arbitrary (adversarially chosen) ε-fraction of entries of $\boldsymbol{T}$ is replaced by adversarially chosen values.

Proof. The proof is similar to the asymmetric $\boldsymbol{N}$, but we apply Theorem 4.1 for input

$$\boldsymbol{Y} = \boldsymbol{T}'/\lambda = X' + \boldsymbol{N}'\,,$$

where $\boldsymbol{T}'$, X' and $\boldsymbol{N}'$ are the restrictions of $\boldsymbol{T}$, $v^{\otimes p}$ and $\frac{1}{\lambda}\boldsymbol{N}$ to the entries $\boldsymbol{T}_{i_1\ldots i_p}$, $(v^{\otimes})_{i_1\ldots i_p}$ and $\frac{1}{\lambda}\boldsymbol{N}_{i_1\ldots i_p}$ with indices $i_1 \leqslant i_2 \leqslant \ldots \leqslant i_p$, respectively. We also use the set

$$\tilde{\Omega} = \left\{\left(\tilde{\mathbb{E}}_{x\sim\mu} x_{i_1} \cdots x_{i_p}\right)_{i_1 \leqslant \ldots \leqslant i_p} : \mu \in \mathcal{E}\right\},$$

where $\mathcal{E}$ is as in Theorem 5.4.

We define $\zeta = \frac{1}{\lambda}$ so that for every $1 \leqslant i_1 \leqslant \ldots \leqslant i_p \leqslant n$, we have

$$\mathbb{P}\left[\left|\boldsymbol{N}'_{i_1\ldots i_p}\right| \leqslant \zeta\right] = \mathbb{P}\left[\left|\boldsymbol{N}_{i_1\ldots i_p}\right| \leqslant 1\right] \geqslant \alpha\,.$$

By defining $r = 1$ and $b = \frac{1}{\lambda}$, we can show similarly to Theorem 5.4 that $\max_{X \in \tilde{\Omega}} \|X\|_{\max} \leqslant b$ and $\max_{X\in\tilde{\Omega}} \|X\|_2 \leqslant r$. By defining $h = \frac{3}{\lambda}$, we can see that

$$h \geqslant \zeta + 2b\,.$$

Also similarly to Theorem 5.4, we can show that for some $\gamma = \Theta\left(\left(p \ln n\right)^{1/4} \cdot n^{p/4}\right)$, the Gaussian complexity of $\tilde{\Omega}$ can be bounded[22] as

$$\mathbb{E}\left[\sup_{X\in\tilde{\Omega}} \langle X, \boldsymbol{W}\rangle\right] \leqslant \gamma\,,$$

[22] We use Corollary 5.2 and Fact 5.3, together with Fact C.1 which implies that the Gaussian complexity does not increase if we restrict to a subset of coordinates.

where $(\boldsymbol{W})_{i_1\ldots i_p}$ are i.i.d. standard Gaussian $N(0,1)$ for $1 \leqslant i_1 \leqslant \ldots \leqslant i_p \leqslant n$.

If follows from Theorem 4.1 that with probability at least $1-2^{-n}$, the pseudo-distribution $\tilde{\mathbb{E}}$ that minimizes the Huber loss satisfies

$$\sum_{1\leqslant i_1\leqslant\ldots\leqslant i_p\leqslant n}\left(v_{i_1}\cdots v_{i_p}-\tilde{\mathbb{E}}x_{i_1}\cdots x_{i_p}\right)^2 \leqslant O\left(\sqrt{\frac{h}{\alpha}\left(\gamma+r\sqrt{\log(1/2^{-n})}\right)}\right)$$
$$\leqslant O\left(\sqrt{\frac{3}{\lambda\alpha}\left(\left(p\ln n\right)^{1/4}\cdot n^{p/4}+1\sqrt{n\log 2}\right)}\right)$$
$$\leqslant O\left(\frac{1}{Cp!}\right).$$

Therefore, with probability at least $1-2^{n}$, we have

$$\left\|v^{\otimes p}-\tilde{\mathbb{E}}x^{\otimes p}\right\|_2 = p!\cdot\sum_{1\leqslant i_1\leqslant\ldots\leqslant i_p\leqslant n}\left(v_{i_1}\cdots v_{i_p}-\tilde{\mathbb{E}}x_{i_1}\cdots x_{i_p}\right)^2 \leqslant O\left(\frac{1}{C}\right).$$

The remaining of the proof is the same as in the asymmetric case. □

We can similarly modify the proof of Theorem 5.5 to get the theorem for symmetric tensor noise:

Theorem 5.7 (Symmetric Tensor Noise of even order)**.** *Let $p \geqslant 2$ be an even. Let $n \in \mathbb{N}$, $n \geqslant 2$, $\lambda > 0$ and $\alpha \in (0,1]$. Let $\boldsymbol{T} = \lambda\cdot v^{\otimes p}+\mathbf{N}$, where $v \in \mathbb{R}^n$ is a unit vector and $\mathbf{N}$ is a random symmetric tensor whose entries $\mathbf{N}_{i_1\ldots i_p}$ with indices $i_1 \leqslant i_2 \leqslant \ldots \leqslant i_p$ are independent (but not necessarily identically distributed), symmetric about zero and satisfy $\mathbb{P}\left[\left|\mathbf{N}_{i_1\ldots i_p}\right| \leqslant 1\right] \geqslant \alpha$.*

There exists an absolute constant $C > 1$ and an algorithm such that if

$$\lambda \geqslant \frac{Cp!}{\alpha}\cdot n^{p/4}$$

and

$$\|v\|_{\max} \leqslant \frac{\left(\alpha/(Cp!)\right)^{1/p}}{n^{1/4}},$$

then the algorithm on input $\boldsymbol{T}$ runs in time $(n^p)^{O(1)}$ and outputs a unit vector $\hat{\boldsymbol{v}} \in \mathbb{R}^n$ satisfying

$$|\langle v,\hat{\boldsymbol{v}}\rangle| \geqslant 0.99$$

with probability at least $1-2^{-n}$.

Furthermore, for $\varepsilon \leqslant \left(Cn^{p/2}\cdot p\ln n\right)^{-1}$, the same result holds if an arbitrary (adversarially chosen) ε-fraction of entries of $\boldsymbol{T}$ is replaced by adversarially chosen values.

5.2 Sparse PCA with oblivious outliers

We will use the system of constraints for sparse PCA from [dKNS20].

Let $t \leqslant k$ and let $\mathcal{S}_t$ be the set of all n-dimensional vectors with values in $\{0,1\}$ that have exactly t nonzero coordinates.

We start with the following definition.

Definition 5.8. For every $u \in \mathcal{S}_t$ we define the following polynomial in variables $s := (s_1, \ldots, s_n)$

$$p_u(s) = \binom{k}{t}^{-1} \cdot \prod_{i \in \operatorname{supp}\{u\}} s_i \,.$$

Note that if v is a k-sparse vector and s is the indicator of its support, then for every $u \in \mathcal{S}_t$, we have

$$p_u(s) = \begin{cases} \binom{k}{t}^{-1} & \text{if } \operatorname{supp}\{u\} \subseteq \operatorname{supp}\{v\}\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Now consider the following system $\mathcal{C}_{s,x}$ of polynomial constraints.

$$\mathcal{C}_{s,x}\colon \left\{ \begin{aligned} &\forall i \in [n], & s_i^2 &= s_i \\ & & \sum_{i \in [n]} s_i &= k \\ &\forall i \in [n], & s_i \cdot x_i &= x_i \\ & & \sum_{i \in [n]} x_i^2 &= 1 \\ & & \sum_{u \in \mathcal{S}_t} p_u(s) &= 1 \\ &\forall i \in [n], & \sum_{u \in \mathcal{S}_t} u_i p_u(s) &= \frac{t}{k} \cdot s_i \end{aligned} \right\} \tag{5.1}$$

It is easy to see that if x is k-sparse and s is the indicator of its support, then x and s satisfy these constraints.

In [dKNS20] a different model of Sparse PCA is considered than the one we study here. There, a bound on $v^\top \boldsymbol{M} v$ is certified where $\boldsymbol{M}$ is the centered Wishart matrix, while we need to certify the bound for standard Gaussian matrix $\boldsymbol{M}$. The proofs of [dKNS20] can be easily adapted for our case. In Appendix B.2 we show that the Gaussian complexity of the set of degree $4t$ pseudo-distributions that satisfy the constraints $\mathcal{C}_{s,x}$ is bounded by $O\left(k\sqrt{\frac{\log n}{t}}\right)$.

Now we are able to show how the algorithm from Theorem 4.1 can be used to solve the sparse PCA problem with general noise with symmetric independent entries.

Theorem 5.9. *Let $n, k \in \mathbb{N}$, $k \leqslant n$, $\lambda > 0$ and $\alpha \in (0,1]$. Let $\boldsymbol{M} = \lambda \cdot vv^\top + \boldsymbol{N}$, where $v \in \mathbb{R}^n$ is a k-sparse unit vector and $\boldsymbol{N}$ is a random matrix with independent (but not necessarily identically distributed) symmetric about zero entries that satisfy $\mathbb{P}\left[\left|\boldsymbol{N}_{ij}\right| \leqslant 1\right] \geqslant \alpha$.*

There exists an absolute constant $C > 1$ and an algorithm such that if $\lambda \geqslant k \geqslant C\ln(n)/\alpha^2$ and $\|v\|_{\max} \leqslant 100/\sqrt{k}$, then the algorithm on input $\boldsymbol{M}$ runs in time $n^{O(\log(n)/\alpha^2)}$ and outputs $\hat{v} \in \mathbb{R}^n$ satisfying

$$|\langle v, \hat{v} \rangle| \geqslant 0.99$$

with probability at least $1 - n^{-100}$.

Moreover, the same result holds if we only get the upper triangle (without the diagonal) of the matrix $\boldsymbol{M}$ as input.

Furthermore, for $\varepsilon \leqslant \frac{k^2\alpha^2}{Cn^2 \ln n}$, the same result holds if an arbitrary (adversarially chosen) ε-fraction of entries of $\boldsymbol{M}$ is replaced by adversarially chosen values.

Proof. We can apply Theorem 4.1 for input

$$\boldsymbol{Y} = \boldsymbol{M}/\lambda = vv^\top + \boldsymbol{N}' ,$$

where $\boldsymbol{N}' = \frac{1}{\lambda}\boldsymbol{N}$. We also use the set

$$\tilde{\Omega} = \left\{\tilde{\mathbb{E}}_{x\sim\mu} xx^\top : \mu \in \mathcal{E}\right\} ,$$

where $\mathcal{E}$ is the set of pseudo-distributions over $\mathbb{R}[x_1, \dots, x_n]$ of degree $4t$ that satisfy the constraints Eq. (5.1) and additional constraints $\tilde{\mathbb{E}}x_i^2 \leqslant 100^2/k$ for all $i \in [n]$. Note that we choose $t = \lceil C\ln(n)/\alpha^2\rceil \leqslant k$.

If we define $b = \frac{100}{k}$ and $r = 1$, it is not hard to see from the constraints Eq. (5.1) and $\tilde{\mathbb{E}}x_i^2 \leqslant 100^2/k$ for all $i \in [n]$ that $\max_{X\in\tilde{\Omega}}\|X\|_{\max} \leqslant b$ and $\max_{X\in\tilde{\Omega}}\|X\|_2 \leqslant r$. Let $h = \frac{201}{k}$, so that

$$h \geqslant \zeta + 2b .$$

From Lemma B.6, we can see that there exists $\gamma = \Theta\left(k\sqrt{\frac{\ln n}{t}}\right)$ such that the Gaussian complexity of $\tilde{\Omega}$ can be bounded by γ. We conclude from Theorem 4.1 that with probability at least $1 - n^{-100}$, the pseudo-distribution $\tilde{\mathbb{E}}$ that minimizes the Huber loss satisfies

$$\begin{aligned}
\left\|vv^\top - \tilde{\mathbb{E}}xx^\top\right\|_2 &\leqslant O\left(\sqrt{\frac{h}{\alpha}\left(\gamma + r\sqrt{\log(1/n^{-100})}\right)}\right) \\
&\leqslant O\left(\sqrt{\frac{201}{k\alpha}\left(k\sqrt{\frac{\ln n}{t}} + 1\sqrt{100\log n}\right)}\right) \\
&\leqslant O\left(\sqrt{\frac{1}{\alpha}\sqrt{\frac{\ln n}{t}} + \frac{\sqrt{\log n}}{k\alpha}}\right) \leqslant O\left(\sqrt{\frac{1}{\sqrt{C}} + \frac{\alpha}{C\sqrt{\log n}}}\right) \\
&\leqslant O\left(\frac{1}{C^{1/4}}\right),
\end{aligned}$$

where the last inequality follows from the fact that $t \geqslant C\ln(n)/\alpha^2$ and $k \geqslant C\ln(n)/\alpha^2$. If we choose $C > 1$ to be large enough, we can make the above bound on $\left\|vv^\top - \tilde{\mathbb{E}}xx^\top\right\|_2$ to be arbitrarily small. Therefore, by Lemma B.2, the top eigenvector $\hat{v}$ of $\tilde{\mathbb{E}}xx^\top$ satisfies the desired bound with probability at least $1 - n^{-100}$. Note that Theorem Theorem 4.1 also implies that we can also afford an ε fraction of arbitrary adversarial changes in the observed matrix as long as

$$\varepsilon \leqslant \frac{\gamma^2}{r^2 n^2 \log(n^2)} = \frac{O\left(k^2\frac{\ln n}{t}\right)}{2n^2\log(n)} = O\left(\frac{k^2}{tn^2}\right) = O\left(\frac{k^2\alpha^2}{n^2\ln(n)}\right).$$

If we only get the upper triangle of $\boldsymbol{M}$ as input, we can optimize the Huber loss over

$$\tilde{\Omega} = \left\{\left(\tilde{\mathbb{E}}_{x\sim\mu}x_i x_j\right)_{i<j} : \mu \in \mathcal{E}\right\} .$$

The Gaussian complexity of $\tilde{\Omega}$ is bounded[23] by $O\left(k\sqrt{\frac{\ln n}{t}}\right)$, hence the pseudo-distribution $\tilde{\mathbb{E}}$ that minimizes the Huber loss satisfies

$$\sum_{i\neq j}\left(v_iv_j - \tilde{\mathbb{E}}x_ix_j\right)^2 \leqslant 2\sum_{1\leqslant i<j\leqslant n}\left(v_iv_j - \tilde{\mathbb{E}}x_ix_j\right)^2 \leqslant O\left(\frac{1}{\sqrt{C}}\right)$$

with probability at least $1 - n^{-100}$. Moreover,

$$\sum_{1\leqslant i\leqslant n}\left(v_i^2 - \tilde{\mathbb{E}}x_i^2\right)^2 \leqslant \max_{1\leqslant j\leqslant n}\left(v_j^2 + \tilde{\mathbb{E}}x_j^2\right)\sum_{1\leqslant i\leqslant n}\left(v_i^2 + \tilde{\mathbb{E}}x_i^2\right) \leqslant O(1/k)\,.$$

Hence, with probability at least $1 - n^{-100}$, we have

$$\left\|vv^\top - \tilde{\mathbb{E}}xx^\top\right\|_2 \leqslant O\left(\sqrt{\frac{1}{\sqrt{C}} + \frac{1}{k}}\right).$$

The remaining of the proof is the same as when the input is the whole matrix $\boldsymbol{M}$. □

6 Reduction from the planted clique problem

6.1 Sparse PCA

In this section we show that the running time $n^{O(\log n)}$ for sparse PCA with symmetric noise is likely to be inherent. We will use a reduction from the planted clique problem. Reductions from the planted clique problem to different models of sparse PCA were studied in [BR13a, BR13b, WBS16, GMZ17, BBH18, BB19]. Our analysis is simpler since our noise model is less restrictive than models considered in prior works. In fact, the planted clique problem can be seen as a special case of sparse PCA with symmetric noise (when only upper triangle without the diagonal is given as input).

Recall that the instance of planted clique problem is a random graph sampled according to the following distribution $G(n,q,k)$: First, some graph is sampled from Erdős-Rényi distribution $G(n,q)$ (where $q \in (0,1)$ is the probability of including an edge), and then a random subset of vertices of size $k \leqslant n$ is chosen and the clique corresponding to these vertices is added to the graph. The goal is to find the clique. It is possible to find in time $n^{O(\log n)}$ for constant[24] q if $k \geqslant \omega(\log n)$, but no polynomial time algorithm is known for $\omega(\log n) \leqslant k \leqslant o(\sqrt{n})$.

In this section we assume that $\omega(\log n) < k < n^{0.49}$. Currently no $n^{o(\log n)}$-time algorithm is known to solve this problem in this regime (for constant q), and for $q = 1/2$ and for some $k = n^{\Omega(1)}$ it is conjectured to be impossible to solve it in time $n^{o(\log n)}$ (see [MRS21] for more details).

Let $\boldsymbol{M} = \lambda \cdot vv^\top + \boldsymbol{N}$, where v is a k-sparse unit vector whose nonzero entries are equal to $1/\sqrt{k}$, $\boldsymbol{N}$ is a random matrix with independent (but not necessarily identically distributed) entries that satisfy $\mathbb{P}\left[\left|\boldsymbol{N}_{ij}\right| \leqslant 1\right] = 1$.

Also suppose that we get only the upper triangle (without the diagonal) of the matrix $\boldsymbol{M}$ as input. There are algorithms that can solve sparse PCA problem that only observe the upper triangle and match (up to a constant factor) current best known guarantees (if $\|v\|_4^4 \ll 1$ which is true if

[23] We use Fact C.1 which implies that the Gaussian complexity does not increase if we restrict to a subset of coordinates.
[24] I.e., $\Omega(1) \leqslant q \leqslant 1 - \Omega(1)$.

$\|v\|_{\max} \leqslant 100/\sqrt{k}$). Hence we assume that for flat vectors the problem does not become harder if we get only the upper triangle of $\boldsymbol{M}$ as input. We denote the upper triangle matrix of $\boldsymbol{M}$ as $\boldsymbol{\mathcal{U}}(\boldsymbol{M})$.

Now let $\boldsymbol{G} \sim G(n, 1/2, k)$ be a random graph with a planted clique of size k. Let $\boldsymbol{A}$ be the adjacency matrix of $\boldsymbol{G}$. Let J be the matrix with all entries equal to 1 and let $\boldsymbol{C} = 2\boldsymbol{A} - J$. Note that $\boldsymbol{\mathcal{U}}(\boldsymbol{C}) = \boldsymbol{\mathcal{U}}(k \cdot vv^\top + \boldsymbol{N})$, where $\sqrt{k} \cdot v$ is the indicator vector of the vertices of the clique (so it is k-sparse), and $\boldsymbol{N}$ is the noise whose entries that correspond to the vertices of the clique are zero, and other entries are iid $\{\pm 1\}$.

If we could recover v from $\boldsymbol{\mathcal{U}}(k \cdot vv^\top + \boldsymbol{N})$ in time $n^{o(\log n)}$, we would be able to find the planted clique in $\boldsymbol{G} \sim G(n, 1/2, k)$ in time $n^{o(\log n)}$.

Moreover, we can make the noise even smaller and the problem is likely to remain hard. That is, if we could recover v from $\boldsymbol{\mathcal{U}}(k \cdot vv^\top + \boldsymbol{N})$ where for all $(i, j) \in [n]^2$, $\mathbb{P}\left[\boldsymbol{N}_{ij} = 0\right] \geqslant \alpha$, then we would be able to find the planted clique in $G(n, (1-\alpha)/2, k)$. Indeed, for $p = (1-\alpha)/2$ let $\boldsymbol{G} \sim G(n, p, k)$ and let $\boldsymbol{A}$ be the adjacency matrix of $\boldsymbol{G}$. Let $\boldsymbol{B}$ be a random matrix such that $\boldsymbol{B}_{ij} = 0$ for all (i, j) such that $\boldsymbol{A}_{ij} = 1$, and for other (i, j), $\boldsymbol{B}_{ij}$ is 0 with probability $p/(1-p)$ and 0.5 with probability $1 - p/(1-p) = \alpha/(1-p)$. Let J be the matrix with all entries equal to 1 and let $\boldsymbol{C} = 2\boldsymbol{A} + 2\boldsymbol{B} - J$. Note that $\boldsymbol{\mathcal{U}}(\boldsymbol{C}) = \boldsymbol{\mathcal{U}}(k \cdot vv^\top + \boldsymbol{N})$, where $\sqrt{k} \cdot v$ is the indicator vector of the vertices of the clique, and $\boldsymbol{N}$ is the noise matrix with independent entries that satisfy $\mathbb{P}\left[\boldsymbol{N}_{ij} = 0\right] \geqslant \alpha$.

Hence if we could recover v from $\boldsymbol{\mathcal{U}}(k \cdot vv^\top + \boldsymbol{N})$, where $\mathbb{P}\left[\boldsymbol{N}_{ij} = 0\right] \geqslant 0.99$ in time $n^{o(\log n)}$, then we could find the planted clique in $\boldsymbol{G} \sim G(n, 0.005, k)$ in time $n^{o(\log n)}$, which currently known algorithms cannot do.

Remark 6.1. Exact recovery of v by the sparse PCA algorithm is not necessary in order for the reduction to work: As we shall see, if we only get unit $\hat{\boldsymbol{v}}$ that has correlation $\rho = \Omega(1)$ with v, we can still find the clique. First notice that since $\sum_{i \in [n]} |v_i| \hat{\boldsymbol{v}}_i \geqslant \sum_{i \in [n]} v_i \hat{\boldsymbol{v}}_i$, we can assume without loss of generality that the entries of $\hat{\boldsymbol{v}}$ are nonnegative. Now consider the set $S \subseteq [n]$ containing the indices the of top $4k/\rho^2$ entries of $\hat{\boldsymbol{v}}$. Then

$$\sum_{i \notin S} v_i \hat{\boldsymbol{v}}_i = \sum_{i \in \operatorname{supp}(v) \setminus S} \frac{1}{\sqrt{k}} \hat{\boldsymbol{v}}_i \overset{(*)}{\leqslant} k \cdot \frac{1}{\sqrt{k}} \cdot \frac{\rho}{2\sqrt{k}} = \rho/2\,,$$

where $(*)$ follows from the fact that $|\operatorname{supp}(v) \setminus S| \leqslant |\operatorname{supp}(v)| = k$ and that for every $i \notin S$, we have[25] $\hat{\boldsymbol{v}}_i^2 \leqslant \frac{\rho^2}{4k}$. We conclude that

$$\sum_{i \in S} v_i \hat{\boldsymbol{v}}_i = \langle v, \hat{\boldsymbol{v}} \rangle - \sum_{i \notin S} v_i \hat{\boldsymbol{v}}_i \geqslant \rho/2\,.$$

Now let $S' = S \cap \operatorname{supp}(v)$. We have

$$\sqrt{k}\rho/2 \leqslant \sqrt{k} \sum_{i \in S} v_i \hat{\boldsymbol{v}}_i = \sqrt{k} \sum_{i \in S \cap \operatorname{supp}(v)} \frac{1}{\sqrt{k}} \hat{\boldsymbol{v}}_i = \sum_{i \in S'} \hat{\boldsymbol{v}}_i \leqslant \sqrt{|S'|}\,,$$

i.e.,

$$|S'| \geqslant k\rho^2/4\,.$$

[25] This is a consequence of $\sum_{i \in [n]} \hat{\boldsymbol{v}}_i^2 = 1$ and the fact that $S \subseteq [n]$ contains the indices the of top $4k/\rho^2$ entries of $\hat{\boldsymbol{v}}$.

So if we restrict the graph to the vertices corresponding to S, we can find[26] the clique corresponding to S' in polynomial time as long as $|S'| \geqslant \omega\left(\sqrt{|S| \log n}\right)$, which is true for $\rho = \Omega(1)$ and $k \geqslant \omega(\log n)$. Then we can then easily find the remaining of the clique by searching for all vertices that are adjacent to every vertex in S'.

6.2 Tensor PCA

In this section, we provide evidence that the assumption on $\|v\|_{\max}$ in Theorem 5.4 is likely to be inherent, at least for our SoS-based approach.

First, we notice that exactly the same reasoning as that of Section 6.1 for obtaining a reduction from the planted clique problem to sparse PCA, can also be applied to get a reduction from the problem of recovering a planted clique in a random p-hypergraph to the problem of recovering a k-sparse unit vector v from the upper simplex of $\boldsymbol{Y} = k^{p/2} \cdot v^{\otimes p} + \boldsymbol{N}$, i.e., from the entries $\boldsymbol{Y}_{i_1,\ldots,i_p}$ such that $i_1 < \ldots < i_p$. It is conjectured that for every constant p, if $k < n^{0.49}$, then the problem of recovering a planted clique in a random p-hypergraph cannot be solved in polynomial time (see [LZ20] for more details). Hence, we expect that if $k < n^{0.49}$ then it is not possible to efficiently recover a k-sparse unit vector v from the upper simplex of $\boldsymbol{Y} = k^{p/2} \cdot v^{\otimes p} + \boldsymbol{N}$.

Second, we show that recovering from the upper simplex is not harder than recovering from the entire tensor $\boldsymbol{Y} = k^{p/2} \cdot v^{\otimes p} + \boldsymbol{N}$, at least for the algorithmic approach that is provided in Theorem 5.4. We proceed similarly to how we showed in Theorem 5.9 that recovering from the upper triangle matrix (without the diagonal) is not harder than recovering from the entire matrix. More precisely, we show that if it is possible to get a sum-of-squares certificate of the bound on the Gaussian complexity in such a way that shows that the algorithm in Theorem 5.4 can recover a k-sparse vector v from the entire $\boldsymbol{Y}$, then by slightly modifying the algorithm in Theorem 5.4 we can also recover v from the upper simplex of $\boldsymbol{Y}$.

Let us start by considering the case $p = 3$. Let $b = O(1/\sqrt{k})$ be a bound on the entries of v. Since we know that v is k-sparse, we can restrict the optimization problem in the algorithm of Theorem 5.4 to the pseudodistributions satisfying the constraint $\sum_{j=1}^{n} |\tilde{\mathbb{E}} x_j| \leqslant bk$. Similarly to the proof of Theorem 5.9, we notice that

$$\sum_{1\leqslant i\leqslant n} \left(v_i^3 - \tilde{\mathbb{E}} x_i^3\right)^2 \leqslant \max_{1\leqslant j\leqslant n}\left(|v_j^3| + |\tilde{\mathbb{E}} x_j^3|\right) \sum_{1\leqslant i\leqslant n} \left(|v_i^3| + |\tilde{\mathbb{E}} x_i^3|\right) \overset{(*)}{\leqslant} O(kb^6) \leqslant o(1)\,,$$

and

$$\begin{aligned}\sum_{1\leqslant i,j\leqslant n} \left(v_i^2 v_j - \tilde{\mathbb{E}} x_i^2 x_j\right)^2 &\leqslant \max_{1\leqslant i',j'\leqslant n}\left(\left|v_{i'}^2 v_{j'}\right| + \left|\tilde{\mathbb{E}} x_{i'}^2 x_{j'}\right|\right) \cdot \sum_{1\leqslant i,j\leqslant n}\left(\left|v_i^2 v_j\right| + \left|\tilde{\mathbb{E}} x_i^2 x_j\right|\right) \\ &\leqslant 2b^3 \sum_{1\leqslant j\leqslant n}\left(\left|v_j\right| + \left|\tilde{\mathbb{E}} x_j\right|\right) \overset{(\dagger)}{\leqslant} O(kb^4) \leqslant o(1)\,,\end{aligned}$$

[26] We apply the well-known spectral algorithm for the planted clique problem. It is worth mentioning that the $\log n$ factor comes from the fact that S is not independent from the graph, and hence the distribution of the graph that is induced by the vertices in S does not exactly match that of the planted clique problem. By taking a union bound over all sets of size $|S|$ and using standard concentration bounds for the spectral norm of symmetric matrices with i.i.d. subgaussian entries, one can show that the maximal spectral norm among all submatrices of the centered adjacency matrix of the random graph is bounded by $O(|S| \log n)$ with high probability.

where $(*)$ and $(\dagger)$ follow from the constraint $\sum_{j=1}^{n}|\tilde{\mathbb{E}}x_j| \leqslant bk$. The remaining of the proof is similar to the proof of Theorem 5.9.

For general p-order tensors with $3 \leqslant p \leqslant O(1)$, we can get a similar bound if we add the constraints $\sum_{1\leqslant i_1,\ldots,i_r\leqslant n}|\tilde{\mathbb{E}}x_{i_1}\cdots x_{i_r}| \leqslant b^r k^r$ for all $1 \leqslant r \leqslant p-2$, from which we can deduce that

$$\sum_{\substack{1\leqslant i_1,\ldots,i_p\leqslant n:\\ \exists j\neq j',\ i_j=i_{j'}}} \left((v^{\otimes p})_{i_1\ldots i_p} - \tilde{\mathbb{E}}(x^{\otimes p})_{i_1\ldots i_p}\right)^2 \leqslant o(1)\,.$$

The above implies that for pseudodistributions that satisfy the added constraints, we have

$$\|v^{\otimes p} - \tilde{\mathbb{E}}x^{\otimes p}\|_2^2 \leqslant o(1) + p!\cdot \sum_{1\leqslant i_1<\ldots<i_p\leqslant n} \left((v^{\otimes p})_{i_1\ldots i_p} - \tilde{\mathbb{E}}(x^{\otimes p})_{i_1\ldots i_p}\right)^2. \tag{6.1}$$

If we could get a sum-of-squares certificate of the bound on the Gaussian complexity showing that our degree-ℓ SoS-based algorithm in Theorem 5.4 can recover k-sparse flat vectors from $\boldsymbol{Y} = k^{p/2}\cdot v^{\otimes p} + \boldsymbol{N}$, then the same bound would imply that in the case where we only observe the upper simplex of $\boldsymbol{Y}$, the algorithm of Theorem 5.4 restricted to pseudoexpectations on the upper simplex would give[27] a pseudodistribution satisfying

$$\sum_{1\leqslant i_1<\ldots<i_p\leqslant n} \left((v^{\otimes p})_{i_1\ldots i_p} - \tilde{\mathbb{E}}(x^{\otimes p})_{i_1\ldots i_p}\right)^2 \ll 1\,.$$

If we also require that the pseudodistributions satisfy the constraints $\sum_{1\leqslant i_1,\ldots,i_r\leqslant n}|\tilde{\mathbb{E}}x_{i_1}\cdots x_{i_r}| \leqslant b^r k^r$ for all $1 \leqslant r \leqslant p-2$, then (6.1) implies that we can get a pseudodistribution satisfying

$$\|v^{\otimes p} - \tilde{\mathbb{E}}x^{\otimes p}\|_2^2 \ll 1\,,$$

from which we can recover v. Since we know that this is not likely to be possible if $k \leqslant n^{0.49}$, we can see that the assumption on $\|v\|_{\max}$ in Theorem 5.4 is likely to be inherent, at least for our SoS-based approach.

[27] This is mainly because of Fact C.1 which implies that the Gaussian complexity does not increase if we restrict to a subset of coordinates.

References

[AY21] Arnab Auddy and Ming Yuan, *On estimating rank-one spiked tensors in the presence of heavy tailed errors*, CoRR **abs/2107.09660** (2021). 3

[BB19] Matthew S. Brennan and Guy Bresler, *Optimal average-case reductions to sparse PCA: from weak assumptions to strong hardness*, Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA (Alina Beygelzimer and Daniel Hsu, eds.), Proceedings of Machine Learning Research, vol. 99, PMLR, 2019, pp. 469–470. 28

[BBH18] Matthew Brennan, Guy Bresler, and Wasim Huleihel, *Reducibility and computational lower bounds for problems with planted sparse structure*, Proceedings of the 31st Conference On Learning Theory (SÃľbastien Bubeck, Vianney Perchet, and Philippe Rigollet, eds.), Proceedings of Machine Learning Research, vol. 75, PMLR, 06–09 Jul 2018, pp. 48–166. 28

[BDJ$^+$22] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M. Kane, Pravesh K. Kothari, and Santosh S. Vempala, *Robustly learning mixtures of* k *arbitrary gaussians*, STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022, 2022, pp. 1234–1247. 1

[BJKK17] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar, *Consistent robust regression*, NIPS, 2017, pp. 2107–2116. 1

[BR13a] Quentin Berthet and Philippe Rigollet, *Complexity theoretic lower bounds for sparse principal component detection*, COLT, JMLR Workshop and Conference Proceedings, vol. 30, JMLR.org, 2013, pp. 1046–1066. 28

[BR13b] ——— , *Computational lower bounds for sparse PCA*, CoRR **abs/1304.0828** (2013). 28

[Cd22] Hongjie Chen and Tommaso d'Orsi, *On the well-spread property and its relation to linear regression*, Conference on Learning Theory, 2-5 July 2022, London, UK, 2022, pp. 3905–3935. 1

[CdO21] Davin Choo and Tommaso d'Orsi, *The complexity of sparse tensor pca*, Advances in Neural Information Processing Systems (M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, eds.), vol. 34, Curran Associates, Inc., 2021, pp. 7993–8005. 5

[CLMW11] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright, *Robust principal component analysis?*, J. ACM **58** (2011), no. 3, 11:1–11:37. 1

[DdNS21] Jingqiu Ding, Tommaso d'Orsi, Rajai Nasser, and David Steurer, *Robust recovery for stochastic block models*, 62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022, 2021, pp. 387–394. 1

[DHS20] Jingqiu Ding, Samuel B. Hopkins, and David Steurer, *Estimating rank-one spikes from heavy-tailed noise via self-avoiding walks*, Proceedings of the 34th International Conference on Neural Information Processing Systems (Red Hook, NY, USA), NIPS'20, Curran Associates Inc., 2020. 3

[dKNS20] Tommaso d'Orsi, Pravesh K. Kothari, Gleb Novikov, and David Steurer, *Sparse PCA: algorithms, adversarial perturbations and certificates*, 61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020, 2020, pp. 553–564. 1, 5, 10, 11, 25, 26

[DKWB19] Yunzi Ding, Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira, *Subexponential-time algorithms for sparse pca*, 2019. 5

[dLN$^+$21] Tommaso d'Orsi, Chih-Hung Liu, Rajai Nasser, Gleb Novikov, David Steurer, and Stefan Tiegel, *Consistent estimation for PCA and sparse regression with oblivious outliers*, Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, 2021, pp. 25427–25438. 1, 6, 7, 9, 43

[DM16] Yash Deshpande and Andrea Montanari, *Sparse PCA via covariance thresholding*, Journal of Machine Learning Research **17** (2016), 141:1–141:41. 5

[dNS21] Tommaso d'Orsi, Gleb Novikov, and David Steurer, *Consistent regression when oblivious outliers overwhelm*, 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, Proceedings of Machine Learning Research, PMLR, 2021. 1, 6

[FP07] Delphine Féral and Sandrine Péché, *The largest eigenvalue of rank one deformation of large wigner matrices*, Communications in mathematical physics **272** (2007), no. 1, 185–228. 5

[GLS81] M. Grötschel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica **1** (1981), no. 2, 169–197. MR 625550 13

[GMZ17] Chao Gao, Zongming Ma, and Harrison H. Zhou, *Sparse cca: Adaptive estimation and computational barriers*, The Annals of Statistics **45** (2017), no. 5, 2074–2101. 28

[HKP$^+$17] Samuel B. Hopkins, Pravesh K. Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer, *The power of sum-of-squares for detecting hidden structures*, 58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017, 2017, pp. 720–731. 3, 9

[Hop20] Samuel B Hopkins, *Mean estimation with sub-gaussian rates in polynomial time*, The Annals of Statistics **48** (2020), no. 2, 1193–1213. 1

[HSS15] Samuel B. Hopkins, Jonathan Shi, and David Steurer, *Tensor principal component analysis via sum-of-square proofs*, Proceedings of The 28th Conference on Learning Theory (Paris, France) (Peter GrÃijnwald, Elad Hazan, and Satyen Kale, eds.), Proceedings of Machine Learning Research, vol. 40, PMLR, 03–06 Jul 2015, pp. 956–1006. 3, 4, 9, 20, 36

[JL09] Iain M Johnstone and Arthur Yu Lu, *On consistency and sparsity for principal components analysis in high dimensions*, Journal of the American Statistical Association **104** (2009), no. 486, 682–693. 5

[KKM18] Adam R. Klivans, Pravesh K. Kothari, and Raghu Meka, *Efficient algorithms for outlier-robust regression*, Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018, 2018, pp. 1420–1430. 1

[KWB19] Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira, *Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio*, 2019. 3

[Las01] Jean B. Lasserre, *New positive semidefinite relaxations for nonconvex quadratic programs*, Advances in convex analysis and global optimization (Pythagorion, 2000), Nonconvex Optim. Appl., vol. 54, Kluwer Acad. Publ., Dordrecht, 2001, pp. 319–331. MR 1846160 13

[Lau09] Monique Laurent, *Sums of squares, moment matrices and optimization over polynomials*, 2009. 36, 37

[LZ20] Yuetian Luo and Anru R Zhang, *Open problem: Average-case hardness of hypergraphic planted clique detection*, Proceedings of Thirty Third Conference on Learning Theory (Jacob Abernethy and Shivani Agarwal, eds.), Proceedings of Machine Learning Research, vol. 125, PMLR, 09–12 Jul 2020, pp. 3852–3856. 30

[MR14] Andrea Montanari and Emile Richard, *A statistical model for tensor PCA*, CoRR **abs/1411.1076** (2014). 3

[MRS21] Pasin Manurangsi, Aviad Rubinstein, and Tselil Schramm, *The strongish planted clique hypothesis and its consequences*, 12th Innovations in Theoretical Computer Science Conference, ITCS 2021, January 6-8, 2021, Virtual Conference (James R. Lee, ed.), LIPIcs, vol. 185, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021, pp. 10:1–10:21. 11, 28

[Nes00] Yurii Nesterov, *Squared functional systems and optimization problems*, High performance optimization, Appl. Optim., vol. 33, Kluwer Acad. Publ., Dordrecht, 2000, pp. 405–440. MR 1748764 13

[Par00] Pablo A Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, Ph.D. thesis, California Institute of Technology, 2000. 13

[PJL20] Ankit Pensia, Varun Jog, and Po-Ling Loh, *Robust regression with covariate filtering: Heavy tails and adversarial contamination*, arXiv preprint arXiv:2009.12976 (2020). 1

[PWB20] Amelia Perry, Alexander S. Wein, and Afonso S. Bandeira, *Statistical limits of spiked tensor models*, Annales de l'Institut Henri PoincarÃl', ProbabilitÃl's et Statistiques **56** (2020), no. 1, 230 – 264. 8

[SBRJ19] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain, *Adaptive hard thresholding for near-optimal consistent robust regression*, Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA, 2019, pp. 2892–2897. 1

[Sho87] N. Z. Shor, *Quadratic optimization problems*, Izv. Akad. Nauk SSSR Tekhn. Kibernet. (1987), no. 1, 128–139, 222. MR 939596 13

[SZF20] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan, *Adaptive huber regression*, Journal of the American Statistical Association **115** (2020), no. 529, 254–265. 1

[TJSO14] Efthymios Tsakonas, Joakim Jaldén, Nicholas D Sidiropoulos, and Björn Ottersten, *Convergence of the huber regression m-estimate in the presence of dense outliers*, IEEE Signal Processing Letters **21** (2014), no. 10, 1211–1214. 1

[Wai19] Martin J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019. 35, 41, 42, 43

[WBS16] Tengyao Wang, Quentin Berthet, and Richard J. Samworth, *Statistical and computational trade-offs in estimation of sparse principal components*, The Annals of Statistics **44** (2016), no. 5, 1896 – 1930. 28

[ZLW+10] Zihan Zhou, Xiaodong Li, John Wright, Emmanuel J. Candès, and Yi Ma, *Stable principal component pursuit*, ISIT, IEEE, 2010, pp. 1518–1522. 1

A Additional tools

This section contain tools used throughout the rest of the paper.

Fact A.1. *[Wai19] Let $\boldsymbol{W}$ be an $n \times d$ random matrix with iid standard Gaussian entries. Then with probability at least $1 - \exp(-\tau/2)$, we have*

$$\|\boldsymbol{W}\| \leqslant \sqrt{n} + \sqrt{d} + \sqrt{\tau}\,.$$

Fact A.2. *[Wai19] Let B be a unit m-dimensional Euclidean ball. Then for every $\varepsilon \in (0,1)$ there exists an ε-net in B of size $(3/\varepsilon)^m$.*

Fact A.3. *Let $\boldsymbol{\xi}$ be a nonnegative random variable. Then $\mathbb{E}\,\boldsymbol{\xi} = \int_0^\infty \mathbb{P}[\boldsymbol{\xi} > \tau]d\tau$.*

Lemma A.4. *Let $\boldsymbol{\eta}$ be a random variable such that for some $a \in \mathbb{R}$ and for all $\tau > 0$, $\boldsymbol{\eta} \leqslant a + \tau$ with probability at least $1 - f(\tau)$ for some nonnegative $f \in \mathcal{L}_1((0,\infty))$. Then*

$$\mathbb{E}\,\boldsymbol{\eta} \leqslant a + \int_0^\infty f(\tau)d\tau\,.$$

Proof. Denote $\boldsymbol{\xi} = \mathbf{1}_{[\boldsymbol{\eta} \geqslant a]}\left(\boldsymbol{\eta} - a\right)$. Note that $\boldsymbol{\eta} \leqslant a + \boldsymbol{\xi}$ and $\boldsymbol{\xi}$ is nonnegative, hence by Fact A.3 we have

$$\mathbb{E}\,\boldsymbol{\xi} = \int_0^\infty \mathbb{P}[\boldsymbol{\xi} > \tau]d\tau = \int_0^\infty \mathbb{P}\left[\boldsymbol{\eta} - a > \tau\right]d\tau = \int_0^\infty f(\tau)d\tau < \infty\,.$$

Hence either $\mathbb{E}\,\boldsymbol{\eta} = -\infty$ and the bound is trivially satisfied, or we can take the expectations form both sides of $\boldsymbol{\eta} \leqslant a + \boldsymbol{\xi}$. □

B Sum-of-squares toolbox

B.1 Rounding

Lemma B.1. *Let $v \in \mathbb{R}^n$ be a unit vector. Let $p \geqslant 3$ be an odd number and let $\tilde{\mathbb{E}}$ be a pseudo-distribution over $\mathbb{R}[x_1, \ldots, x_n]$ of degree $t \geqslant p+1$ such that $\tilde{\mathbb{E}}\|x\|_2^2 = 1$. If for some $\varepsilon > 0$*

$$\left\|v^{\otimes p} - \tilde{\mathbb{E}}x^{\otimes p}\right\|_2 \leqslant \varepsilon\,,$$

then $\tilde{v} = \tilde{\mathbb{E}}x / \|\tilde{\mathbb{E}}x\|_2$ satisfies

$$\langle v, \tilde{v}\rangle \geqslant 1 - 2\varepsilon\,.$$

Proof. The proof is almost the same as the proof of Lemma 50 from [HSS15].

Consider the univariate polynomial $f(u) = 1 - 2u^p + u$. It is easy to verify that $f(u) \geqslant 0$ for all $u \in [-1, 1]$. Hence by Theorem 3.23 from [Lau09], f can be written as

$$f(u) = s_1(u)(1+u) + s_2(u)(1-u)\,,$$

where s_1 and s_2 are SoS polynomials of degree at most $p-1$.

Now consider $\tilde{\mathbb{E}}f(\langle v, x\rangle)$. Since $\|v\|_2^2 + \|x\|_2^2 \pm 2\langle v, x\rangle$ are SoS polynomials of degree 2 in variables $x_1, \ldots, x_n$, for every SoS polynomial s of degree at most $p-1$, we have

$$\left|\tilde{\mathbb{E}}[s(\langle v, x\rangle)\langle v, x\rangle]\right| \leqslant \frac{1}{2}\tilde{\mathbb{E}}\left[s(\langle v, x\rangle)\left(\|v\|_2^2 + \|x\|_2^2\right)\right] \leqslant \tilde{\mathbb{E}}[s(\langle v, x\rangle)]\,.$$

Hence

$$\tilde{\mathbb{E}}f(\langle v, x\rangle) = \tilde{\mathbb{E}}[s_1(\langle v, x\rangle)(1 + \langle v, x\rangle)] + \tilde{\mathbb{E}}[s_2(\langle v, x\rangle)(1 - \langle v, x\rangle)] \geqslant 0\,,$$

which implies that $\tilde{\mathbb{E}}\langle v, x\rangle \geqslant 2\tilde{\mathbb{E}}\langle v, x\rangle^p - 1$. On the other hand, since

$$\|\tilde{\mathbb{E}}x^{\otimes p}\|_2 \geqslant \|v^{\otimes p}\|_2 - \left\|v^{\otimes p} - \tilde{\mathbb{E}}x^{\otimes p}\right\|_2 \geqslant 1 - \varepsilon,$$

we have

$$\tilde{\mathbb{E}}\langle v, x\rangle^p = \frac{1}{2}\left(\|v\|_2^{2p} + \|\tilde{\mathbb{E}}x^{\otimes p}\|_2^2 - \|v^{\otimes p} - \tilde{\mathbb{E}}x^{\otimes p}\|_2^2\right) \geqslant 1 - \varepsilon\,.$$

We conclude that

$$\tilde{\mathbb{E}}\langle v, x\rangle \geqslant 2\tilde{\mathbb{E}}\langle v, x\rangle^p - 1 \geqslant 1 - 2\varepsilon\,.$$

□

Lemma B.2. *Let $v \in \mathbb{R}^n$ be a unit vector. Let $p \geqslant 2$ be an even number and let $\tilde{\mathbb{E}}$ be a pseudo-distribution over $\mathbb{R}[x_1, \ldots, x_n]$ of degree $t \geqslant p$ such that $\tilde{\mathbb{E}}\|x\|_2^2 = 1$. If for some $\varepsilon > 0$*

$$\left\|v^{\otimes p} - \tilde{\mathbb{E}}x^{\otimes p}\right\|_2 \leqslant \varepsilon\,,$$

then the top (unit) eigenvector $\tilde{v}$ of $\tilde{\mathbb{E}}xx^\top$ satisfies

$$\langle v, \tilde{v}\rangle^2 \geqslant 1 - 4\varepsilon\,.$$

Proof. Consider the polynomial $f(u) = 1 - 2u^p + u^2$. It is easy to verify that $f(u) \geqslant 0$ for all $u \in [-1, 1]$. Hence by Theorem 3.23 from [Lau09], f can be written as

$$f(u) = s_3(u) + s_4(u)(1 - u^2)\,,$$

where s_3 and s_4 are SoS polynomials satisfying $\deg(s_3) \leqslant p$ and $\deg(s_4) \leqslant p - 2$. It is easy to see that $\tilde{\mathbb{E}} f(\langle v, x\rangle) \geqslant 0$, and so $\tilde{\mathbb{E}}\langle v, x\rangle^2 \geqslant 2\tilde{\mathbb{E}}\langle v, x\rangle^p - 1$. On the other hand, since

$$\|\tilde{\mathbb{E}} x^{\otimes p}\|_2 \geqslant \|v^{\otimes p}\|_2 - \left\|v^{\otimes p} - \tilde{\mathbb{E}} x^{\otimes p}\right\|_2 \geqslant 1 - \varepsilon,$$

we have

$$\tilde{\mathbb{E}}\langle v, x\rangle^p = \frac{1}{2}\left(\|v\|_2^{2p} + \|\tilde{\mathbb{E}} x^{\otimes p}\|_2^2 - \|v^{\otimes p} - \tilde{\mathbb{E}} x^{\otimes p}\|_2^2\right) \geqslant 1 - \varepsilon\,.$$

Therefore,

$$v^\top\left(\tilde{\mathbb{E}} x x^\top\right)v = \tilde{\mathbb{E}}\langle v, x\rangle^2 \geqslant 2\tilde{\mathbb{E}}\langle v, x\rangle^p - 1 \geqslant 1 - 2\varepsilon\,.$$

Hence, by Fact B.3, the top eigenvector of $\tilde{\mathbb{E}} x x^\top$ satisfies the desired bound. □

Fact B.3. *Let $M \in \mathbb{R}^{d\times d}$ be such that $M \succeq 0$ and $\operatorname{Tr} M = 1$, and let $z \in \mathbb{R}^d$ be a unit vector such that $z^\mathsf{T} M z \geqslant 1 - \varepsilon$. Then the top eigenvector v_1 of M satisfies $\langle v_1, z\rangle^2 \geqslant 1 - 2\varepsilon$.*

Proof. Write $z = \alpha v_1 + \sqrt{1 - \alpha^2} v_\perp$ where $v_\perp$ is a unit vector orthogonal to v_1. Let $\lambda_1 \geqslant \ldots \geqslant \lambda_d \geqslant 0$ be the eigenvalues of M. We have

$$\begin{aligned} 1 - \varepsilon &\leqslant z^\mathsf{T} M z \\ &= \alpha^2 {v_1}^\mathsf{T} M v_1 + \left(1 - \alpha^2\right) {v_\perp}^\mathsf{T} M v_\perp \\ &= \alpha^2\left(\lambda_1 - {v_\perp}^\mathsf{T} M v_\perp\right) + {v_\perp}^\mathsf{T} M v_\perp \\ &\leqslant \alpha^2 + {v_\perp}^\mathsf{T} M v_\perp\,, \end{aligned}$$

where the last inequality follows from $M \succeq 0$ and $\operatorname{Tr} M = 1$, which imply that ${v_\perp}^\mathsf{T} M v_\perp \geqslant 0$ and $\lambda_1 \leqslant 1$.

Now since $M \preceq \lambda_1 I_d$, we have $\lambda_1 \geqslant z^\mathsf{T} M z \geqslant 1 - \varepsilon$, and

$$\lambda_2 + \ldots + \lambda_d = \operatorname{Tr} M - \lambda_1 \leqslant 1 - (1 - \varepsilon) = \varepsilon\,.$$

Therefore, ${v_\perp}^\mathsf{T} M v_\perp \leqslant \varepsilon$ and

$$\begin{aligned} \langle v_1, z\rangle^2 &= \alpha^2 \\ &\geqslant 1 - \varepsilon - {v_\perp}^\mathsf{T} M v_\perp \\ &\geqslant 1 - 2\varepsilon. \end{aligned}$$

□

B.2 Sum-of-squares certificates for sparse PCA

Let $t \in \mathbb{N}$ be such that $1 \leqslant t \leqslant k$ and let $\mathcal{S}_t$ be the set of all n-dimensional vectors with values in $\{0, 1\}$ that have exactly t nonzero coordinates.

We start with the following definition.

Definition B.4. For every $u \in \mathcal{S}_t$ we define the following polynomial in variables $s := (s_1, \dots, s_n)$

$$p_u(s) = \binom{k}{t}^{-1} \cdot \prod_{i \in \operatorname{supp}\{u\}} s_i \,.$$

Note that if x is a k-sparse vector and s is the indicator of its support, then for every $u \in \mathcal{S}_t$, we have

$$p_u(s) = \begin{cases} \binom{k}{t}^{-1} & \text{if } \operatorname{supp}\{u\} \subseteq \operatorname{supp}\{x\}\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Now consider the following system $\mathcal{C}_{s,x}$ of polynomial constraints.

$$\mathcal{C}_{s,x} \colon \left\{ \begin{array}{ll} \forall i \in [n], & s_i^2 = s_i \\ & \sum_{i \in [n]} s_i = k \\ \forall i \in [n], & s_i \cdot x_i = x_i \\ & \sum_{i \in [n]} x_i^2 = 1 \\ & \sum_{u \in \mathcal{S}_t} p_u(s) = 1 \\ \forall i \in [n], & \sum_{u \in \mathcal{S}_t} u_i p_u(s) = \frac{t}{k} \cdot s_i \end{array} \right\} \tag{B.1}$$

It is easy to see that if x is k-sparse and s is the indicator of its support, then x and s satisfy these constraints.

Lemma B.5. *Let $M \in \mathbb{R}^{n \times n}$ be a matrix. Denote by m_t the maximal spectral norm among all $2t \times 2t$ principal submatrices of M. Then*

$$\mathcal{C}_{s,x} \,\Big|\frac{s,x}{2t+2}\; \left\{ x^\top M x \leqslant 2 \cdot m_t \cdot k/t \right\}.$$

Proof. Without loss of generality we can assume that M is symmetric, since otherwise we can replace M by its symmetrisation. Indeed, $x^\top M x = x^\top \left(\frac{1}{2}M + \frac{1}{2}M^\top\right)x$, and symmetrisation does not increase the spectral norms of principal submatrices. Note that

$$\mathcal{C}_{s,x} \,\Big|\frac{s}{2t}\; \left\{ ss^\top = \frac{k^2}{t^2} \sum_{u,u' \in \mathcal{S}_t} u' u^\top p_{u'}(s) p_u(s) \right\}.$$

For $x, y \in \mathbb{R}^n$ we denote the Hadamard product of x and y as $x \odot y$, i.e., $x \odot y$ is the vector in $\mathbb{R}^n$ with entries $(x \odot y)_i = x_i \cdot y_i$ for all $i \in [n]$. It follows that

$$\mathcal{C}_{s,x} \,\Big|\frac{s,x}{4}\; \left\{ xx^\top = (x \odot s)(x \odot s)^\top \right\}$$

$$\begin{aligned}
&\vdash^{s,x}_{2t+2} \left\{ xx^{\mathsf{T}} = \frac{k^2}{t^2} \sum_{u,u' \in \mathcal{S}_t} (x \odot u')(x \odot u)^{\mathsf{T}} p_{u'}(s) p_u(s) \right\} \\
&\vdash^{s,x}_{2t+2} \left\{ x^{\mathsf{T}} M x = \frac{k^2}{t^2} \sum_{u,u' \in \mathcal{S}_t} (x \odot u)^{\mathsf{T}} M (x \odot u') p_{u'}(s) p_u(s) \right\} \\
&\vdash^{s,x}_{2t+2} \left\{ x^{\mathsf{T}} M x = \frac{k^2}{2t^2} \sum_{u,u' \in \mathcal{S}_t} \big((x \odot u)^{\mathsf{T}} M (x \odot u') + (x \odot u')^{\mathsf{T}} M (x \odot u) \big) p_{u'}(s) p_u(s) \right\} .
\end{aligned}$$

Now for every $u, u' \in \mathcal{S}_t$, let $M_{u,u'}$ be the matrix that coincides with M at the entries (i,j) such that both i and j are from the union of the supports of u and u', and is zero at other entries. Note that $(x \odot u)^{\top} M (x \odot u') = (x \odot u)^{\top} M_{u,u'} (x \odot u')$. Since $M_{u,u'}$ is symmetric, it is a difference of two PSD matrices $M^{+}_{u,u'}$ and $M^{-}_{u,u'}$ whose spectral norms are at most $\|M_{u,u'}\| \leqslant m_t$. Since for every PSD matrix S we have $\vdash^{a,b}_{2} \{\langle ab^{\top} + ba^{\top}, S\rangle \leqslant \langle aa^{\top} + bb^{\top}, S\rangle\}$ for variables $a, b \in \mathbb{R}^n$, we get

$$\begin{aligned}
\vdash^{x}_{2} \Bigg\{ & (x \odot u)^{\top} M_{u,u'} (x \odot u') + (x \odot u)^{\top} M_{u,u'} (x \odot u') \\
& = (x \odot u)^{\top} M^{+}_{u,u'} (x \odot u') + (x \odot u')^{\top} M^{+}_{u,u'} (x \odot u) \\
& \qquad\qquad + (-(x \odot u))^{\top} M^{-}_{u,u'} (x \odot u') + (x \odot u')^{\top} M^{-}_{u,u'} (-(x \odot u)) \\
& \leqslant (x \odot u)^{\top} M^{+}_{u,u'} (x \odot u) + (x \odot u')^{\top} M^{+}_{u,u'} (x \odot u') \\
& \qquad\qquad + (x \odot u)^{\top} M^{-}_{u,u'} (x \odot u) + (x \odot u')^{\top} M^{-}_{u,u'} (x \odot u') \\
& \leqslant 2 \cdot m_t \cdot \left(\|(x \odot u)\|^2 + \|(x \odot u')\|^2 \right) \Bigg\} .
\end{aligned}$$

Since $C_{s,x} \vdash^{s}_{t} \{p_u(s) \geqslant 0\}$, it follows that

$$\begin{aligned}
C_{s,x} &\vdash^{s,x}_{2t+2} \left\{ x^{\mathsf{T}} M x \leqslant \frac{k^2}{t^2} \sum_{u,u' \in \mathcal{S}_t} \left(m_t \cdot \|(x \odot u)\|^2 + m_t \cdot \|(x \odot u')\|^2 \right) p_{u'}(s) p_u(s) \right\} \\
&\vdash^{s,x}_{2t+2} \left\{ x^{\mathsf{T}} M x \leqslant m_t \tfrac{k^2}{t^2} \left(\sum_{u \in \mathcal{S}_t} \|(x \odot u)\|^2 p_u(s) \left(\sum_{u' \in \mathcal{S}_t} p_{u'}(s) \right) + \sum_{u' \in \mathcal{S}_t} \|(x \odot u')\|^2 p_{u'}(s) \left(\sum_{u \in \mathcal{S}_t} p_u(s) \right) \right) \right\} \\
&\vdash^{s,x}_{2t+2} \left\{ x^{\mathsf{T}} M x \leqslant m_t \frac{k^2}{t^2} \left(\sum_{u \in \mathcal{S}_t} \|(x \odot u)\|^2 p_u(s) + \sum_{u' \in \mathcal{S}_t} \|(x \odot u')\|^2 p_{u'}(s) \right) \right\} \\
&\vdash^{s,x}_{2t+2} \left\{ x^{\mathsf{T}} M x \leqslant 2 m_t \frac{k^2}{t^2} \sum_{u \in \mathcal{S}_t} \|(x \odot u)\|^2 p_u(s) \right\} .
\end{aligned}$$

Now observe that

$$\begin{aligned}
C_{s,x} &\vdash^{s,x}_{t+2} \left\{ \sum_{u \in \mathcal{S}_t} \|(x \odot u)\|^2 p_u(s) = \sum_{u \in \mathcal{S}_t} \sum_{i=1}^{n} x_i^2 u_i^2 \cdot p_u(s) \right\} \\
&\vdash^{s,x}_{t+2} \left\{ \sum_{u \in \mathcal{S}_t} \|(x \odot u)\|^2 p_u(s) = \sum_{i=1}^{n} x_i^2 \sum_{u \in \mathcal{S}_t} u_i \cdot p_u(s) \right\}
\end{aligned}$$

$$\Big|\!\frac{s,x}{t+2}\left\{\sum_{u\in\mathcal{S}_t}\|(x\odot u)\|^2 p_u(s)=\frac{t}{k}\sum_{i=1}^{n}x_i^2 s_i\right\}$$

$$\Big|\!\frac{s,x}{t+2}\left\{\sum_{u\in\mathcal{S}_t}\|(x\odot u)\|^2 p_u(s)=\frac{t}{k}\right\}.$$

□

Lemma B.6. *Suppose that $n \geqslant k \geqslant t \geqslant 2$. Let $\mathcal{P}_\ell$ be the set of all pseudo-distributions of degree ℓ on $\mathbb{R}[x,s]=\mathbb{R}[x_1,\dots,x_n,s_1,\dots,s_n]$ and let $\boldsymbol{W}\in\mathbb{R}^{n\times n}$ be a random matrix with i.i.d. standard Gaussian $N(0,1)$ entries. We have*

$$\mathbb{E}\left[\sup_{\substack{\mu\in\mathcal{P}_{2t+2}:\\ \mu\models\mathcal{C}_{s,x}}}\tilde{\mathbb{E}}_{x\sim\mu}x^\top\boldsymbol{W}x\right]\leqslant O\left(k\sqrt{\frac{\log n}{t}}\right).$$

Proof. Fix a pseudo-distribution $\mu\in\mathcal{P}_{2t+2}$ that satisfies $\mathcal{C}_{s,x}$. By Lemma B.5, we have

$$\tilde{\mathbb{E}}_{x\sim\mu}x^\top\boldsymbol{W}x\leqslant 2\cdot\boldsymbol{w}_t\cdot k/t\,,$$

where $\boldsymbol{w}_t$ is the maximal spectral norm among all $2t\times 2t$ principal submatrices of $\boldsymbol{W}$. Since this is true for every $\mu\in\mathcal{P}_{2t+2}$ satisfying $\mathcal{C}_{s,x}$, we get

$$\mathbb{E}\left[\sup_{\substack{\mu\in\mathcal{P}_{2t+2}:\\ \mu\models\mathcal{C}_{s,x}}}\tilde{\mathbb{E}}_{x\sim\mu}x^\top\boldsymbol{W}x\right]\leqslant 2\cdot\mathbb{E}[\boldsymbol{w}_t]\cdot k/t\,. \tag{B.2}$$

For every $A\subseteq[n]$, let $\boldsymbol{W}_A$ be the principal submatrix of $\boldsymbol{W}$ that is obtained by taking the rows and columns with indices in A. By Fact A.1, we know that for every fixed $A\subseteq[n]$ of size $2t$ and every $0<\delta'<1$, we have

$$\|\boldsymbol{W}_A\|\leqslant\sqrt{2t}+\sqrt{2t}+\sqrt{2\log(1/\delta')}\,,$$

with probability at least $1-\delta'$. By taking a union bound over all $\binom{n}{2t}$ subsets $A\subseteq[n]$ of size $2t$, we can see that for every $0<\delta<1$, the following holds with probability at least $1-\delta$:

$$\begin{aligned}\boldsymbol{w}_t &= \max_{\substack{A\subseteq[n]:\\ |A|=2t}}\|\boldsymbol{W}_A\|\leqslant 2\sqrt{2t}+\sqrt{2\log\left(\frac{\binom{n}{2t}}{\delta}\right)}\leqslant 2\sqrt{2t}+\sqrt{2\log\left(\frac{\left(\frac{ne}{2t}\right)^{2t}}{\delta}\right)}\\ &= 2\sqrt{2t}+\sqrt{4t\log(n)+4t-4t\log(2t)+2\log(1/\delta)}\\ &\leqslant 10\sqrt{t\log(n)}+\sqrt{2\log(1/\delta)}\,.\end{aligned}$$

In other words, for every $\tau>0$, with probability at least $1-\exp(-\tau^2/2)$, we have

$$\boldsymbol{w}_t\leqslant 10\sqrt{t\log(n)}+\tau\,.$$

By applying Lemma A.4, we get

$$\mathbb{E}[\boldsymbol{w}_t] \leqslant 10\sqrt{t\log(n)} + \int_0^\infty \exp(-\tau^2/2)d\tau \leqslant 10\sqrt{t\log(n)} + O(1)\,.$$

Combining this with (B.2), we get the result.

□

C Facts about Gaussian and Rademacher complexities

Recall that for a bounded set $A \in \mathbb{R}^m$, the Gaussian complexity $\mathcal{G}(A)$ is defined as

$$\mathcal{G}(A) = \mathop{\mathbb{E}}_{\boldsymbol{w}\sim N(0,\mathrm{Id}_m)} \sup_{a\in A} \sum_{i=1}^m a_i \boldsymbol{w}_i\,,$$

and Rademacher complexity $\mathcal{R}(A)$ is defined as

$$\mathcal{R}(A) = \mathop{\mathbb{E}}_{\boldsymbol{s}\sim U(\{\pm 1\}^m)} \sup_{a\in A} \sum_{i=1}^m a_i \boldsymbol{s}_i\,,$$

where $U(\{\pm 1\}^m)$ is the uniform distribution over $\{+1,-1\}^m$.

Fact C.1 ([Wai19], Proposition 5.28)**.** *Let $A \subset \mathbb{R}^m$ be a bounded set, and let $\phi_1,\ldots,\phi_m : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz functions that satisfy $\phi_i(0) = 0$ for all $i \in [m]$. Denote $\phi : \mathbb{R}^m \to \mathbb{R}^m$, $\phi(x_1,\ldots,x_m) = (\phi_1(x_1),\ldots,\phi_m(x_m))$. Then*

$$\mathcal{G}(\phi(A)) \leqslant \mathcal{G}(A) \quad \textit{and} \quad \mathcal{R}(\phi(A)) \leqslant 2\mathcal{R}(A)\,.$$

Fact C.2. *[Wai19] For every bounded set $A \subset \mathbb{R}^m$,*

$$\mathcal{R}(A) \leqslant \sqrt{\pi/2}\cdot\mathcal{G}(A)\,.$$

Proof. Let $\boldsymbol{w}_1,\ldots \boldsymbol{w}_m$ be iid standard Gaussian variables. Denote $\boldsymbol{s}_i = \operatorname{sign}(\boldsymbol{w}_i)$ and $\boldsymbol{z}_i = |\boldsymbol{w}_i|$. Since $\boldsymbol{w}_i$ are symmetric, $\boldsymbol{s}_i$ and $\boldsymbol{z}_i$ are independent. Since $\mathbb{E}\,\boldsymbol{z}_i = \sqrt{2/\pi}$,

$$\mathcal{G}(A) = \mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^m a_i\boldsymbol{w}_i\right] = \mathbb{E}\,\mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^m a_i\boldsymbol{z}_i\boldsymbol{s}_i \;\middle|\; \boldsymbol{s}\right] \geqslant \sqrt{2/\pi}\,\mathbb{E}\left[\sup_{a\in A}\sum_{i=1}^m a_i\boldsymbol{s}_i\right] = \mathcal{R}(A)\,.$$

□

Fact C.3 ([Wai19], Theorem 3.4)**.** *Let $m \in \mathbb{N}$, $h > 0$ and let $\xi_1,\ldots,\xi_m$ be independent random variables such that $\forall i \in [m]$, $|\xi_i| \leqslant h$ with probability 1. Let $L > 0$ and let $f : \mathbb{R}^m \to \mathbb{R}$ be a convex L-Lipschitz function. Then for all $t > 0$,*

$$\mathbb{P}\big[f(\xi_1,\ldots,\xi_m) \geqslant \mathbb{E}\,f(\xi_1,\ldots,\xi_m) + t\big] \leqslant \exp\left(-\frac{t^2}{16\cdot L^2\cdot h^2}\right).$$

Lemma C.4. *Let $m \in \mathbb{N}$, $h > 0$ and let $\xi_1, \ldots, \xi_m$ be independent, symmetric about zero random variables such that $\forall i \in [m]$, $|\xi_i| \leqslant h$ with probability 1. Let $A \subset \mathbb{R}^m$ be a bounded set and denote $r_A = \sup_{a \in A} \|a\|_2$. Let*

$$S_A = \sup_{a \in A} \sum_{i=1}^{m} a_i \xi_i \,.$$

Then

$$\mathbb{E}\, S_A \leqslant 2 \cdot h \cdot \mathcal{R}(A) \leqslant 3 \cdot h \cdot \mathcal{G}(A)\,,$$

and for all $t > 0$,

$$\mathbb{P}[S_A \geqslant \mathbb{E}\, S_A + t] \leqslant \exp\left(-\frac{t^2}{16 \cdot r_A^2 \cdot h^2}\right).$$

Proof. Let us first show the concentration bound. Consider the function $f : \mathbb{R}^m \to \mathbb{R}$ defined as $f(x) = \sup_{a \in A} \langle x, a \rangle$. It is a convex function (as the supremum of convex functions), and r_A-Lipschitz since for all $x, y \in \mathbb{R}^m$,

$$\langle x, a \rangle - f(y) \leqslant \langle x - y, a \rangle \leqslant \|a\|_2 \|x - y\|_2\,,$$

and if we take sup over $a \in A$, we get $f(x) - f(y) \leqslant r_A \|x - y\|_2$. The desired bound follows from Fact C.3.

Now let us bound the expectation. Denote $s_i = \operatorname{sign}(\xi_i)$ and $\eta_i = \frac{1}{h}|\xi_i|$ so that $\xi_i = h \cdot s_i \cdot \eta_i$. Since ξ_i are symmetric, s_i and η_i are independent. We have

$$\mathbb{E}\left[\sup_{a \in A} \sum_{i=1}^{m} a_i \xi_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sup_{a \in A} \sum_{i=1}^{m} a_i \cdot h \cdot \eta_i s_i \,\middle|\, \eta\right]\right] = h \cdot \mathbb{E}\left[\mathbb{E}\left[\sup_{a \in A} \sum_{i=1}^{m} \phi_i(a_i) \cdot s_i \,\middle|\, \eta\right]\right],$$

where $\phi_i : \mathbb{R} \to \mathbb{R}$ is defined as $\phi_i(x_i) = \eta_i x_i$. Since $0 \leqslant \eta_i \leqslant 1$ for all $i \in [m]$, we can see that $\phi_1, \ldots, \phi_m$ are all 1-Lipschitz. It follows fromFact C.1 that

$$\mathbb{E}\left[\sup_{a \in A} \sum_{i=1}^{m} \phi_i(a_i) \cdot s_i \,\middle|\, \eta\right] \leqslant 2\mathcal{R}(A)\,,$$

and hence

$$\mathbb{E}\, S_A \leqslant 2 \cdot h \cdot \mathcal{R}(A) \leqslant 3 \cdot h \cdot \mathcal{G}(A)\,,$$

where the last inequality follows from Fact C.2. □

Fact C.5 (Sudakov's Minoration, [Wai19], Theorem 5.30)**.** *Let $A \subset \mathbb{R}^m$ be a bounded set. Then*

$$\sup_{\varepsilon > 0} \frac{\varepsilon}{2} \sqrt{\log |\mathcal{N}_\varepsilon(A)|} \leqslant \mathcal{G}(A)\,,$$

where $|\mathcal{N}_\varepsilon(A)|$ is the minimal size of ε-net in A with respect to Euclidean distance.

D Decomposability

In [dLN+21] the Huber loss minimization was studied with ℓ_1-norm and nuclear norm regularizers. The authors of [dLN+21] used a well-known property of these norms that is called *decomposability* (this property has been extensively used in the literature, see [Wai19]).

A norm $\|\cdot\|_\circ$ in $\mathbb{R}^m$ is said to be *decomposable* with respect to a pair of vector subspaces $\left(V, \bar{V}\right)$ such that $V \subseteq \bar{V}$, if for all $v \in V, u \in \bar{V}^\perp$, we have $\|v+u\|_\circ = \|v\|_\circ + \|u\|_\circ$. In the context of PCA, where the signal is rank-one symmetric matrix $X^* = \lambda \cdot vv^\top$, and if $X^* \in V$ and both V and $\bar{V}$ contain only rank $O(1)$ matrices, then decomposability of the nuclear norm implies $\mathbf{\Delta} \in \{M : \|M\|_{\text{nuc}} \leqslant O(\|M\|_{\text{F}})\}$. This fact can be used to get a good bound on the error. It is easy to see that for the nuclear norm there are natural spaces V and $\bar{V}$ that satisfy these properties: $V = \operatorname{span}\{vv^\top\}$ and $\bar{V} = \operatorname{span}\{uv^\top + vu^\top : u \in \mathbb{R}^m\}$.

Now assume that the signal is a tensor $X^* = \lambda \cdot v^{\otimes 3}$. One could try to apply the same approach for tensors: That is, to minimize the Huber loss with the dual norm of the injective tensor norm as a regularizer (let us ignore in this discussion computational aspects of the problem for simplicity). Recall that the injective tensor norm of order 3 symmetric tensor T is defined as

$$\|T\|_{\text{inj}} := \sup_{\|x\|_2=1} \langle x^{\otimes 3}, T\rangle \,.$$

Let us check whether its dual norm $\|\cdot\|_{\text{inj}}^*$ is decomposable with respect to some natural subspaces of V and $\bar{V}$ of low-rank tensors. The choice of V is simple: it is always better if it is as small as possible, so we should just take $V = \operatorname{span}\{v^{\otimes 3}\}$. And there are two candidates for $\bar{V}$ similar to the corresponding subspace in the matrix case: $\bar{V}_1 = \operatorname{span}\{\operatorname{Sym}(u \otimes u \otimes v) : u \in \mathbb{R}^m\}$ and $\bar{V}_2 = \operatorname{span}\{\operatorname{Sym}(u \otimes w \otimes v) : u, w \in \mathbb{R}^m\}$.

$\bar{V}_2$ is not a good choice: It contains some tensors of rank n, and hence we cannot a get good bound if we use it (additional $\sqrt{n}$ factor appears in the error bound if we use it).

Let us now show that $\bar{V}_1$ is also not a good choice: The norm $\|\cdot\|_{\text{inj}}^*$ is not decomposable with respect to $\left(V, \bar{V}_1\right)$. Indeed, for unit vectors $u, w \in \mathbb{R}^m$ such that all u, v, w are orthogonal to each other, the tensor $S = \operatorname{Sym}(v \otimes u \otimes w)$ is in $\bar{V}^\perp$. By definition of $\|\cdot\|_{\text{inj}}^*$,

$$\|v^{\otimes 3} + S\|_{\text{inj}}^* = \max_{\|T\|_{\text{inj}} \leqslant 1} \langle T, v^{\otimes 3} + S\rangle.$$

We can assume without loss of generality that $T = \lambda v^{\otimes 3} + \mu S$ for some $\lambda, \mu \in \mathbb{R}$. Then $\langle T, v^{\otimes 3} + S\rangle \leqslant \lambda + \mu/6$.

Note that $|\lambda| \leqslant 1$, otherwise $\|T\|_{\text{inj}} > 1$. Now consider

$$x = \frac{1}{\sqrt{3}}\left(\operatorname{sign}(\mu) \cdot u + \operatorname{sign}(\lambda) \cdot v + \operatorname{sign}(\lambda) \cdot w\right),$$

and note that

$$\langle x^{\otimes 3}, T\rangle = 3^{-3/2}|\lambda| + 3^{-3/2}|\mu| \leqslant 1\,.$$

The maximal value of the linear function $\lambda + \mu/6$ on the polygon

$$\left\{(\lambda, \mu) \in \mathbb{R}^2 : |\lambda| \leqslant 1\,,\ |\lambda| + |\mu| \leqslant 3^{3/2}\right\}$$

is achieved at one of the vertices of this polygon, hence

$$\lambda + \mu/6 \leqslant \max\left\{\sqrt{3}/2, 1 + \sqrt{3}/2 - 1/6\right\} \leqslant 5/6 + \sqrt{3}/2\,,$$

and $\|v^{\otimes 3} + S\|_{\text{inj}}^{*} \leqslant 5/6 + \sqrt{3}/2$.

It is easy to verify that $\|v^{\otimes 3}\|_{\text{inj}}^{*} = 1$ and $\|S\|_{\text{inj}}^{*} = \sqrt{3}/2$, hence

$$\|v^{\otimes 3} + S\|_{\text{inj}}^{*} < \|v^{\otimes 3}\|_{\text{inj}}^{*} + \|S\|_{\text{inj}}^{*}$$

and the norm $\|\cdot\|_{\text{inj}}^{*}$ is not decomposable with respect to $\left(V, \bar{V}_1\right)$.

This shows that naive approach fails and either we need to look for other sets $\bar{V}$ and try to prove decomposability for them, or not to use decomposability at all. We show that decomposability is not necessary for obtaining vanishing error[28], and hence we can study Huber loss minimization over more complicated sets than nuclear norm ball or ℓ_1 ball.

[28] The only advantage of the analysis that uses decomposability compared to our approach is that it guarantees better error convergence: when the decomposability guarantees the error bound $O(\varepsilon)$, our analysis guarantees the bound $O(\sqrt{\varepsilon})$.