Computational-Statistical Gaps for Improper Learning in Sparse Linear Regression

Rares-Darius Buhai RARES.BUHAI@INF.ETHZ.CH
Jingqiu Ding JINGQIU.DING@INF.ETHZ.CH
Stefan Tiegel STEFAN.TIEGEL@INF.ETHZ.CH
ETH Zürich

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study computational-statistical gaps for improper learning in sparse linear regression. More specifically, given n samples from a k-sparse linear model in dimension d, we ask what is the minimum sample complexity to efficiently (in time polynomial in d, k, and n) find a potentially dense estimate for the regression vector that achieves non-trivial prediction error on the n samples. Information-theoretically this can be achieved using $\Theta(k \log(d/k))$ samples. Yet, despite its prominence in the literature, there is no polynomial-time algorithm known to achieve the same guarantees using less than $\Theta(d)$ samples without additional restrictions on the model. Similarly, existing hardness results are either restricted to the proper setting, in which the estimate must be sparse as well, or only apply to specific algorithms.

We give evidence that efficient algorithms for this task require at least (roughly) $\Omega(k^2)$ samples. In particular, we show that an improper learning algorithm for sparse linear regression can be used to solve sparse PCA problems (with a negative spike) in their Wishart form, in regimes in which efficient algorithms are widely believed to require at least $\Omega(k^2)$ samples. We complement our reduction with low-degree and statistical query lower bounds for the sparse PCA problems from which we reduce.

Our hardness results apply to the (correlated) random design setting in which the covariates are drawn i.i.d. from a mean-zero Gaussian distribution with unknown covariance.

Keywords: computational-statistical gaps, sparse linear regression, reduction-based hardness

1. Introduction

We study computational-statistical gaps in sparse linear regression models with (correlated) random designs. In particular, on receiving n samples $(\boldsymbol{a}_i, \boldsymbol{y}_i)$ for $\boldsymbol{y}_i = \langle \boldsymbol{a}_i, x^* \rangle + \boldsymbol{w}_i$ where $\boldsymbol{w}_i$ is Gaussian noise and x^* is an unknown sparse vector, we wish to find an estimate $\hat{\boldsymbol{x}}$ for x^* that produces predictions $\langle \boldsymbol{a}_i, \hat{\boldsymbol{x}} \rangle$ close to $\langle \boldsymbol{a}_i, x^* \rangle$ for the given samples. We consider the setting in which the covariates $\boldsymbol{a}_i$ are drawn i.i.d. from a (correlated) Gaussian distribution (cf. model 1) – this is referred to as the *random design* setting.

Despite its prominence, the computational complexity of sparse linear regression with respect to general estimators remains poorly understood. Without further restrictions on the model, no efficient algorithms are known that use $o(d)$ samples. Similarly, there is a dearth of hardness results: Known lower bounds apply only against specific algorithms or rule out so-called *proper* learners, in which the output of the estimator needs to be sparse as well. In particular, to the best of our knowledge, there is no evidence ruling out efficient algorithms producing potentially dense estimates that use only the information-theoretically minimal number of samples. We call such estimators *improper*. The distinction between proper and improper learners can be substantial: There are learning tasks that are NP-hard in the proper setting but polynomial-time solvable using an improper learner Valiant (1984); Pitt and Valiant (1988).[1] Thus, hardness against proper learners only gives limited evidence for the hardness of a learning task.

Similarly, many of the known lower bounds only apply to the more stringent setting in which the covariates are allowed to be worst-case (worst-case designs) Zhang et al. (2014, 2015). A priori the random design setting may seem much more benign than the worst-case setting. Computational hardness results for worst-case designs, albeit restricted to proper learners, have been known for nearly a decade Zhang et al. (2014), yet obtaining any computational lower bounds for random Gaussian designs is a major open question explored in Kelner et al. (2022b,a). In this work, we show hardness results for improper learners that hold even in the restrictive random design setting.

More specifically, we consider the following model:

Model 1 (Sparse linear regression with Gaussian design) *Let $d, k, n \in \mathbb{N}$ and $\sigma \in \mathbb{R}_{\geq 0}$ be known. Let Σ be an unknown positive semi-definite matrix and $x^* \in \mathbb{R}^d$ be some unknown k-sparse vector, i.e., that has at most k non-zero entries. We draw n i.i.d. samples $(\boldsymbol{a}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{a}_n, \boldsymbol{y}_n) \in \mathbb{R}^d \times \mathbb{R}$ from the following linear model: Independently draw $\boldsymbol{a}_i \sim N(0, \Sigma)$ and $\boldsymbol{w}_i \sim N(0, \sigma^2)$. Output the $(\boldsymbol{a}_i, \boldsymbol{y}_i)$, where $\boldsymbol{y}_i = \langle \boldsymbol{a}_i, x^* \rangle + \boldsymbol{w}_i$.*

Let $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ be the matrix that has the $\boldsymbol{a}_i$ as rows and let $\boldsymbol{y} \in \mathbb{R}^n$ be the vector that has entries $\boldsymbol{y}_i$. We say an algorithm achieves prediction error ρ *if it outputs a (potentially dense) vector $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ such that $\frac{1}{n}\|\boldsymbol{A}x^* - \boldsymbol{A}\hat{\boldsymbol{x}}\|^2 \leq \rho$.*

We refer to the matrix $\boldsymbol{A}$ as the design matrix. In what follows, we will focus on algorithms achieving prediction error $\rho = 0.01$.

Information-theoretically, this problem is well-understood: $\Theta(k \log(d/k))$ samples are both necessary and sufficient to achieve prediction error 0.01 with, say, probability 0.99 Raskutti et al. (2011). Unfortunately, the algorithm achieving the upper bound – known as the best-subset-selection estimator – relies on an exhaustive search over k-sparse vectors and requires time $d^{\Omega(k)}$. This exponential-in-k running time is prohibitively large as soon as k is only slightly larger than constant. Instead, we would like algorithms running in time $\mathrm{poly}(d, n, k)$. We call such algorithms efficient. All efficient algorithms for this task use $\Omega(d)$ samples, yet there is no hardness result giving evidence that even $\omega(k \log(d/k))$ samples are necessary. This leads us to the main question of this work:

Question 2 *What is the best-possible sample complexity for computationally efficient algorithms in Definition 1 achieving prediction error 0.01?*

We make progress on this question by presenting evidence that efficient algorithms need $\Omega(k^2)$ samples, even in the improper setting. This leads to an intriguing state of affairs: Many problems involving sparsity exhibit a k-vs-k^2 statistical-computational gap. That is, information-theoretically the problem can be solved with (roughly) k samples, yet computationally efficient algorithms likely require k^2 samples (we refer to Brennan and Bresler (2020) and the references therein for a more detailed treatment). Our results provide evidence that this might also be the case for sparse linear

[1] The learning tasks are 3-Term DNFs that are known to be efficiently learnable via 3-CNFs yet are NP-hard to learn properly.

regression.[2] Our hardness results come in the form of reductions from problems that are believed to be computationally intractable.

Known lower bounds and our approach Known lower bounds have focused on showing impossibility results for the harder task of (semi-)proper learning in which the algorithm has to output a k'-sparse estimate for $k \leq k' \leq k \cdot d^{o(1)}$. Based on worst-case complexity assumptions, Zhang et al. (2014); Foster et al. (2015) show that for worst-case design matrices polynomial-time algorithms require $\omega(k \log d)$ samples. While this settles Question 2 for (semi-)proper learners, this does not say anything about the improper and/or random design case. To the best of our knowledge, the only known lower bounds for this more general setting apply to (very) restricted families of algorithms Zhang et al. (2015); Kelner et al. (2022b,a). While these include popular sparse regression algorithms such as LASSO and variations thereof, this only provides weak evidence of hardness: It remains unclear if an algorithm from a different family can circumvent the results. In particular, in recent years algorithms based on spectral methods and the sum-of-squares hierarchy have seen an immense success in the design of computationally efficient algorithms for statistical estimation problems. Such a general class of algorithms is not ruled out.

This calls for hardness evidence against more general classes of algorithms. Unfortunately, and in contrast to the proper learning setting, there seem to be inherent barriers to basing hardness of improper learning (in average-case problems) on classical worst-case assumptions such as P $\neq$ NP Applebaum et al. (2008). By now, the two most prevalent techniques for showing average-case hardness are the following: First, show unconditional lower bounds in a restricted model of computation (that captures most known algorithms for the problem), such as statistical query (SQ) Kearns (1998); Feldman et al. (2017) or low-degree algorithms Hopkins and Steurer (2017); Hopkins et al. (2017); Hopkins (2018). These algorithms capture most known algorithms for statistical inference problems (with the notable exception of Gaussian elimination and the LLL Algorithm Zadik et al. (2022); Diakonikolas and Kane (2022)). For many average-case problems (such as graph recovery problems Barak et al. (2019); Hopkins and Steurer (2017), mean estimation and learning Gaussian Mixture Models Diakonikolas et al. (2017) and spiked matrix models Hopkins et al. (2017); Ding et al. (2022)), these frameworks have been successfully used to trace out a computational phase transition for a broad class of computational problems: the statistical error rate achieved by existing polynomial-time algorithms can be matched by algorithms in the class, while obtaining better error rate is impossible for these algorithms. The second approach is to show a reduction from a problem believed to be computationally hard Brennan and Bresler (2020); Bruna et al. (2021); Gupte et al. (2022). We remark that the latter can also be interpreted in a positive way as understanding the connections between different computational problems: If, say, we find a better sparse linear regression algorithm, we also find a better algorithm for another problem that we did not know before.

We follow the second approach. In particular, we show a reduction from a slight variant of a widely studied PCA problem:

Model 3 (Sparse spiked Wishart model) *Let $d, k, n \in \mathbb{N}$ with $d \geq k$ and $\theta \in (0, 1)$. We define the $\mathrm{NegSPCA}_\theta$ problem as the following distinguishing problem: We are given n i.i.d. samples from either $\mathcal{P}$ or $\mathcal{Q}$ defined as follows and want to decide whether they came from $\mathcal{P}$ or $\mathcal{Q}$.*

[2] However, we remark that for sparse linear regression there is no known algorithm using $O(k^2)$ or even $o(d)$ samples.

- ***Planted distribution*** *$\mathcal{P}$: Sample a unit vector $\boldsymbol{x} \in \mathbb{R}^{d+1}$ with $\boldsymbol{x}_1 = -\frac{1}{\sqrt{k+1}}$ and $\boldsymbol{x}_{\setminus 1}$ sampled uniformly from k-sparse $\left\{\pm\frac{1}{\sqrt{k+1}}, 0\right\}^d$ vectors and fix it for the rest of the sampling procedure. Produce samples by sampling from $N(0, \mathrm{Id}_{d+1} - \theta \cdot \boldsymbol{x}\boldsymbol{x}^\top)$.*
- ***Null distribution*** *$\mathcal{Q}$: $N(0, \mathrm{Id}_{d+1})$.*

We remark that, except for the condition that the first coordinate of $\boldsymbol{x}$ under $\mathcal{P}$ is non-zero, this coincides with the classical problem of sparse PCA in the Wishart model with a negative spike that has a sparse prior. Several variants of this problem are known to be hard using (roughly) less than k^2 samples: in the low-degree model, inside the sum-of-squares framework, and under a reduction from a variant of the planted clique assumption Hopkins et al. (2017); Ding et al. (2021); Brennan and Bresler (2020). We make the following hardness assumption about Definition 3, which we formally verify in the low-degree and statistical query model in Section **??** (our proof follows closely the proof of Ding et al. (2021)).

Assumption 4 *Let $d, n, k \in \mathbb{N}$ with $d \geq k$ and let $0 < \delta \leq 0.1$ be an arbitrary absolute constant. Suppose that $n = o(\min(d, k^{2-\delta}))$. Then, for any $\theta \in (0, 1)$, any algorithm that solves $\mathrm{NegSPCA}_\theta$ in dimension d using n samples and has success probability at least $1 - o(1)$ requires running time $d^{k^{\Omega(1)}}$.*

Results We show that a polynomial-time algorithm for improper learning in sparse linear regression that uses n samples can be turned into a polynomial-time algorithm for Definition 3 using the same number of samples. In particular, our main result is the following:

Theorem 5 (Main result, see reduction in Theorem 8) *Let $d, n, k \in \mathbb{N}$ with $d \geq k$ and let $0 < \delta \leq 0.1$ be an arbitrary absolute constant. Suppose that $n = o(\min(d, k^{2-\delta}))$. If there is an improper learner for the Sparse Linear Regression Model with Gaussian Design (cf. Definition 1) that uses n samples and runs in time $d^{k^{o(1)}}$ and achieves prediction error better than 0.01 with probability at least $1 - o(1)$, then Conjecture 4 is false.*

Note that this result implies that efficient algorithms for improperly learning sparse linear regression models likely need (roughly) $\Omega(\min(k^2, d))$ samples, whereas information-theoretically, only $O(k \log(d/k))$ samples are necessary. Without additional assumptions on the model (cf. Section 1.1) no efficient algorithm is known that uses $o(d)$ samples for this task. We remark that when we only want to rule out algorithms that achieve prediction error 0.01 with probability at least $1 - \Omega(\frac{1}{d})$, we can reduce from a version of Definition 3 in which $\boldsymbol{x}$ is sampled uniformly from all $(k+1)$-sparse vectors in $\left\{\pm\frac{1}{\sqrt{k+1}}, 0\right\}$ (without the first coordinate being known).[3]

It turns out that LASSO can solve our hard instance with $O(k^2 \log d)$ samples, so our lower bound is tight up to logarithmic factors. Therefore a different construction would be needed to obtain even stronger lower bounds.

[3] We remark that we cannot combine this last reduction with the reductions in Brennan and Bresler (2020) to obtain a reduction from (secret-leakage) planted clique. The reason is that the parameters in the reduction from the latter to Definition 3 are not sufficiently strong. Specifically, Brennan and Bresler (2020) requires $\theta = o(1)$, while we need θ to be an absolute constant close to 1.

1.1. Relation to known algorithms and other hardness results

We outline here how our result compares to other algorithmic and hardness results that have been obtained in prior works.

As stated before, known algorithms achieve prediction error 0.01 using $o(d)$ samples only with additional assumptions on the model. In particular, under the assumption that the columns are normalized to have squared norm n, the LASSO estimator is known to achieve prediction error 0.01 using $O(\|x^*\|_1^2 \log d)$ samples (Wainwright, 2019, Theorem 7.20). This gives a good sample complexity if the ℓ_1-norm of the secret is bounded. Second, if the design matrix is known to satisfy the so-called *restricted eigenvalue condition* (with the same normalization of the columns), LASSO is known to succeed with few samples. In particular, let A be the design matrix. Assuming that (roughly) for all k-sparse unit vectors u it holds that $\frac{1}{n}\|Au\|^2 \geq \gamma$, LASSO achieves prediction error 0.01 using $O(\frac{k \log d}{\gamma})$ samples. γ is referred to as the RE constant.

Further, Kelner et al. (2022b) showed that in the Gaussian design setting that we consider in this paper, the LASSO estimator combined with a preconditioning step achieves nearly optimal sample complexity whenever the dependency structure of the covariates satisfies a specific regularity condition.

For lower bounds, the results closest to us are the following: Zhang et al. (2014) showed that, assuming $\mathrm{NP} \not\subseteq \mathrm{P}_{/\,\mathrm{poly}}$, the dependence of the RE constant achieved by the LASSO is tight for all polynomial-time proper learners. In particular, they construct instance in which the RE constant can be exponentially small and obtain a corresponding lower bound on the sample complexity of proper learners achieving prediction error 0.01. Under other complexity theoretic assumptions Foster et al. (2015) showed that there is no polynomial-time estimator for this task, even if the output vector is allowed to be $O(k \cdot 2^{\log^{1-\delta}(d)})$-sparse for any constant $\delta > 0$. Note that this notion is still significantly more restrictive than ours, as it does note rule out outputting even a $d^{0.001}$-sparse vector. The design matrices in both the above works are worst-case.

Zhang et al. (2015) showed that a restricted class of convex M-estimators (including the LASSO estimator) cannot achieve optimal prediction error (under worst-case designs). While the algorithms they rule out are improper, their result does not make any claims beyond the specific algorithms they consider. To the best of our knowledge, the only hardness results for (a restricted class of) improper learners in the random design setting are Kelner et al. (2022b,a). They ruled out that the LASSO can achieve optimal sample complexity even when combined with a pre-processing step (called pre-conditioning).

Concurrent work Concurrent and independent work of Gupte et al. (2024) shows hardness for proper learning under random designs based on worst-case problems in lattices. In particular, they give evidence that in this setting, improving the prediction error bound of LASSO, when expressing the error as a function of the RE constant of the design matrix, would give stronger lattice algorithms. As the information-theoretic prediction error has no dependence on the RE constant and the RE constant can be exponentially small, this gives evidence of a possibly large computational-statistical separation.

Most closely to our result, concurrent and independent work of Kelner et al. (2024) has also observed the connection between negative sparse PCA and sparse linear regression that we make in this paper. Using this, they deduce a similar k-to-k^2 gap for sparse linear regression. Their lower bound holds for the *generalization error* in the (correlated) random design setting, and their reduction is similar to the unknown-variance reduction discussed in the first part of Section 2 in our work. Our result in Section 3 establishes hardness in this setting for the *training error* (i.e., the error on the observed samples), which is often an easier quantity to minimize than the generalization error. We remark that Kelner et al. (2024) also contains algorithmic upper bounds under additional assumptions on the model.

2. Technical overview

Notation We use asymptotic notations $O(\cdot), o(\cdot), \Omega(\cdot), \omega(\cdot)$ for $n \to \infty$. We use $\tilde{O}(\cdot)$ to hide $\frac{1}{\mathrm{polylog}(n)}$ factors. We use $\|\cdot\|$ for the Euclidean norm of vectors. We use boldface for random variables. For a vector $x \in \mathbb{R}^d$, we use $x_{\setminus 1} \in \mathbb{R}^{d-1}$ for the vector obtained by removing the first coordinate. For a matrix $A \in \mathbb{R}^{n\times d}$, we use A_j for the j-th column and $A_{j:k}$ for the submatrix containing columns j through k. For two vectors $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$, we denote by $\mathrm{concat}(x, y)$ the vector $(x, y) \in \mathbb{R}^{d_1+d_2}$.

In this section, we present our reduction from PCA with a negative spike (cf. Conjecture 4) to improperly learning sparse linear regression models. Our reduction is similar to the one of Bresler et al. (2018). However, a crucial difference is that we start from sparse PCA with a *negative* spike. Indeed, since the design matrices in the resulting SLR instance in Bresler et al. (2018) satisfy the restricted eigenvalue property with appropriate parameters, they can be solved with error 0.01 using only $O(k \log d) \ll k^2$ samples. Further, the reduction of Bresler et al. (2018) only works assuming access to a *proper* sparse linear regression algorithm, i.e., one that outputs a sparse solution. By giving a slightly different reduction, we show that it is enough to assume access to an improper learner. As a warm-up, we will show that improperly learning sparse linear regression models is hard when the variance of the noise is unknown (but known to be in [0,1]). Second, we show that even when the variance is known and equal to 1 (i.e., Definition 1) the problem remains hard. For the remainder of the paper, unless explicitly specified, when referring to ”solving sparse linear regression” or similar expressions, we always mean improper learners.

Hardness of certifying RIP and a first lower bound We first recall that achieving prediction error 0.01 is possible in polynomial time using $O(k \log d)$ samples under additional assumptions on the design matrix. In particular, if the design matrix satisfies the restricted eigenvalue condition (with a constant), the LASSO algorithm achieves this guarantee. In order to construct a lower bound instance, it is thus necessary that this property does not hold (or only with weaker parameters than necessary for known algorithms). We focus here on the related *restricted isometry property* (RIP). In particular, a matrix $X \in \mathbb{R}^{n\times d}$ is said to satisfy (k, δ)-RIP if for all k-sparse unit vectors v it holds that $\|Xv\|^2 \in [1-\delta, 1+\delta]$.

Suppose that we could *certify* that a matrix satisfies RIP. This would give us a way to convince ourselves that our learning algorithm succeeded in some cases: Check if the design matrix satisfies RIP, and if yes run LASSO. Else, we do not guarantee anything. It is thus natural to wonder whether there is a formal connection between RIP certification and sparse linear regression. Unfortunately, it is believed that certifying RIP is computationally difficult Ding et al. (2021).[4] On the other hand, our work shows that we can exploit this connection to show *hardness results* in the following way: Lower bounds for RIP certification are proved by showing that an associated distinguishing problem

[4] Interestingly, the results in Ding et al. (2021) are tight, in the sense that there is a certification algorithm matching their lower bound Koiran and Zouzias (2014).

between two distributions is hard. One hypothesis corresponds to matrices with good RIP, the other does not. If we can certify RIP, we can distinguish the two. Our main observation is that sparse linear regression solvers can be used to solve this distinguishing problem.

In particular, consider the following (degenerate) instance of the negative sparse PCA problem. This instance will already give us a lower bound for sparse regression if we ask that the algorithm works in the case when the variance of the noise is unknown between 0 and 1. Given samples $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ from either $\mathcal{Q} = N(0, I_d)$ or $\mathcal{P} = N(0, I_d - \boldsymbol{x}\boldsymbol{x}^\top)$, where $\boldsymbol{x}$ is a uniformly random $(k+1)$-sparse unit vector with $\boldsymbol{x}_1 = -\frac{1}{\sqrt{k+1}}$, decide which is the case. Note that this instance being hard implies that RIP certification is hard: Consider the matrix $\boldsymbol{Z} \in \mathbb{R}^{n \times d}$ with rows $\boldsymbol{z}_i$. Under $\mathcal{Q}$ this has good RIP and under $\mathcal{P}$ it does not. Our key observation is that the absence of RIP implies a dependence between the columns of $\boldsymbol{Z}$ that is not present under $\mathcal{Q}$. Luckily for us, in our setting this dependence is linear and only concerns k columns. We can thus hope to detect it with our sparse regression oracle.

Indeed, consider the matrix $\boldsymbol{Z}$ under $\mathcal{P}$: It holds that $\boldsymbol{Z}_1\boldsymbol{x}_1 + \boldsymbol{Z}_{\setminus 1}\boldsymbol{x}_{\setminus 1} = \boldsymbol{Z}\boldsymbol{x} = 0$ and hence $\boldsymbol{Z}_1 = \boldsymbol{Z}_{\setminus 1}(-\frac{1}{\boldsymbol{x}_1}\boldsymbol{x}_{\setminus 1})$. Under $\mathcal{Q}$ on the other hand the columns are independent and $\boldsymbol{Z}_1 = \boldsymbol{Z}_{\setminus 1} \cdot 0 + \boldsymbol{w}$, where $\boldsymbol{w} \sim N(0, \mathrm{Id}_n)$. In either case, let $\boldsymbol{A} = \boldsymbol{Z}_{\setminus 1}$ and $\boldsymbol{y} = \boldsymbol{Z}_1$. We have shown that in both cases $(\boldsymbol{A}, \boldsymbol{y})$ forms a valid input for our regression algorithm since $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{w}$, where $\boldsymbol{x}^*$ is 0 or $-\frac{1}{\boldsymbol{x}_1}\boldsymbol{x}_{\setminus 1}$ under $\mathcal{Q}$ and $\mathcal{P}$ respectively and $\boldsymbol{w}$ is independent of $\boldsymbol{A}$ and either $N(0, \mathrm{Id}_n)$ or 0.[5] Our regression oracle allows us to estimate $\boldsymbol{A}\boldsymbol{x}^*$. In particular, under $\mathcal{Q}$, $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{Z}_{\setminus 1}\boldsymbol{x}^* = 0$ and hence if $\hat{\boldsymbol{x}}$ is our estimator then $\boldsymbol{A}\hat{\boldsymbol{x}}$ should have small norm. On the other hand, under $\mathcal{P}$, $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{Z}_{\setminus 1}\boldsymbol{x}^* = \boldsymbol{Z}_1 \sim N(0, (1 - \frac{1}{k+1}) \cdot \mathrm{Id}_n)$ and we expect our estimate to have large norm. Indeed, suppose our regression oracle outputs $\hat{\boldsymbol{x}}$ achieving prediction error 0.01 with probability at least $1 - \delta$. Then, under $\mathcal{Q}$, $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\hat{\boldsymbol{x}}\| = \frac{1}{\sqrt{n}}\|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{x}^*\| \leq 0.1$ with probability at least $1 - \delta$. However, under $\mathcal{P}$ it holds that

$$\frac{1}{\sqrt{n}}\|\boldsymbol{A}\hat{\boldsymbol{x}}\| \geq \left| \frac{1}{\sqrt{n}}\|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{x}^*\| - \frac{1}{\sqrt{n}}\|\boldsymbol{A}\boldsymbol{x}^*\| \right| .$$

By standard concentration bounds it follows that $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\boldsymbol{x}^*\| = \frac{1}{\sqrt{n}}\|\boldsymbol{Z}_1\|$ is at least, say, 0.6 with high probability and thus $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\hat{\boldsymbol{x}}\| \geq 0.5$ with probability at least $1 - \delta - o(1)$. It follows that thresholding $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\hat{\boldsymbol{x}}\|$ faithfully distinguishes $\mathcal{Q}$ and $\mathcal{P}$ with probability at least $1 - \delta - o(1)$.

Note that if we assume that our sparse linear regression algorithm has success probability at least $1 - O(\frac{1}{d})$, we do not need to assume that the first coordinate of the sparse PCA prior in Definition 3 is known to be $-\frac{1}{\sqrt{k+1}}$. Indeed, instead of the above reduction, we can set $\boldsymbol{y} = \boldsymbol{Z}_i$ and $\boldsymbol{A} = \boldsymbol{Z}_{\setminus i}$ for all $i \in [d]$ and run our regression solver on $(\boldsymbol{A}, \boldsymbol{y})$. Under $\mathcal{Q}$, with probability at least, say, 0.99, it holds that in all of these runs we have that $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\hat{\boldsymbol{x}}\| \leq 0.1$. Similarly, under $\mathcal{P}$, with probability at least 0.99, there exists at least one i such that $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\hat{\boldsymbol{x}}\| \geq 0.5$. Thus, we can distinguish $\mathcal{Q}$ and $\mathcal{P}$ with at least constant probability, based on whether $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\hat{\boldsymbol{x}}\|$ is large in at least one iteration.

Note that in both cases the ℓ_1-norm of the secret is at most $O(k)$ and thus LASSO would achieve prediction error 0.01 using $O(k^2 \log d)$ samples, under the slow-rate analysis.

Extension to non-degenerate negative sparse PCA So far our reduction used a degenerate planted hypothesis $\mathcal{P} = N(0, I_d - \boldsymbol{x}\boldsymbol{x}^\top)$. In what follows, we argue that the same reduction

[5] Note that the 0 vector is also k-sparse.

can also produce valid instances from non-degenerate hypotheses. The sparse linear regression instances produced will now have unknown variance of the noise between 0.5 and 1.

In particular, let $\theta = \frac{k+1}{k+2} \approx 1 - \frac{1}{k}$. We start with the negative sparse PCA problem in which $\mathcal{Q} = N(0, \mathrm{Id}_d)$ and $\mathcal{P} = N(0, \mathrm{Id}_d - \theta \cdot \boldsymbol{x}\boldsymbol{x}^\top)$, where $\boldsymbol{x}$ is again a uniformly random $(k+1)$-sparse unit vector with $\boldsymbol{x}_1 = -\frac{1}{\sqrt{k+1}}$. The null case does not change (and the variance of the noise in this case is 1), so it suffices to analyze $\mathcal{P}$. Then, for the planted case, we use the following fact proved in Bresler et al. (2018):[6]

Fact 6 (Appendix B.2 in Bresler et al. (2018)) *Let $\boldsymbol{y}, \boldsymbol{A}$ be as above (under $\mathcal{P}$) and $\gamma = \frac{\theta}{1-\theta \cdot \frac{k}{k+1}} = \frac{k+1}{2}$. Then $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{w}$, where $\boldsymbol{A}\boldsymbol{x}^*$ and $\boldsymbol{w}$ are independent with $\boldsymbol{w} \sim N(0, \sigma^2 \mathrm{Id}_n)$, and*

$$\boldsymbol{x}^* = \frac{\gamma}{\sqrt{k+1}} \cdot \boldsymbol{x}_{\setminus 1} = \frac{\sqrt{k+1}}{2} \cdot \boldsymbol{x}_{\setminus 1}\,, \qquad \sigma^2 = 1 - \frac{\gamma}{k+1} = \frac{1}{2}\,.$$

This establishes that the variance of the noise is always either 0.5 or 1. It follows using the same techniques that under $\mathcal{P}$ it still holds that $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\hat{\boldsymbol{x}}\| \geq 0.5$ with probability at least $1 - \delta - o(1)$.

Known variance of the noise Our goal is to reduce $\mathrm{NegSPCA}_\theta$ to sparse linear regression instances that have *known* variance of the noise.

In the discussions so far our strategy has been to set $\boldsymbol{y} = \boldsymbol{Z}_1$. Then under $\mathcal{Q}$ we have $\boldsymbol{y} \sim N(0, \mathrm{Id}_n)$, but under $\mathcal{P}$ we have $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{w}$ where $\boldsymbol{w}$ is Gaussian with variance *less than* 1. Indeed, if $\boldsymbol{w}$ had variance exactly 1, we could distinguish $\mathcal{Q}$ and $\mathcal{P}$ by thresholding the norm of $\boldsymbol{y}$. Therefore, in this strategy the variance of the noise cannot be the same under $\mathcal{Q}$ and $\mathcal{P}$. More generally, in $\mathrm{NegSPCA}_\theta$ the null case and the planted case are asymmetric, so it seems difficult to obtain some $\boldsymbol{y}$ for which the variance of the noise is the same.

Instead, we introduce a new, symmetric, distinguishing problem $\mathrm{PairNegSPCA}_\theta$ which requires distinguishing between a sample from $\mathcal{P} \times \mathcal{Q}$ and a sample from $\mathcal{Q} \times \mathcal{P}$, where $\mathcal{P}$ and $\mathcal{Q}$ are the planted and null distributions in $\mathrm{NegSPCA}_\theta$:

Definition 7 (Paired spiked Wishart model) *Let $d, k, n \in \mathbb{N}$ with $d \geq k$ and $\theta \in (0, 1)$. Let $\mathcal{P}$ and $\mathcal{Q}$ be the planted and null distributions in a $\mathrm{NegSPCA}_\theta$ problem, respectively. We define the $\mathrm{PairNegSPCA}_\theta$ problem as the following distinguishing problem: We are given n i.i.d. samples from either $\mathcal{P} \times \mathcal{Q}$ or $\mathcal{Q} \times \mathcal{P}$ defined as follows and we want to decide whether they came from $\mathcal{P} \times \mathcal{Q}$ or $\mathcal{Q} \times \mathcal{P}$.*

- *$\mathcal{P} \times \mathcal{Q}$: Sample independently $z \sim \mathcal{P}$ and $z' \sim \mathcal{Q}$ and return (z, z').*
- *$\mathcal{Q} \times \mathcal{P}$: Sample independently $z \sim \mathcal{Q}$ and $z' \sim \mathcal{P}$ and return (z, z').*

Then we reduce $\mathrm{PairNegSPCA}_\theta$ to sparse linear regression by setting $\boldsymbol{y}$ to be the sum of a column as in $\mathcal{P}$ and a column as in $\mathcal{Q}$ – without knowing which is which. This ensures that the variance of the noise is identical under $\mathcal{P} \times \mathcal{Q}$ and $\mathcal{Q} \times \mathcal{P}$. In addition, we reduce $\mathrm{NegSPCA}_\theta$ to $\mathrm{PairNegSPCA}_\theta$, so by composition we obtain the desired reduction from $\mathrm{NegSPCA}_\theta$ to sparse linear regression with known variance of the noise.

[6] Formally, Bresler et al. (2018) only shows this fact for sparse PCA instances with *positive* spikes. The proof they give also works for the setting with negative spikes.

3. Sparse linear regression reduction

Theorem 8 gives our main result.

Theorem 8 *Let $d, k, n \in \mathbb{N}$ with $d \geq k$ and $\theta \in (0, 1)$. Assume there exists an efficient algorithm that, given n samples from the sparse linear regression model with variance of the noise $\sigma^2 = 1$, achieves prediction error $\frac{1}{n} \|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{x}^*\|^2$ at most 0.01 with probability $1 - \delta$. Then there exists an efficient algorithm that solves $\mathrm{NegSPCA}_\theta$ for $\theta = \frac{k+1}{k+2}$ with probability $1 - \sqrt{2\delta} - \exp(-\Omega(n))$.*

We first state a reduction from $\mathrm{NegSPCA}_\theta$ to $\mathrm{PairNegSPCA}_\theta$. Informally, this says that if it is hard to decide whether a sample comes from $\mathcal{P}$ or $\mathcal{Q}$, then it is also hard, given a sample from $\mathcal{P}$ and one from $\mathcal{Q}$, to decide which is which. The result follows from a more general reduction from distinguishing distributions to distinguishing paired distributions that we give in Lemma 10.

Lemma 9 *Let $d, k, n \in \mathbb{N}$ with $d \geq k$ and $\theta \in (0, 1)$. Assume there exists an efficient algorithm that solves $\mathrm{PairNegSPCA}_\theta$ with probability at least $1 - \delta$. Then there exists an efficient algorithm that solves $\mathrm{NegSPCA}_\theta$ with probability at least $1 - \sqrt{2\delta}$.*

The remaining step is to reduce $\mathrm{PairNegSPCA}_\theta$ to sparse linear regression.

The reduction Let $\theta = \frac{k+1}{k+2}$. Given n samples $(\boldsymbol{z}_1, \boldsymbol{z}'_1), \ldots, (\boldsymbol{z}_n, \boldsymbol{z}'_n)$ from $\mathrm{PairNegSPCA}_\theta$, let $\boldsymbol{Z} \in \mathbb{R}^{n \times 2d}$ be the matrix with rows $\mathrm{concat}(\boldsymbol{z}_1, \boldsymbol{z}'_1), \ldots, \mathrm{concat}(\boldsymbol{z}_n, \boldsymbol{z}'_n)$. We do the following:

1. Set $\boldsymbol{y} = \boldsymbol{Z}_1 + \boldsymbol{Z}_{d+1}$ and $\boldsymbol{A} = \boldsymbol{Z}_{\setminus\{1,d+1\}}$. That is, $\boldsymbol{A}$ is the $n \times (2d-2)$ matrix that is obtained by removing columns $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_{d+1}$ from $\boldsymbol{Z}$.
2. Invoke the sparse linear regression solver on $(\boldsymbol{A}, \boldsymbol{y})$ to obtain an estimator $\hat{\boldsymbol{x}}$.
3. If $\frac{1}{\sqrt{n}} \|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{Z}_1\| \leq \frac{1}{\sqrt{n}} \|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{Z}_{d+1}\|$ output $\mathcal{P} \times \mathcal{Q}$. Else output $\mathcal{Q} \times \mathcal{P}$.

Consider the case $\mathcal{P} \times \mathcal{Q}$. As in Fact 6, let $\gamma = \frac{\theta}{1-\theta \cdot k/(k+1)} = \frac{k+1}{2}$. Then, by the same argument as in Fact 6, we can write $\boldsymbol{Z}_1 = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{w}'$, where $\boldsymbol{A}\boldsymbol{x}^*$ and $\boldsymbol{w}'$ are independent with $\boldsymbol{w}' \sim N(0, \sigma^2 \mathrm{Id}_n)$, and

$$\boldsymbol{x}^* = \frac{\sqrt{k+1}}{2} \cdot \mathrm{concat}(\boldsymbol{x}_{\setminus 1}, \boldsymbol{0}^{d-1}) \,, \qquad \sigma^2 = \frac{1}{2} \,.$$

Also, $\boldsymbol{Z}_{d+1} \sim N(0, \mathrm{Id}_n)$ is independent of $\boldsymbol{A}\boldsymbol{x}^*$ and $\boldsymbol{w}'$, so we can write $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{w}$, where $\boldsymbol{A}\boldsymbol{x}^*$ and $\boldsymbol{w}$ are independent with $\boldsymbol{w} \sim N(0, (\sigma^2 + 1)\mathrm{Id}_n)$. Note that by symmetry the variance of the noise is $\sigma^2 + 1$ also in the case $\mathcal{Q} \times \mathcal{P}$.

We now prove Theorem 8.

Proof (Proof of Theorem 8) Consider the case $\mathcal{P} \times \mathcal{Q}$. We have $\boldsymbol{Z}_1 = \boldsymbol{A}\boldsymbol{x}^* + \boldsymbol{w}$, where $\boldsymbol{w} \sim N(0, 1.5\mathrm{Id}_n)$. By the assumption on our sparse linear regression estimator, by scaling up the guarantees such that the known variance of the noise is 1.5, we have with probability $1 - \delta$ that $\frac{1}{\sqrt{n}} \|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{x}^*\| \leq 0.01 \cdot \sqrt{1.5} \leq 0.13$. By the triangle inequality, we get

$$\frac{1}{\sqrt{n}} \|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{Z}_1\| \leq \frac{1}{\sqrt{n}} \|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{x}^*\| + \frac{1}{\sqrt{n}} \|\boldsymbol{w}'\| \,.$$

Recall that $\boldsymbol{w}'$ has covariance matrix $0.5 \cdot \mathrm{Id}_n$. Then we have by Fact 11 with probability $1 - \exp(-\Omega(n))$ that $\frac{1}{\sqrt{n}} \|\boldsymbol{w}'\| \leq \sqrt{0.5} + 0.001 \leq 0.71$. Then

$$\frac{1}{\sqrt{n}} \|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{Z}_1\| \leq 0.13 + 0.71 = 0.84\,.$$

On the other hand

$$\frac{1}{\sqrt{n}} \|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{Z}_{d+1}\| \geq \left| \frac{1}{\sqrt{n}} \|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{x}^*\| - \frac{1}{\sqrt{n}} \|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{Z}_{d+1}\| \right| .$$

The entries of $\boldsymbol{Z}_{d+1}$ are i.i.d. Gaussian with mean zero and variance 1, independent of the multivariate Gaussian $\boldsymbol{A}\boldsymbol{x}^*$. Then we have by Fact 11 with probability $1 - \exp(-\Omega(n))$ that $\frac{1}{\sqrt{n}} \|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{Z}_{d+1}\|$ is at least 0.99. Then

$$\frac{1}{\sqrt{n}} \|\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{Z}_{d+1}\| \geq 0.99 - 0.13 = 0.86\,,$$

and overall we output $\mathcal{P} \times \mathcal{Q}$ with probability at least $1 - \delta - \exp(-\Omega(n))$.

Analogously, by symmetry, in the case $\mathcal{Q} \times \mathcal{P}$ we also output $\mathcal{Q} \times \mathcal{P}$ with probability $1 - (\delta + \exp(-\Omega(n)))$. Then, by the reduction in Lemma 9, we also solve $\mathrm{NegSPCA}_\theta$ with probability at least $1 - \sqrt{2(\delta + \exp(-\Omega(n)))} \geq 1 - \sqrt{2\delta} - \exp(-\Omega(n))$.

■

4. Distinguishing paired distributions reduction

In this section we show that an algorithm that distinguishes a sample from the paired distributions $\mathcal{P} \times \mathcal{Q}$ and $\mathcal{Q} \times \mathcal{P}$ also distinguishes a sample from the distributions $\mathcal{P}$ and $\mathcal{Q}$, assuming that $\mathcal{Q}$ is known.

Lemma 10 *Let $\mathcal{P}$ and $\mathcal{Q}$ be probability distributions over a domain $\mathcal{D}$, and suppose that it is possible to sample independently from $\mathcal{Q}$ in time T_{sample}. Suppose that there exists a (randomized) algorithm $\mathcal{A}$ that takes as input a pair $(x, x') \in \mathcal{D} \times \mathcal{D}'$ and runs in time T_{dist} with the following guarantees:*

- $\mathbb{P}_{\boldsymbol{x} \sim \mathcal{P}, \boldsymbol{x}' \sim \mathcal{Q}}(\mathcal{A}(\boldsymbol{x}, \boldsymbol{x}') = 0) \geq 1 - \delta$*, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are sampled independently and where the probability is also over any internal randomness of $\mathcal{A}$.*
- $\mathbb{P}_{\boldsymbol{x} \sim \mathcal{Q}, \boldsymbol{x}' \sim \mathcal{P}}(\mathcal{A}(\boldsymbol{x}, \boldsymbol{x}') = 1) \geq 1 - \delta$*, where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are sampled independently and where the probability is also over any internal randomness of $\mathcal{A}$.*

Then there also exists a randomized algorithm $\mathcal{B}$ that takes as input $x \in \mathcal{D}$ and runs in time $O((T_{\text{sample}} + T_{\text{dist}})/\sqrt{\delta})$ with the following guarantees:

- $\mathbb{P}_{\boldsymbol{x} \sim \mathcal{P}}(\mathcal{B}(\boldsymbol{x}) = 0) \geq 1 - \sqrt{2\delta}$*, where the probability is also over any internal randomness of $\mathcal{B}$.*
- $\mathbb{P}_{\boldsymbol{x} \sim \mathcal{Q}}(\mathcal{B}(\boldsymbol{x}) = 1) \geq 1 - \sqrt{2\delta}$*, where the probability is also over any internal randomness of $\mathcal{B}$.*

Proof Let $\boldsymbol{x} \in \mathcal{D}$ such that either $\boldsymbol{x} \sim \mathcal{P}$ or $\boldsymbol{x} \sim \mathcal{Q}$. Let M be a positive integer that we will fix later. The randomized algorithm $\mathcal{B}$ is the following:

```
1. Initialize $\Delta = 0$.
2. Repeat $M$ times:
    (a) Sample $\boldsymbol{x}' \sim \mathcal{Q}$.
    (b) If $\mathcal{A}(\boldsymbol{x}, \boldsymbol{x}')$ returns 0 and $\mathcal{A}(\boldsymbol{x}', \boldsymbol{x})$ returns 1, increment $\Delta$.
3. If $\Delta = M$, return 0. Else, return 1.
```

We now analyze the algorithm.

If $\boldsymbol{x} \sim \mathcal{P}$, then $\mathbb{P}_{\boldsymbol{x} \sim \mathcal{P}}(\mathcal{B}(\boldsymbol{x}) = 1) = \mathbb{P}(\Delta < M) \leq 2M\delta$. This is because the probability that $\mathcal{A}$ is wrong in an iteration is at most δ, so the probability that $\mathcal{A}$ is wrong in any of the $2M$ iterations is at most $2M\delta$.

On the other hand, if $\boldsymbol{x} \sim \mathcal{Q}$, we show that $\mathbb{P}_{\boldsymbol{x} \sim \mathcal{Q}}(\mathcal{B}(\boldsymbol{x}) = 0) = \mathbb{P}(\Delta = M) \leq \frac{1}{M+1}$. From now on, we condition on the internal randomness of $\mathcal{A}$, which makes $\mathcal{A}$ deterministic, and show that the same probability upper bound holds irrespective of this internal randomness. This implies that the upper bound also holds unconditionally. To begin, let $\boldsymbol{x}_1, ..., \boldsymbol{x}_M$ be the M random variables corresponding to the x' that are generated in the M iterations. Then we are interested in the event $E(\boldsymbol{x}, (\boldsymbol{x}_1, ..., \boldsymbol{x}_M))$ that Δ is incremented in each iteration. First, note that $E(\boldsymbol{x}, (\boldsymbol{x}_1, ..., \boldsymbol{x}_M))$ has the same probability as $E(\boldsymbol{x}_i, (\boldsymbol{x}, \boldsymbol{x}_1, ..., \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, ..., \boldsymbol{x}_M))$ for all $i \in [M]$, because $\boldsymbol{x}, \boldsymbol{x}_1, ..., \boldsymbol{x}_M$ are independently and identically distributed according to $\mathcal{Q}$. Second, note that all the events $E(\boldsymbol{x}, (\boldsymbol{x}_1, ..., \boldsymbol{x}_M))$ and $E(\boldsymbol{x}_i, (\boldsymbol{x}, \boldsymbol{x}_1, ..., \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, ..., \boldsymbol{x}_M))$ for all $i \in [M]$ are mutually exclusive (e.g., for $i \neq j$, $E(\boldsymbol{x}_i, (\boldsymbol{x}, \boldsymbol{x}_1, ..., \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, ..., \boldsymbol{x}_M))$ implies that $\mathcal{A}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ returns 0, but $E(\boldsymbol{x}_j, (\boldsymbol{x}, \boldsymbol{x}_1, ..., \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, ..., \boldsymbol{x}_M))$ implies that $\mathcal{A}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ returns 1, which is a contradiction). Hence, the probability that Δ is incremented in every iteration is at most $\frac{1}{M+1}$. Note that this argument did not use the value of the internal randomness of $\mathcal{A}$, so the same probability upper bound also holds unconditionally.

Taking $M = 1/\sqrt{2\delta}$ gives the desired probability bounds. Finally, the time complexity of algorithm $\mathcal{B}$ is $O((T_{\text{sample}} + T_{\text{dist}})M) = O((T_{\text{sample}} + T_{\text{dist}})/\sqrt{\delta})$. ■

Acknowledgments

The authors thank David Steurer for helpful discussions. 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).

References

Thomas D Ahle. Sharp and simple bounds for the raw moments of the binomial and poisson distributions. *Statistics & Probability Letters*, 182:109306, 2022.

Benny Applebaum, Boaz Barak, and David Xiao. On basing lower-bounds for learning on worst-case assumptions. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 211–220. IEEE, 2008.

Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019.

Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage, 2020.

Matthew Brennan, Guy Bresler, Samuel B Hopkins, Jerry Li, and Tselil Schramm. Statistical query algorithms and low-degree tests are almost equivalent. *arXiv preprint arXiv:2009.06107*, 2020.

Guy Bresler, Sung Min Park, and Madalina Persu. Sparse pca from sparse linear regression. *Advances in Neural Information Processing Systems*, 31, 2018.

Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. Continuous lwe. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 694–707, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3451000. URL https://doi.org/10.1145/3406325.3451000.

Ilias Diakonikolas and Daniel Kane. Non-gaussian component analysis via lattice basis reduction. In *Conference on Learning Theory*, pages 4535–4547. PMLR, 2022.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

Yunzi Ding, Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. The average-case time complexity of certifying the restricted isometry property. *IEEE Transactions on Information Theory*, 67(11):7355–7361, 2021. doi: 10.1109/TIT.2021.3112823.

Yunzi Ding, Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. Subexponential-time algorithms for sparse pca, 2022.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S. Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. volume 64, New York, NY, USA, apr 2017. Association for Computing Machinery. doi: 10.1145/3046674. URL https://doi.org/10.1145/3046674.

Dean Foster, Howard Karloff, and Justin Thaler. Variable selection is hard. In *Conference on Learning Theory*, pages 696–709. PMLR, 2015.

Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Continuous lwe is as hard as lwe & applications to learning gaussian mixtures. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1162–1173. IEEE, 2022.

Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Sparse linear regression and lattice problems. *arXiv preprint arXiv:2402.14645*, 2024.

Justin Holmgren and Alexander S Wein. Counterexamples to the low-degree conjecture. *arXiv preprint arXiv:2004.08454*, 2020.

Samuel Hopkins. *Statistical Inference and the Sum of Squares Method.* PhD thesis, 2018.

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

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. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

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

Jonathan Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. Lower bounds on randomly preconditioned lasso via robust sparse designs. *Advances in Neural Information Processing Systems*, 35:24419–24431, 2022a.

Jonathan Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. Lasso with latents: Efficient estimation, covariate rescaling, and computational-statistical gaps. *arXiv preprint arXiv:2402.15409*, 2024.

Jonathan A Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. On the power of preconditioning in sparse linear regression. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 550–561. IEEE, 2022b.

Pascal Koiran and Anastasios Zouzias. Hidden cliques and the certification of the restricted isometry property. *IEEE transactions on information theory*, 60(8):4999–5006, 2014.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019.

Leonard Pitt and Leslie G Valiant. Computational limitations on learning from examples. *Journal of the ACM (JACM)*, 35(4):965–984, 1988.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10): 6976–6994, 2011.

Oded Regev. On lattices, learning with errors, random linear codes, and cryptography, 2024.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Martin J. Wainwright. *Graphical models for high-dimensional data*, page 347–382. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.011.

Ilias Zadik, Min Jae Song, Alexander S Wein, and Joan Bruna. Lattice-based methods surpass sum-of-squares in clustering. In *Conference on Learning Theory*, pages 1247–1248. PMLR, 2022.

Yuchen Zhang, Martin J. Wainwright, and Michael I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 921–948, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/zhang14.html.

Yuchen Zhang, Martin J. Wainwright, and Michael I. Jordan. Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators, 2015.

Appendix A. Concentration bounds

The following fact is taken from (Wainwright, 2019, Example 2.11).

Fact 11 *Let $\boldsymbol{X} \sim N(0, \mathrm{Id}_n)$. Then*

$$\mathbb{P}\left(\left|\frac{1}{n}\left\|\boldsymbol{X}\right\|^2 - 1\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{8}\right).$$

Appendix B. Low-degree lower bound for negative-spike sparse Wishart model

B.1. Low-degree framework

One significant barrier for proving computational lower bounds in high dimension statistics is that we need to show average-case hardness: we want to show that no efficient algorithms can succeed with high probability over random inputs. It is notoriously difficult to show average-case hardness based on worst-case hardness assumptions such as $\mathrm{P} \neq \mathrm{NP}$, except in a few notable examples Regev (2024); Bruna et al. (2021); Brennan and Bresler (2020).

The low-degree method, which is developed in Hopkins and Steurer (2017); Hopkins (2018); Hopkins et al. (2017), provides a way to give evidence for average-case hardness by showing that no efficient algorithms based on low-degree polynomials can succeed with high probability over random inputs. In the context of hypothesis testing problems, it has been applied to a wide range of fundamental problems, such as community detection Hopkins and Steurer (2017), spiked matrix models Kunisky et al. (2019), sparse principal component analysis Hopkins et al. (2017); Ding et al. (2022), and certifying restricted isometry property Ding et al. (2021).

The crux of the techniques is to bound the projection of the likelihood ratio function onto the space of low-degree polynomials:

Definition 12 (Definition 1.14 in Kunisky et al. (2019)) *Let $\mathsf{V}_n^{\leq D} \subset L^2(\mathcal{Q}_n)$ denote the linear subspace of polynomials $\mathsf{S}_n \to \mathbb{R}$ of degree at most D. Let $\mathsf{P}^{\leq D} : L^2(\mathcal{Q}_n) \to \mathsf{V}_n^{\leq D}$ denote the orthogonal projection onto this linear subspace.*[7] *Finally, define the D-*low-degree likelihood ratio *(D-LDLR) as $L_n^{\leq D} := \mathsf{P}^{\leq D} L_n$.*

Then in hypothesis testing a low-degree lower bound shows that efficient algorithms based on thresholding low-degree polynomials are inherently obstructed from distinguishing the two distributions. Particularly, we have the following proposition connecting the low-degree likelihood ratio and the distribution separation by low-degree polynomials:

[7] To clarify, the orthogonal projection is with respect to the inner product induced by $\mathcal{Q}_n$ operator on this subspace.

Proposition 13 (Hopkins and Steurer (2017); Hopkins (2018); Kunisky et al. (2019)) *Let $\mathcal{P}$ and $\mathcal{Q}$ be two sequences of probability measures, and $L_n = \frac{dP}{dQ}$ be their likelihood ratio. Then there exists $\varepsilon > 0$ and $D = D(n) \geq (\log n)^{1+\varepsilon}$ for which $\|L_n^{\leq D}\|$ remains bounded as $n \to \infty$ if and only if for any degree $\leq D$ polynomials $f(\cdot)$ we have*

$$\frac{\mathbb{E}_{\mathcal{P}} f}{\sqrt{\mathbb{E}_{\mathcal{Q}} f^2}} \leq O(1)\,.$$

The low-degree conjecture states that, for ”sufficiently nice” distributions, low-degree lower bounds rule out all polynomial-time algorithms that strongly distinguish the distributions:

Conjecture 14 (Informal, Hopkins (2018); Kunisky et al. (2019)) *For ”sufficiently nice” sequences of probability measures $\mathcal{P}$ and $\mathcal{Q}$, let $L_n = \frac{dP}{dQ}$ be their likelihood ratio. If there exists $\varepsilon > 0$ and $D = D(n) \geq (\log n)^{1+\varepsilon}$ for which $\|L_n^{\leq D}\|$ remains bounded as $n \to \infty$, then there is no polynomial-time algorithm that* strongly distinguishes *$\mathcal{P}$ and $\mathcal{Q}$, i.e., distinguishes the two distributions with $1 - o(1)$ probability.*

The formal version of the conjecture, given originally as Conjecture 2.2.4 in Hopkins (2018) and further updated in Holmgren and Wein (2020), only rules out polynomial time-algorithms that distinguish between $U_\delta \mathcal{P}$ and $\mathcal{Q}$, for any $\delta > 0$, where U_δ is the Ornstein-Uhlenbeck noise operator: $U_\delta \mathcal{P}$ is sampled by drawing $x \sim \mathcal{P}$ and $y \sim \mathcal{Q}$ and outputting $\sqrt{1-\delta}x + \sqrt{\delta}y$. For our sparse linear regression result we need $\delta \leq ck$ for some small enough constant $c > 0$ such that $U_\delta \mathcal{P}$ continues to encode a negative spike of magnitude $1 - \Theta(1/k)$.

B.2. LD lower bound for negative-spike sparse Wishart model

In this section, we provide evidence of hardness for the sparse negative-spike Wishart model, based on the low-degree likelihood ratio (LDLR) method.

We show that the low-degree likelihood ratio is bounded for the sparse negative-spike Wishart model in the hard regime conjectured in Conjecture 4.

Our proof follows closely the proof of Ding et al. (2021). The main differences are that in our planted distribution the prior of $\boldsymbol{x}^*$ is the uniform distribution over k-sparse vectors (instead of a sparse Rademacher distribution, in which the exact sparsity can vary) and that the first coordinate of $\boldsymbol{x}^*$ is non-zero.

Theorem 15 (Low-degree lower bound for $\mathrm{NegSPCA}_\theta$) *Fix $\delta \in (0, 0.1]$. Suppose that $n = o(\min(d, k^{2-\delta}))$. Consider the planted distribution $\mathcal{P}_n$ formed from n samples from $\mathcal{P}$ and null distribution $\mathcal{Q}_n$ formed from n samples from $\mathcal{Q}$ in $\mathrm{NegSPCA}_\theta$ defined in Definition 3. Then for any degree-k^δ polynomial $f(z_1, z_2, \ldots, z_n) : \mathbb{R}^{n \times d} \to \mathbb{R}$ such that $\mathbb{E}_{\mathcal{Q}} f = 0$, we have*

$$\frac{\mathbb{E}_{\mathcal{P}_n} f}{\sqrt{\mathbb{E}_{\mathcal{Q}_n} f^2}} \leq 1 + o(1)\,.$$

Proof For degree-D polynomials, the maximum value of $\frac{\mathbb{E}_{\mathcal{P}_n} f}{\sqrt{\mathbb{E}_{\mathcal{Q}_n} f^2}}$ is given by (see Hopkins (2018))

$$\|L_{n,\gamma,\theta,\chi}\|^2 := \mathbb{E}\left(\left(\frac{d\mathcal{P}_n}{d\mathcal{Q}_n}\right)^{\leq D}\right)^2 .$$

Equations (8) and (10) in Ding et al. (2021) provide the bound

$$\begin{aligned}\|L_{n,\gamma,\theta,\chi}\|^2 &\leq \sum_{\ell=0}^{\lfloor D/2 \rfloor} \frac{(2n+4D)^\ell}{\ell!} \cdot \frac{\theta^{2\ell}}{4^\ell} \cdot \mathbb{E}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle^{2\ell} \\ &\leq 1 + \sum_{\ell=1}^{\lfloor D/2 \rfloor} \left(\frac{4\theta^2 \max(n,D)}{\ell} \right)^\ell \mathbb{E}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle^{2\ell} \\ &\leq 1 + \sum_{\ell=1}^{\lfloor D/2 \rfloor} \left(\frac{4\max(n,D)}{\ell} \right)^\ell \mathbb{E}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle^{2\ell},\end{aligned}$$

where $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$ are independently sampled from the prior of the spike vector in the planted distribution, i.e., the first coordinates $\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_1^{(2)}$ are sampled uniformly from $\left\{ \pm\sqrt{\frac{1}{k+1}} \right\}$ and the rest of the coordinates $\boldsymbol{x}_{\setminus 1}^{(1)}, \boldsymbol{x}_{\setminus 1}^{(2)}$ are sampled uniformly from k-sparse $\left\{ \pm\sqrt{\frac{1}{k+1}}, 0 \right\}^d$ vectors.

Now we note that

$$\begin{aligned}\mathbb{E}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle^{2\ell} &= \mathbb{E}\left(\langle \boldsymbol{x}_{\setminus 1}^{(1)}, \boldsymbol{x}_{\setminus 1}^{(2)} \rangle + \boldsymbol{x}_1^{(1)} \boldsymbol{x}_1^{(2)} \right)^{2\ell} \\ &\leq 2^{2\ell} \cdot \left(\mathbb{E}\langle \boldsymbol{x}_{\setminus 1}^{(1)}, \boldsymbol{x}_{\setminus 1}^{(2)} \rangle^{2\ell} + \mathbb{E}\left(\boldsymbol{x}_1^{(1)} \boldsymbol{x}_1^{(2)} \right)^{2\ell} \right) \\ &= 2^{2\ell} \cdot \mathbb{E}\langle \boldsymbol{x}_{\setminus 1}^{(1)}, \boldsymbol{x}_{\setminus 1}^{(2)} \rangle^{2\ell} + \left(\frac{4}{(k+1)^2} \right)^\ell,\end{aligned}$$

so

$$\|L_{n,\gamma,\theta,\chi}\|^2 \leq 1 + \sum_{\ell=1}^{\lfloor D/2 \rfloor} \left(\frac{16\max(n,D)}{\ell} \right)^\ell \cdot \mathbb{E}\langle \boldsymbol{x}_{\setminus 1}^{(1)}, \boldsymbol{x}_{\setminus 1}^{(2)} \rangle^{2\ell} + \sum_{\ell=1}^{\lfloor D/2 \rfloor} \left(\frac{16\max(n,D)}{(k+1)^2 \ell} \right)^\ell.$$

Using that $n = o(k^2)$ and $D = o(k^2)$, we have

$$\sum_{\ell=1}^{\lfloor D/2 \rfloor} \left(\frac{16\theta^2 \max(n,D)}{(k+1)^2 \ell} \right)^\ell \leq o(1).$$

For the remaining term, we have by Lemma 16 that there is an absolute constant $C > 0$ such that

$$\sum_{\ell=1}^{\lfloor D/2 \rfloor} \left(\frac{16\max(n,D)}{\ell} \right)^\ell \cdot \mathbb{E}\langle \boldsymbol{x}_{\setminus 1}^{(1)}, \boldsymbol{x}_{\setminus 1}^{(2)} \rangle^{2\ell} \leq \sum_{\ell=1}^{\lfloor D/2 \rfloor} \left(C \cdot \max(n,D) \cdot \left(\frac{1}{d} + \frac{\ell}{k^2} \right) \right)^\ell.$$

For this sum to be bounded by $o(1)$, we want both $\max(n,D) = o(d)$ and $\max(n,D) = o(k^2/D)$. This is achieved if $n = o(d)$ and $nD = o(k^2)$ and $D = o(k)$, so it is achieved under the conditions of the theorem with $D \leq k^\delta$. Then

$$\|L_{n,\gamma,\theta,\chi}\|^2 \leq 1 + o(1).$$

■

Lemma 16 (Expectation bound) *In the setting of Theorem 15, there exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\langle \boldsymbol{x}^{(1)}_{\setminus 1}, \boldsymbol{x}^{(2)}_{\setminus 1}\rangle^{2\ell} \le (C\ell)^{\ell}\left(\frac{1}{d}+\frac{\ell}{k^2}\right)^{\ell} . \tag{1}$$

Proof Let $\boldsymbol{S}_1, \boldsymbol{S}_2 \subset [d]$ be the sets of non-zero indices of $\boldsymbol{x}^{(1)}_{\setminus 1}$ and $\boldsymbol{x}^{(2)}_{\setminus 1}$ respectively. We have

$$\mathbb{E}\langle \boldsymbol{x}^{(1)}_{\setminus 1}, \boldsymbol{x}^{(2)}_{\setminus 1}\rangle^{2\ell} = \frac{1}{(k+1)^{2\ell}} \mathop{\mathbb{E}}_{\boldsymbol{S}_1,\boldsymbol{S}_2} \mathbb{E}\left[\left(\sum_{i\in \boldsymbol{S}_1\cap \boldsymbol{S}_2} \boldsymbol{\rho}_i\right)^{2\ell} \middle| \boldsymbol{S}_1, \boldsymbol{S}_2\right] ,$$

where $\boldsymbol{\rho}_i$ are i.i.d. Rademacher variables. Then $\sum_{i\in \boldsymbol{S}_1\cap \boldsymbol{S}_2} \boldsymbol{\rho}_i$ is a sub-Gaussian random variable with variance proxy $|\boldsymbol{S}_1\cap \boldsymbol{S}_2|$, so for an absolute constant $C > 0$,

$$\mathbb{E}\langle \boldsymbol{x}^{(1)}_{\setminus 1}, \boldsymbol{x}^{(2)}_{\setminus 1}\rangle^{2\ell} \le \frac{(C\ell)^{\ell}}{(k+1)^{2\ell}}\, \mathbb{E}\,|\boldsymbol{S}_1\cap \boldsymbol{S}_2|^{\ell} .$$

Let $\boldsymbol{a}_i \in \{0,1\}$ for $i \in [k]$ be the indicator that the i-th element of $\boldsymbol{S}_1$ is also in $\boldsymbol{S}_2$. Then $|\boldsymbol{S}_1\cap \boldsymbol{S}_2| = \boldsymbol{a}_1 + \ldots + \boldsymbol{a}_k$. Also let $\boldsymbol{b}_i$ for $i \in [k]$ be i.i.d. $\mathrm{Ber}(\frac{k}{d})$ variables. Then for any $S \subseteq [k]$ we have

$$\mathbb{E}\prod_{i\in S} \boldsymbol{a}_i = \frac{k}{d}\frac{k-1}{d-1}\cdots\frac{k-|S|+1}{d-|S|+1} \le \left(\frac{k}{d}\right)^{|S|} = \mathbb{E}\prod_{i\in S}\boldsymbol{b}_i .$$

Then also $\mathbb{E}(\boldsymbol{a}_1+\ldots+\boldsymbol{a}_k)^{\ell} \le \mathbb{E}\,\boldsymbol{B}^{\ell}$ where $\boldsymbol{B}$ is a $\mathrm{Bin}(k, \frac{k}{d})$ variable. (Ahle, 2022, Corollary 1) gives that $\mathbb{E}\,\boldsymbol{B}^{\ell} \le \left(\frac{k^2}{d}+\frac{\ell}{2}\right)^{\ell}$. Therefore, overall

$$\mathbb{E}\langle \boldsymbol{x}^{(1)}_{\setminus 1}, \boldsymbol{x}^{(2)}_{\setminus 1}\rangle^{\ell} \le \frac{(C\ell)^{\ell}}{(k+1)^{2\ell}}\left(\frac{k^2}{d}+\frac{\ell}{2}\right)^{\ell} \le (C\ell)^{\ell}\left(\frac{1}{d}+\frac{\ell}{k^2}\right)^{\ell} .$$

■

Appendix C. Statistical query lower bound for negative-spike sparse Wishart model

In this section, we prove a statistical query (SQ) lower bound for negative spike sparse Wishart model. Our proof heavily relies on an almost equivalence established between LDLR lower bounds and SQ lower bounds Brennan et al. (2020), and resembles the proof in section 8.3 of their paper.

C.1. SQ framework

The SQ framework is a restricted computational model where a learning algorithm can make certain types of queries to an oracle and get answers that are subject to a certain degree of noise Kearns (1998). We will focus on the SQ model with VSTAT queries which is used in Brennan et al. (2020), where the learning algorithm has access to the VSTAT oracle as defined below:

Definition 17 (VSTAT oracle) *Given a query $\phi : \mathbb{R}^d \to [0,1]$ and a distribution D over $\mathbb{R}^d$, the VSTAT(n) oracle returns $\mathbb{E}_{x\sim D}[\phi(x)] + \zeta$ for an adversarially chosen $\zeta \in \mathbb{R}$ such that $|\zeta| \leq \max\left(\frac{1}{n}, \sqrt{\frac{\mathbb{E}[\phi](1-\mathbb{E}[\phi])}{n}}\right)$.*

One way to show an SQ lower bound is by computing the statistical dimension of the hypothesis testing problem, which is a measure on the complexity of the testing problem. In this paper, we use the following definition of statistical dimension introduced by Feldman et al. (2017):

Definition 18 (Statistical dimension) *Let $\mu_\emptyset$ be some distribution with $\mathcal{D}_\emptyset$ as density function. Let $\mathcal{S} = \{\mu_u\}$ be some family of distributions indexed by u, such that μ_u has density function given by $\mathcal{D}_u$. Consider the hypothesis testing problem between:*

- *Null hypothesis: sample i.i.d. from $\mu_\emptyset$,*
- *Alternative hypothesis: sample i.i.d. from μ_u where u is sampled from some prior distribution μ.*

For $D_u \in \mathcal{S}$, define the relative density $\bar{D}_u(x) = \frac{D_u(x)}{D_\emptyset(x)}$ and the inner product $\langle f, g\rangle = \mathbb{E}_{x\sim D_\emptyset}[f(x)g(x)]$. The statistical dimension $SDA(\mathcal{S}, \mu, n)$ measures the tail of $\langle \bar{D}_u, \bar{D}_v\rangle - 1$ with u, v drawn independently from μ:

$$\begin{aligned}&SDA(\mathcal{S}, \mu, n)\\ &= \max\left\{q \in \mathbb{N} : \mathop{\mathbb{E}}_{u,v\sim\mu}\left[|\langle \bar{D}_u, \bar{D}_v\rangle - 1| \mid A\right] \leq \frac{1}{n} \text{ for all events } A \text{ s.t. } \mathop{\mathbb{P}}_{u,v\in\mu}(A) \geq \frac{1}{q^2}\right\}.\end{aligned}$$

We will use $SDA(n)$ or $SDA(\mathcal{S}, n)$ when $\mathcal{S}$ or μ are clear from the context. In Feldman et al. (2017), it was shown that the statistical dimension is a lower bound on the SQ complexity of the hypothesis test using VSTAT oracles:

Theorem 19 (Theorem 2.7 of Feldman et al. (2017), Theorem A.5 of Brennan et al. (2020)) *Let $D_\emptyset$ be a null distribution and $\mathcal{S}$ be a set of alternative distributions. Then any (randomized) statistical query algorithm which solves the hypothesis testing problem between $D_\emptyset$ and $\mathcal{S}$ with probability at least $1-\gamma$ requires at least $(1-\gamma)SDA(\mathcal{S}, n)$ queries to $VSTAT(\frac{n}{3})$.*

The almost equivalence between SQ lower bounds and low-degree lower bounds is established in Brennan et al. (2020):

Theorem 20 (Theorem 3.1 in Brennan et al. (2020), LDLR to SDA) *Let $\ell \in \mathbb{N}$ with ℓ even and $\mathcal{S} = \{D_v\}_{v\in S}$ be a collection of probability distributions with prior μ over $\mathcal{S}$. Suppose that $\mathcal{S}$ satisfies:*

- *The ℓ-sample high-degree part of the likelihood ratio is bounded by $\left\|\mathbf{E}_{u\sim\mathcal{S}}\left(\bar{D}_u^{>\ell}\right)^{\otimes\ell}\right\| \leqslant \gamma$.*
- *For some $n \in \mathbb{N}$, the degree-ℓ likelihood ratio is bounded by $\left\|\mathbf{E}_{u\sim\mathcal{S}}\left(\bar{D}_u^{\otimes n}\right)^{\leqslant\ell} - 1\right\| \leqslant \varepsilon$.*

Then for any $q \geqslant 1$, it follows that

$$\mathrm{SDA}\left(\mathcal{S}, \frac{n}{q^{2/\ell}\left(\ell\varepsilon^{2/\ell} + \gamma^{2/\ell} n\right)}\right) \geqslant q\,.$$

C.2. SQ lower bound for sparse negative-spike Wishart model

The sparse negative-spike Wishart model $\mathrm{NegSPCA}_\theta$ under the statistical query framework has the following form:

Model 21 (Sparse spiked Wishart model in SQ model) *Let $d, k \in \mathbb{N}$ with $d \geq k$ and $\theta \in (0,1)$. We define the $\mathrm{NegSPCA}_\theta$ problem in the statistical query case as the following distinguishing problem: We want to distinguish between the distributions $\mathcal{P}$ or $\mathcal{Q}$ defined as follows, using VSTAT oracle calls to them.*

- ***Planted distribution** $\mathcal{P}$: Sample a unit vector $\boldsymbol{x} \in \mathbb{R}^{d+1}$ with $\boldsymbol{x}_1 = -\frac{1}{\sqrt{k+1}}$ and $\boldsymbol{x}_{\setminus 1}$ sampled uniformly from k-sparse $\left\{\pm\frac{1}{\sqrt{k+1}}, 0\right\}^d$ vectors and fix it for the rest of the sampling procedure. Produce samples by sampling from $N(0, \mathrm{Id}_{d+1} - \theta \cdot \boldsymbol{x}\boldsymbol{x}^\top)$.*
- ***Null distribution** $\mathcal{Q}$: $N(0, \mathrm{Id}_{d+1})$.*

Corresponding to Theorem 20, the set of distributions $\mathcal{S}$ is parameterized by the random unit vector $\boldsymbol{x}$, and is given by $\left\{N(0, \mathrm{Id}_{d+1} - \theta \cdot \boldsymbol{x}\boldsymbol{x}^\top)\right\}$ in our setting. The prior distribution μ is determined by the distribution of $\boldsymbol{x}$.

Under the statistical query framework, we have the following lower bound:

Theorem 22 (SQ lower bound for $\mathrm{NegSPCA}_\theta$) *Consider the $\mathrm{NegSPCA}_\theta$ model. Let $0 < \delta \leq 0.1$ be an arbitrary absolute constant. Then, for $n = o(\min(d/k^\delta, k^{2-2\delta}))$, we have the SQ lower bound SDA$(\mathcal{S}, n) \geq 2^{k^\delta}$.*

To apply the almost equivalence relation between low-degree lower bounds and statistical query lower bounds, we need to prove the following lemma:

Lemma 23 *Consider the $\mathrm{NegSPCA}_\theta$ model. Fix $\delta \in (0, 0.1]$. Then for any $\ell \leq k^\delta$, we have*

$$\left\| \mathop{\mathbb{E}}_{u\sim\mathcal{S}} \left(\bar{D}_u^{>\ell}\right)^{\otimes \ell} \right\|^2 \leq k^{-\ell^2/3}\,.$$

Proof We follow the same proof as Lemma 8.21 in Brennan et al. (2020)[8], and get

$$\left\| \mathop{\mathbb{E}}_{u\sim\mathcal{S}} (\bar{D}_x^{>\ell})^{\otimes \ell} \right\|^2 \leq \ell^{2\ell^2}\left(1 - 4\theta^2\right)^{2\ell^2} \mathbb{E}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\rangle^{2\ell\cdot(\ell+1)}\,,$$

where $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$ are independently sampled from the prior of the spike vector in the planted distribution, i.e., the first coordinates $\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_1^{(2)}$ are sampled uniformly from $\left\{\pm\sqrt{\frac{1}{k+1}}\right\}$ and the rest of the coordinates $\boldsymbol{x}_{\setminus 1}^{(1)}, \boldsymbol{x}_{\setminus 1}^{(2)}$ are sampled uniformly from k-sparse $\left\{\pm\sqrt{\frac{1-\frac{1}{k+1}}{k}}, 0\right\}^d$ vectors.

Following the proof of Theorem 15, we have

$$\mathbb{E}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\rangle^{2\ell(\ell+1)} \leq 2^{2\ell(\ell+1)} \cdot (C\ell(\ell+1))^{\ell(\ell+1)} \cdot \left(\frac{1}{d} + \frac{\ell(\ell+1)}{k^2}\right)^{\ell(\ell+1)} + \left(\frac{4}{(k+1)^2}\right)^{\ell(\ell+1)}\,.$$

[8] Particularly, we set their k and d to ℓ, and n to d in our setting.

As a result, when $\ell \leq k^{\delta}$, we have

$$\ell^{2\ell^2} \left(1 - 4\theta^2\right)^{2\ell^2} \mathbb{E}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle^{2\ell(\ell+1)} \leq (1/k^{1/3})^{\ell^2} ,$$

which finishes the proof. ■

Proof (Proof of Theorem 22) We apply Theorem 20 with $n = o(\min(d, k^{2-\delta}))$, $\ell = k^{\delta}$ and $q = 2^{\ell}$. By Lemma 23, we can take $\gamma = k^{-\frac{\ell^2}{3}}$. By Theorem 15, we can take $\epsilon = o(1)$. As a result, we have $\mathrm{SDA}(\mathcal{S}, \frac{n}{100k^{\delta}}) \geq 2^{k^{\delta}}$. ■