Proceedings of Machine Learning Research vol 247:1–47, 2024 37th Annual Conference on Learning Theory

Lasso with Latents: Efficient Estimation, Covariate Rescaling, and Computational-Statistical Gaps

Jonathan A. Kelner KELNER@MIT.EDU
MIT

Frederic Koehler FKOEHLER@UCHICAGO.EDU
University of Chicago

Raghu Meka RAGHUM@CS.UCLA.EDU
UCLA

Dhruv Rohatgi DROHATGI@MIT.EDU
MIT

Editors: Shipra Agrawal and Aaron Roth

Abstract

It is well-known that the statistical performance of Lasso can suffer significantly when the covariates of interest have strong correlations. In particular, the prediction error of Lasso becomes much worse than computationally inefficient alternatives like Best Subset Selection. Due to a large conjectured computational-statistical tradeoff in the problem of sparse linear regression, it may be impossible to close this gap in general.

In this work, we propose a natural sparse linear regression setting where strong correlations between covariates arise from unobserved latent variables. In this setting, we analyze the problem caused by strong correlations and design a surprisingly simple fix. While Lasso with standard normalization of covariates fails, there exists a heterogeneous scaling of the covariates with which Lasso will suddenly obtain strong provable guarantees for estimation. Moreover, we design a simple, efficient procedure for computing such a "smart scaling."

The sample complexity of the resulting "rescaled Lasso" algorithm incurs (in the worst case) quadratic dependence on the sparsity of the underlying signal. While this dependence is not information-theoretically necessary, we give evidence that it is optimal among the class of polynomial-time algorithms, via the method of low-degree polynomials. This argument reveals a new connection between sparse linear regression and a special version of sparse PCA with a *near-critical negative spike*. The latter problem can be thought of as a real-valued analogue of learning a sparse parity. Using it, we also establish the first computational-statistical gap for the closely related problem of learning a Gaussian Graphical Model.

1. Introduction

Sparse linear regression (SLR) is one of the most fundamental problems in high-dimensional statistics. In this paper, we study algorithmic aspects of the problem. For simplicity, we focus on the following setting with Gaussian random design (though our results should be generalizable to e.g., sub-Gaussian data, misspecification via oracle inequalities, etc.):

Definition 1 *Let $\Sigma \in \mathbb{R}^{n\times n}$ be positive semi-definite, $w^\star \in \mathbb{R}^n$ be k-sparse, and $\sigma \geq 0$. We define $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$ to be the distribution of (X, y) where $X \sim N(0, \Sigma)$ and $y \sim N(\langle X, w^\star\rangle, \sigma^2)$.*

Given m independent samples $(X^{(j)}, y^{(j)})_{j=1}^m$ from $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$, the goal of sparse linear regression is to produce an estimate $\hat{w}$ with low *out-of-sample (clean) prediction error*, defined as:

$$\mathbb{E}(\langle X^{(0)}, \hat{w}\rangle - \langle X^{(0)}, w^\star\rangle)^2 = (\hat{w} - w^\star)^\top \Sigma (\hat{w} - w^\star) =: \|\hat{w} - w^\star\|_\Sigma^2$$

where $(X^{(0)}, y^{(0)})$ is a fresh sample from $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$.

Despite significant effort, there is a vast gap in our understanding of the *computational complexity* of sparse linear regression – and, in particular, how computational efficiency interplays with sample efficiency. On the one hand, the natural *Best Subset Selection* estimator (Hocking and Leslie, 1967) achieves prediction error $O(\sigma^2 k(\log n)/m)$ with m samples, so long as $m = \Omega(k \log n)$. Note that the sample complexity scales only logarithmically with the ambient dimension n, and no further assumptions on Σ or $w^\star$ are needed. Unfortunately, this estimator is computationally intractable. On the other hand, classical estimators such as Lasso (Tibshirani, 1996) can be computed in polynomial time. However, they are known to have poor statistical performance (e.g., sample complexity *linear* in n) in many settings where the covariates have strong correlations. In particular, Lasso is only statistically efficient when Σ satisfies a restricted condition-number assumption such as the *compatibility condition* (Van De Geer et al., 2009). While there are several special cases where Lasso fails but other polynomial-time algorithms are known to succeed, these are (thus far) the exceptions to the rule.[1] See Section 3.1 for further discussion about Lasso and other estimators.

Given the dearth of strong algorithmic guarantees for SLR, it's natural to speculate that some choices of Σ make SLR computationally hard for any sample complexity $m = o(n)$. But proving such a lower bound via average-case reduction or in any standard restricted computational model (e.g., low-degree polynomials or statistical queries) seems out of reach at present (see Section 3.2 for discussion of prior attempts). We lack even a *conjecture* about which families of Σ might induce computational hardness: obviously, Σ must be ill-conditioned, but little else is clear.

In this paper, we make progress on this problem by identifying the fundamental computational limits for a subclass of SLR problems. Informally, the subclass captures common situations where strong correlations are due to the existence of a few latent confounders or a few directions of unusually small variance in the data; see below for more details. For this subclass, we give mathematical evidence that no efficient estimator can succeed with significantly less than $O(k^2 \log n)$ samples (even though $O(k \log n)$ samples suffice information-theoretically), and we design a new polynomial-time algorithm that matches the lower bound. Our efficient algorithm is based on a simple but surprisingly powerful *smart scaling* procedure that we use as a preprocessing step to "fix" the Lasso. Our lower bound is based on a new connection between SLR and what we call the *near-critical* regime of negatively spiked sparse principal component analysis (PCA).

1.1. Upper bounds

We start by describing two natural settings where Σ may be arbitrarily ill-conditioned (and Lasso has poor sample complexity and performs poorly empirically), but the degeneracies among covariates are sufficiently few or structured so that one may still hope for an efficient SLR algorithm.

Setting 1 (Latent variable models) *Correlations are often induced via latent confounders. Thus, as is common in econometrics, causal inference, and other fields (see e.g. Hoyle (1995); Pearl*

1. Note that even if Σ were known, it's not clear whether the problem would become any easier. One could precondition the covariates by $\Sigma^{-1/2}$, but this typically destroys the sparsity guarantee. While preconditioning is a useful algorithmic tool, finding the *right* preconditioner is often challenging (whether Σ is known or unknown).

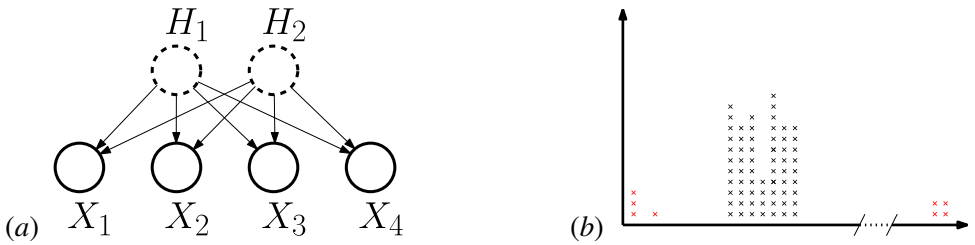

Figure 1: (a) Example graphical model with $X_1, \ldots, X_4$ observed and H_1, H_2 latent. (b) Example eigenspectrum that is well-conditioned aside from a few "outliers" (displayed in red).

(2009)), we can posit that covariates follow a Structural Equation Model *(SEM) with $h \ll n$ latent variables. Formally, we suppose that each observed covariate vector $X^{(j)}$ can be written as $X^{(j)} := AH^{(j)} + Z^{(j)}$ where A is an unknown, fixed $n \times h$ matrix, $H^{(j)}$ is an unobserved Gaussian random vector, and independently $Z^{(j)} \sim N(0, D)$ is Gaussian noise where $D \succ 0$ is diagonal. Thus, there is a rank-h matrix $L \succeq 0$ such that each $X^{(j)}$ is a multivariate Gaussian with covariance matrix*

$$\Sigma := \mathbb{E}\, XX^\top = L + D. \tag{1}$$

Setting 2 (Eigenspectrum with outliers) *Alternatively, we may restrict the degeneracies in Σ by explicitly controlling the eigenspectrum. In this setting, originally introduced by Kelner et al. (2023), we assume that the spectrum of Σ is well-concentrated, aside from a small number of "outlier" eigenvalues. That is, suppose that the eigenvalues of Σ are $\lambda_1 \geq \cdots \geq \lambda_n$, and there is some $d \ll n$ such that $\lambda_{d+1}/\lambda_{n-d}$ is small (see Figure 1 for a depiction). Unlike in Kelner et al. (2023), we do not assume that Σ is known.*

Note that the latter setting generalizes the classical, well-conditioned setting (where $\lambda_1/\lambda_n = O(1)$). Both settings allow for a small number of approximate linear dependencies among the covariates, which is a natural case where Lasso may provably fail, requiring as many as $\Omega(n)$ samples to achieve non-trivial prediction error (Kelner et al., 2021, Theorem 6.5).

Challenge: adapting to unknown structure. In both settings, the covariates are drawn from a highly structured distribution, but one of the main challenges is that the structure is *unknown*. In the latent variable model setting, Σ has a "low-rank plus diagonal" decomposition. However, even if Σ is known, efficiently computing such a decomposition is a well-studied open problem (Saunderson et al., 2012; Bertsimas et al., 2017; Wu et al., 2020) with some evidence of intractability (Tunçel et al., 2023). In the outlier setting, note that the eigendecomposition of Σ isn't even identifiable from a sublinear number of samples. Thus, efficient sparse linear regression in these settings requires exploiting unknown structure without completely learning it.

1.1.1. AN EFFICIENT ALGORITHM VIA SMART SCALING

Conventional wisdom when applying the Lasso (and in statistics and machine learning more broadly) is to scale all covariates to unit variance (Ahrens et al., 2020). While this is a good idea in many cases, it is not always the *optimal* choice! In fact, in both settings described above, even if Lasso

has poor performance with the standard scaling, there always exists a clever rescaling after which Lasso would achieve near-optimal sample complexity. We formalize this existence criterion via the following notion of (α, h)-*rescalability*. It essentially states that after rescaling by some diagonal matrix, covariates from $N(0, \Sigma)$ satisfy a restricted eigenvalue condition (similar to Raskutti et al. (2010)) modulo a low-rank subspace $\mathrm{span}(L)$.

Definition 2 *For any $n \in \mathbb{N}$ and $\gamma > 1$, let $\mathcal{C}_n(\gamma) := \{x \in \mathbb{R}^n : \|x\|_1 \le \gamma \|x\|_\infty\}$ be the set of* γ-quantitatively sparse vectors.

Definition 3 *Let $n \in \mathbb{N}$, and let $\Sigma \in \mathbb{R}^{n \times n}$ be a positive semi-definite matrix. For $k, h \in \mathbb{N}$ and $\alpha > 0$, we say that Σ is* (α, h)-rescalable at sparsity k *if there are matrices $D, L \in \mathbb{R}^{n \times n}$ such that $D \succ 0$ is diagonal, $L \succeq 0$ has rank at most h, and*

$$I_n \preceq_{\mathcal{C}_n(32k)} D^{-1/2} \Sigma D^{-1/2} \preceq \alpha I_n + L \tag{2}$$

where $I_n \preceq_{\mathcal{C}_n(32k)} D^{-1/2}\Sigma D^{-1/2}$ means that $v^\top v \le v^\top D^{-1/2}\Sigma D^{-1/2} v$ for all $v \in \mathcal{C}_n(32k)$. If Equation (2) holds for all k, then we simply say that Σ is (α, h)-rescalable.

In Setting 1 (the latent variable model with h latent variables), it's immediate from Definition 3 that Σ is $(1, h)$-rescalable. In Setting 2, the implication is less obvious, but we are able to show that Σ is (α, h)-rescalable at sparsity k with $\alpha = O(k^2 \lambda_{d+1}/\lambda_{n-d})$ and $h = O(k^2 d)$ (see Lemma 29). Thus, the notion of rescalability unifies both settings.

If D were known, then one could simply rescale each sample via $X \mapsto D^{-1/2}X$. By standard analyses, Lasso with this "oracle rescaling" would have sample complexity $O((\alpha k + h)\log n)$, which is information-theoretically optimal for $\alpha, h = O(1)$. However, as discussed above, it is unreasonable to assume access to D, which in Setting 1 consists of the conditional variances of the covariates with respect to the unknown latent variables. Our first main result is a computationally efficient SLR algorithm `RescaledLasso()` that doesn't need to know D (or Σ), and nonetheless matches the sample complexity of the "oracle rescaled" Lasso up to a factor of k:

Theorem 4 *Let $n, m, k, h \in \mathbb{N}$ and $\alpha, \delta, \sigma, \lambda > 0$. Suppose that $\Sigma \in \mathbb{R}^{n \times n}$ is (α, h)-rescalable at sparsity k, and $w^\star \in \mathbb{R}^n$ is k-sparse. Let $(X^{(j)}, y^{(j)})_{j=1}^m \sim \mathsf{SLR}_{\Sigma,\sigma}(w^\star)$ be independent samples, and define $\hat{w} := \texttt{RescaledLasso}((X^{(j)}, y^{(j)})_{j=1}^m, k, \lambda)$ (see Algorithm 1).*

If $m = \Omega((\alpha k^2 + h)\log(n/\delta))$ and $\lambda = \Omega(\sigma\sqrt{(\alpha k^2 + h)\log(n/\delta)/(k^2 m)})$, then with probability at least $1 - \delta$ it holds that $\|\hat{w} - w^\star\|_\Sigma^2 \le O(k^2\lambda^2)$. Moreover, the algorithm's time complexity is $\mathrm{poly}(n, \log \max_i \frac{\Sigma_{ii}}{D_{ii}})$, where D is the (unknown) matrix in Definition 3.

In particular, for the optimal choice of λ, the algorithm achieves prediction error $O(\sigma^2(\alpha k^2 + h)\log(n)/m)$ with high probability. Note that the time complexity depends on $\max_i \Sigma_{ii}/D_{ii}$, but only logarithmically; hence, the algorithm runs in $\mathrm{poly}(n)$ time even if this ratio is exponentially large. Applying Theorem 4 to Setting 1 is immediate using Equation (1); simply set $\alpha = 1$, and take h to be the number of latent variables:

Corollary 5 *Let $n, m, k, h \in \mathbb{N}$ and $\delta, \sigma, \lambda > 0$. Let $\Sigma := D + L \in \mathbb{R}^{n \times n}$ for some diagonal matrix $D \succ 0$ and rank-h matrix $L \succeq 0$, and let $w^\star \in \mathbb{R}^n$ be k-sparse. Let $(X^{(j)}, y^{(j)})_{j=1}^m \sim \mathsf{SLR}_{\Sigma,\sigma}(w^\star)$ be independent samples, and define $\hat{w} := \texttt{RescaledLasso}((X^{(j)}, y^{(j)})_{j=1}^m, k, \lambda)$. If $m = \Omega((k^2 + h)\log(n/\delta))$ and $\lambda = \Omega(\sigma\sqrt{(k^2 + h)\log(n/\delta)/(k^2 m)})$, then with probability at least $1 - \delta$ it holds that $\|\hat{w} - w^\star\|_\Sigma^2 \le O(k^2\lambda^2)$. Moreover, the time complexity is $\mathrm{poly}(n, \log \max_i \frac{\Sigma_{ii}}{D_{ii}})$.*

Note the quadratic dependence on k above and in Theorem 4. While not information-theoretically necessary, we give evidence in Section 1.2 that it is the optimal dependence for efficient algorithms.

The more involved application is to Setting 2, where proving rescalability is non-trivial (see Lemma 29). Combining Theorem 4 with Lemma 29 yields the following result.

Corollary 6 *Let $n, m, k \in \mathbb{N}$ and $\delta, \sigma, \lambda > 0$. Suppose that $\Sigma \in \mathbb{R}^{n \times n}$ is positive definite with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, and that $w^\star \in \mathbb{R}^n$ is k-sparse. Let $(X^{(j)}, y^{(j)})_{j=1}^m$ be i.i.d. samples from $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$, and define $\hat{w} := \texttt{RescaledLasso}((X^{(j)}, y^{(j)})_{j=1}^m, k, \lambda)$. If $m = \Omega(\min_{0 \leq d < n}(k^4 \frac{\lambda_{d+1}}{\lambda_{n-d}} + k^2 d) \log(n/\delta))$ and $k\lambda = \Omega(\sigma \min_{0 \leq d < n}(k^2 \sqrt{\frac{\lambda_{d+1}}{\lambda_{n-d}}} + k\sqrt{d}) \sqrt{\log(n/\delta)/m})$, then with probability at least $1 - \delta$ it holds that $\|\hat{w} - w^\star\|_\Sigma^2 \leq O(k^2 \lambda^2)$. Moreover, the algorithm's time complexity is $\mathrm{poly}(n, \log \frac{\lambda_1}{\lambda_n})$.*

Hence for the optimal choice of λ, the algorithm achieves prediction error $O(\sigma^2 \min_{0 \leq d < n}(k^4 \frac{\lambda_{d+1}}{\lambda_{n-d}} + k^2 d) \log(n/\delta)/m)$. If the spectrum of Σ has few outliers, in the sense that there is some $d = O(1)$ with $\lambda_{d+1}/\lambda_{n-d} = O(1)$, then this simplifies to $O(\sigma^2 k^4 \log(n)/m)$. This significantly improves upon the main result of Kelner et al. (2023), which requires knowledge of Σ and incurs exponential dependence on k; see Section 3.1 for more detailed comparison.

1.2. Lower bounds

In light of Theorem 4, (α, h)-rescalable matrices Σ (for small α, h) are likely not the "hardest" covariance matrices, for which one might expect that no computationally efficient algorithm achieves non-trivial prediction error with $o(n)$ samples. However, there is still a polynomial gap between the sample complexity of `RescaledLasso()` and the information-theoretic optimum: even for constant α and h, the sample complexity of `RescaledLasso()` is $O(k^2 \log n)$ (to achieve prediction error $O(\sigma^2)$), whereas the inefficient Best Subset Selection estimator only requires $O(k \log n)$ samples. It is natural to ask whether this gap is inherent.

We prove that, under a plausible conjecture about the power of low-degree polynomials, the quadratic dependence on k incurred by `RescaledLasso()` may indeed be necessary for *any* computationally efficient algorithm. While lower bounds have previously been shown for specific algorithms (such as Lasso and some generalizations), the below result is, to our knowledge, the first broad evidence for a (super-constant) computational-statistical tradeoff in sparse linear regression; see Section 3.2 for further discussion.

Theorem 7 *Let $\epsilon, C > 0$ with $\epsilon \leq 2$. Let $\mathcal{A}$ be a polynomial-time algorithm. Suppose that for any $n, k \in \mathbb{N}$, $\sigma > 0$, positive semi-definite, $(1, k)$-rescalable matrix $\Sigma \in \mathbb{R}^{n \times n}$, k-sparse vector $w^\star \in \mathbb{R}^n$, and $m \geq Ck^{2-\epsilon} \log n$, the output $\hat{w} \leftarrow \mathcal{A}((X^{(j)}, y^{(j)})_{j=1}^m)$ satisfies*

$$\Pr[\|\hat{w} - w^\star\|_\Sigma^2 \leq \sigma^2/10] \geq 1 - o(1)$$

where the probability is over the randomness of $\mathcal{A}$ and m independent samples $(X^{(j)}, y^{(j)})_{j=1}^m$ from $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$. Then Conjecture 33 is false.

Conjecture 33 is an instantiation of the *Low-Degree Hypothesis*: it asserts that low-degree polynomials have optimal power among polynomial-time algorithms for a natural hypothesis testing problem called *negative-spike sparse PCA*. Informally, this is the problem of distinguishing samples

from the standard multivariate Gaussian $N(0, I_n)$ versus samples from the spiked Wishart model $N(0, I_n + \beta ww^\top)$, where w is a random sparse unit vector, and $\beta \in (-1, 0)$ is the spike strength.

The proof of Theorem 7 has two components. First, we analyze the low-degree likelihood ratio for negative spike k-sparse PCA – thereby showing that low-degree polynomials require $\Omega(k^2)$ samples to solve the testing problem. We give additional evidence for the hardness of this problem by proving a lower bound for a natural SDP formulation. Second, we given an efficient reduction from this testing problem (in the *near-critical* regime where β is close to -1) to sparse linear regression with a $(1, k)$-rescalable covariance matrix Σ.[2]

Our analysis of near-critical negative spike PCA also yields the first computational-statistical gap for learning Gaussian Graphical Models (GGMs). While it's information-theoretically possible to learn any κ-nondegenerate, degree-d GGM with only $O(d \log(n)/\kappa^2)$ samples (Misra et al., 2017), the low-degree analysis implies (under Conjecture 33) that any computationally efficient algorithm requires at least $\Omega(d^{2-\epsilon} \log(n))$ samples, for any constant $\epsilon > 0$ and even when $\kappa = \Omega(1)$. See Remark 43 for details. We do not know if this lower bound is tight for learning GGMs (the true computational-statistical gap may be much larger), but it is in fact tight for a natural *testing* problem: testing between an empty graphical model and a sparse graphical model with at least one nonnegligible edge. The matching (computationally efficient) upper bound is given in Section D.

Independent work. In independent and concurrent work, Buhai, Ding, and Tiegel also gave evidence that sparse linear regression exhibits a k-to-k^2 computational-statistical gap Buhai et al. (2024). Their proof proceeds along the same lines as ours (via reduction from negative-spike sparse PCA and analysis of the low-degree likelihood ratio).

1.3. Outline

In Section 2 we sketch the proofs of our main results. In Section 3 we survey related work on algorithms and lower bounds for sparse linear regression and sparse PCA. In Section A, Section B, and Section C we formally prove Theorem 4, Corollary 6, and Theorem 7 respectively. In Section D, we analyze testing between empty and non-empty GGMs, and in Section E we show the results of applying `RescaledLasso()` to a simple simulated dataset.

2. Proof Overview

In this section we give overviews of the proof of Theorem 4 (via a new variable normalization procedure) and Theorem 7 (via a new connection with negative spike sparse PCA).

Throughout the paper, we adopt the following notation. For $n \times n$ symmetric matrices A, B, we write $A \preceq B$ to denote that $B - A$ is positive semi-definite; for a set $S \subseteq \mathbb{R}^n$, we write $A \preceq_S B$ to denote that $v^\top Av \leq v^\top Bv$ for all $v \in S$. For a matrix $A \in \mathbb{R}^{n \times n}$, we write $\mathrm{diag}(A)$ to denote the matrix $D \in \mathbb{R}^{n \times n}$ defined by $D_{ij} = A_{ij} \mathbb{1}[i = j]$. We write I_n to denote the $n \times n$ identity matrix. For a positive semi-definite matrix $A \in \mathbb{R}^{n \times n}$ and vector $v \in \mathbb{R}^n$, $\|v\|_A$ denotes $\sqrt{v^\top Av}$.

2. To contrast, Bresler et al. (2018) studied solving *positive-spike* sparse PCA in the computationally easy regime by solving *well-conditioned* SLR problems where the LASSO is statistically optimal up to constants. Our hardness reduction is crucially based upon the special properties of sparse PCA with a *large (near-critical) negative spike*.

2.1. Upper bounds

Informally, Theorem 4 states that there is a computationally efficient and sample-efficient algorithm for sparse linear regression (as modelled in Definition 1) whenever the covariance matrix Σ is rescalable (see Definition 3). To reiterate, the main algorithmic difficulty is that the diagonal matrix D in Definition 3 is unknown and potentially even unidentifiable, so we cannot simply perform the "oracle rescaling" $X \mapsto D^{-1/2}X$. In particular, in Setting 1, where $\Sigma = D + L$, the diagonal entry D_{ii} measures the conditional variance of the covariate X_i with respect to the latent variables. Even in very simple examples, these conditional variances are unidentifiable:

Example 1 *Consider a model with latent variable $H_1 \sim N(0, 1-\epsilon^2)$, and independent covariates $X_1 = H_1 + N(0, \epsilon^2)$ and $X_2 \sim N(0, 1)$. Then X_1 has conditional variance ϵ^2, whereas X_2 has conditional variance 1. However, from the observed data it is impossible to tell which of X_1 or X_2 is connected to the latent, since either way the joint law is $X \sim N(0, I_2)$.*

Fortunately, we are happy with *any* good rescaling matrix $\hat{D}$, even if it's not the oracle one. Based on our mathematical understanding of the Lasso, a rescaling should be "good" if the rescaled covariates admit no (quantitatively) sparse approximate dependencies – and these are identifiable, even from a small number of samples. This motivates the algorithm described below.

Algorithm description. The procedure `SmartScaling()` takes as input the covariate data $(X^{(j)})_{j=1}^m$ and the sparsity level k, and then initializes $\hat{D}$ to be the diagonal of the empirical covariance matrix $\hat{\Sigma} = \frac{1}{m}\sum_{j=1}^m X^{(j)}(X^{(j)})^\top$. Note that this initialization corresponds to the "standard" covariate normalization, which may be highly sub-optimal in the presence of strong correlations.

To fix this, the procedure iteratively decreases entries of $\hat{D}$ until it can certify that the resulting scaling is good. At each step, for each covariate X_i, the procedure solves the following program to compute how much of the (empirical) variance of X_i cannot be explained by quantitatively sparse combinations of the other covariates:[3]

$$\min_{\substack{v \in \mathbb{R}^n : \|\hat{D}^{1/2}v\|_1 \le 16k \\ (\hat{D}^{1/2}v)_i = 1}} \frac{1}{m}\sum_{j=1}^m \langle X^{(j)}, v\rangle^2. \tag{3}$$

If (3) is at least a constant for all $i \in [n]$, then `SmartScaling()` returns $\hat{D}$. Otherwise, the procedure picks some i for which (3) is small, and then halves $\hat{D}_{ii}$ and repeats.

After `SmartScaling()` returns $\hat{D}$, the main algorithm `RescaledLasso()` simply applies the rescaling $X^{(j)} \mapsto \hat{D}^{-1/2}X^{(j)}$ to each sample covariate, solves Lasso, and unscales the solution (equivalently, it uses a coordinatewise penalty $\|\hat{D}^{1/2}w\|_1$). See Algorithm 1 for pseudocode.

Example 2 *Consider a model where two covariates X_i, X_j are highly correlated, and the remaining covariates are independent. Concretely, suppose all covariates have unit variance, but $X_i|X_j$ has conditional variance ϵ^2. Then `SmartScaling()` will keep alternately halving $\hat{D}_{ii}$ and $\hat{D}_{jj}$ until both entries are roughly $\epsilon^2 > 0$. At this point, the procedure will terminate. This example also illustrates why the time complexity depends on the log condition number (theorem 4).*

3. We remark that this program is similar in spirit to a natural convex relaxation for detecting the sparsest vector in a subspace — c.f. Demanet and Hand (2014); Spielman et al. (2012). See also our related work section.

Algorithm 1: Adaptive variable rescaling

Procedure `SmartScaling` $(\mathbb{X}, k)$

$\mathsf{DIV} \leftarrow 2$; $\mathsf{B} \leftarrow 16k$; $\hat{\Sigma} \leftarrow \frac{1}{m}\mathbb{X}^\top\mathbb{X}$; $\hat{D}^{(1)} \leftarrow \mathrm{diag}(\hat{\Sigma})$

for $t = 1, 2, 3, \ldots$ **do**

For every $1 \le i \le n$, compute

$$v^{(t,i)} \leftarrow \operatorname*{argmin}_{\substack{v\in\mathbb{R}^n:\ \|(\hat{D}^{(t)})^{1/2}v\|_1 \le \mathsf{B} \\ ((\hat{D}^{(t)})^{1/2}v)_i = 1}} \frac{1}{m}\|\mathbb{X}v\|_2^2 . \tag{4}$$

$i_{\mathsf{min}} \leftarrow \mathrm{argmin}_{i\in[n]} \frac{1}{m}\left\|\mathbb{X}v^{(t,i)}\right\|_2^2$

if $\frac{1}{m}\left\|\mathbb{X}v^{(t,i_{\mathsf{min}})}\right\|_2^2 \le 1$ **then** $\hat{D}^{(t+1)} \leftarrow \hat{D}^{(t)}$; $\hat{D}^{(t+1)}_{i_{\mathsf{min}}i_{\mathsf{min}}} \leftarrow \hat{D}^{(t+1)}_{i_{\mathsf{min}}i_{\mathsf{min}}}/\mathsf{DIV}$

else return $\hat{D}^{(t)}$

Procedure `RescaledLasso` $((X^{(j)}, y^{(j)})_{j=1}^m, k, \lambda)$

Define $\mathbb{X} \in \mathbb{R}^{m\times n}$ by $\mathbb{X} \leftarrow \begin{bmatrix} X^{(1)} & X^{(2)} & \ldots & X^{(m)} \end{bmatrix}^\top$.

$\hat{D} \leftarrow$ `SmartScaling` $(\mathbb{X}, k)$.

Compute and return $\hat{w}$, the solution to the modified Lasso:

$$\hat{w} \leftarrow \operatorname*{argmin}_{w\in\mathbb{R}^n} \frac{1}{m}\|\mathbb{X}w - y\|_2^2 + \lambda\left\|\hat{D}^{1/2}w\right\|_1 . \tag{5}$$

Why does this work? In the above example, the rescaling is "good" because (a) all eigenvalues of the covariance matrix of the rescaled covariates are $\Omega(1)$, and (b) there are not many super-constant eigenvalues (in fact there is only one, in direction $X_i + X_j$); in such situations it can be shown that Lasso succeeds. More generally, the heart of our analysis is the following guarantee about the output $\hat{D}$ of `SmartScaling()`. It states that if the empirical covariance matrix $\hat{\Sigma}$ is spectrally lower bounded (on quantitatively sparse vectors) after the "oracle rescaling" $D^{-1/2}$, then it's also lower bounded after the estimated rescaling $\hat{D}^{-1/2}$, and moreover D is an approximate lower bound on $\hat{D}$.[4] Connecting to the above example, the former fact implies a restricted version of (a), and the latter fact, together with the upper bound in eq. (2), implies (b).

Lemma 8 *Let $n, m, k \in \mathbb{N}$. Let $\mathbb{X} \in \mathbb{R}^{m\times n}$. Suppose that $I_n \preceq_{\mathcal{C}_n(32k)} D^{-1/2}\hat{\Sigma}D^{-1/2}$ where $D \succ 0$ is a diagonal matrix, and $\hat{\Sigma} := \frac{1}{m}\mathbb{X}^\top\mathbb{X}$. Then the algorithm $\hat{D} \leftarrow$ `SmartScaling` $(\mathbb{X}, k)$ terminates after at most $T := n\log\max_{i\in[n]}\frac{2\hat{\Sigma}_{ii}}{D_{ii}}$ repetitions, and moreover the output satisfies:*

1. $\hat{D} \succeq \frac{1}{2}D$.
2. $\frac{1}{m}\|\mathbb{X}\hat{D}^{-1/2}v\|_2^2 > 1$ *for all* $v \in \mathbb{R}^n$ *with* $\|v\|_\infty = 1$ *and* $\|v\|_1 \le 16k$.

For any rescalable Σ, the spectral lower bound needed to apply Lemma 8 to samples from $N(0, \Sigma)$ is inherited (with high probability) from the assumed lower bound on Σ (Lemma 22). This crucially

4. Note however that the guarantee with $\hat{D}^{-1/2}$ is qualitatively weaker than the guarantee with the oracle rescaling: in Lemma 8, $\frac{1}{m}\|\mathbb{X}\hat{D}^{-1/2}v\|_2^2$ is lower bounded in terms of $\|v\|_\infty$ rather than $\|v\|_2$. From a technical perspective, this discrepancy is the source of the quadratic dependence on k in the sample complexity of `RescaledLasso()`.

uses a generalization bound derived from the upper bound $D^{-1/2}\Sigma D^{-1/2} \preceq \alpha I_n + L$. Next, let us explain why `RescaledLasso()` has low prediction error assuming Lemma 8.

It is simplest to consider the special case of sparse linear regression where $\sigma = 0$ (i.e. the responses are noiseless). In this setting, instead of solving the rescaled Lasso program (5), one would solve the rescaled basis pursuit:

$$\hat{w} \in \operatorname*{argmin}_{w \in \mathbb{R}^n : \mathbb{X}w = y} \|\hat{D}^{1/2} w\|_1,$$

where $\mathbb{X} \in \mathbb{R}^{m \times n}$ is the matrix with rows $X^{(1)}, \ldots, X^{(m)}$. When does this program fail to return the true solution $w^\star$? Since $y = \mathbb{X}w^\star$, the program only fails when there is some alternative $\hat{w} \in \mathbb{R}^n$ with $\mathbb{X}(\hat{w} - w^\star) = 0$ and $\|\hat{D}^{1/2}\hat{w}\|_1 \le \|\hat{D}^{1/2} w^\star\|_1$. By a standard manipulation, the second inequality (together with k-sparsity of $w^\star$) implies that the error vector $e := \hat{w} - w^\star$ is $O(k)$-quantitatively sparse with respect to $\hat{D}$, i.e.

$$\|\hat{D}^{1/2} e\|_1 \le O(k) \cdot \|\hat{D}^{1/2} e\|_\infty.$$

By the second guarantee of Lemma 8 with $v := \hat{D}^{1/2} e$ (and the fact that $e \neq 0$), it follows that $\frac{1}{m}\|\mathbb{X}e\|_2^2 > 0$. This is a contradiction, so in fact the rescaled basis pursuit must return $w^\star$. Extending this argument to the general, noisy setting follows a similar rough blueprint; we defer the details to Section A. We now sketch the proof of the key Lemma 8 (see Section A for the full proof).

Proof [Proof sketch for Lemma 8] The second guarantee is immediate from the termination condition of `SmartScaling()`. The bound on the number of repetitions in `SmartScaling()` will be immediate once we show that the output satisfies $\hat{D} \succeq \frac{1}{2} D$, since at every repetition, the algorithm halves at least one entry of $\hat{D}$.

The only non-obvious claim (and the heart of the result) is that $\hat{D} \succeq \frac{1}{2} D$ at termination. For intuition, in this sketch we'll only consider the latent variable model setting (i.e. $\Sigma = D + L$) and the large sample limit $\hat{\Sigma} \approx \Sigma$, but the proof generalizes. Say that each covariate has variance 1, so the algorithm has initialized $\hat{D} := I_n$. Since D measures the conditional variances of the covariates, it's clear that $D \preceq \hat{D}$ holds initially. Now suppose there is some vector v with $v_i = 1$ and $\|v\|_\Sigma \ll 1$. Then v describes an approximate dependency involving covariate X_i, so the conditional variance of X_i (which is exactly D_{ii}) *must* be small: formally,

$$D_{ii} = D_{ii} v_i^2 \le \|v\|_D^2 \le \|v\|_\Sigma^2 \ll 1 = \hat{D}_{ii}.$$

Thus, the algorithm can safely set $\hat{D}_{ii} \leftarrow \hat{D}_{ii}/2$ while preserving the invariant $\hat{D} \succeq \frac{1}{2} D$. At each subsequent step, similar logic applies, so at termination $\hat{D} \succeq \frac{1}{2} D$ still holds. ■

2.2. Lower bounds

Theorem 7 asserts that `RescaledLasso()`, which requires only $O(k^2 \log n)$ samples to achieve prediction error $O(\sigma^2)$ whenever Σ is $(1, k)$-rescalable (and $w^\star$ is k-sparse), is essentially optimal among polynomial-time algorithms, under a conjecture about the power of low-degree polynomials. We prove the theorem by studying negative-spike sparse PCA in a "near-critical" regime. Concretely, this refers to a distribution testing problem between a spiked Wishart distribution $\mathbb{P}_{n,k,\beta,m}$ and a null distribution $\mathbb{Q}_{n,m}$ defined below, in the regime where β is negative and close to -1.

Definition 9 *Let $n, k \in \mathbb{N}$ with $k \le n$. The* fixed-size sparse Rademacher prior *$\mathcal{W}_{n,k}$ is the distribution on $\mathbb{R}^n$ where $w \sim \mathcal{W}_{n,k}$ is drawn by: 1. sampling a subset $S \subseteq [n]$ of size k uniformly at random, and 2. setting $w_i \sim \mathrm{Unif}(\{1/\sqrt{k}, -1/\sqrt{k}\})$ for each $i \in S$ and $w_i = 0$ otherwise.*

Definition 10 *Let $n, k, m \in \mathbb{N}$ with $k \le n$ and $\beta \in (-1, \infty)$. The k-*sparse spiked Wishart distribution *$\mathbb{P}_{n,k,\beta,m}$ is the distribution of $(Z^{(j)})_{j=1}^m$ where first we sample $w \sim \mathcal{W}_{n,k}$ (Definition 9), and then $(Z^{(j)})_{j=1}^m \sim N(0, I_n + \beta w w^\top)^{\otimes m}$. The* null distribution *$\mathbb{Q}_{n,m}$ is defined as $N(0, I_n)^{\otimes m}$.*

For the hardness of negative-spike sparse PCA, we show that for any spike strength $\beta \in (-1, 1)$, degree-$\log^{O(1)}(n)$ polynomials require sample complexity $m \ge \tilde{\Omega}(k^2)$ to test between $\mathbb{P}_{n,k,\beta,m}$ and $\mathbb{Q}_{n,m}$ (Theorem 41). This result largely follows similar bounds for positive-spike sparse PCA (Bandeira et al., 2020; Ding et al., 2023). To give further evidence of hardness, we also prove a lower bound for a natural semidefinite programming relaxation (Theorem 45) — this is inspired by analogous results in the positive spike setting (Krauthgamer et al., 2015), although we need to use a different construction since we are minimizing, rather than maximizing, the SDP objective.

We then show that an improved algorithm for sparse linear regression (with rescalable covariances) would yield an improved tester for negative spike sparse PCA when β is close to -1:

Theorem 11 *Let $m_{\mathsf{SLR}} : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ be a function, and suppose that there is a polynomial-time algorithm $\mathcal{A}$ with the following property. For any $n, k \in \mathbb{N}$, $\sigma > 0$, positive semi-definite $(1, k)$-rescalable matrix $\Sigma \in \mathbb{R}^{n \times n}$, k-sparse vector $w^\star \in \mathbb{R}^n$, and $m \ge m_{\mathsf{SLR}}(n, k)$, the estimate $\hat{w} \leftarrow \mathcal{A}((X^{(j)}, y^{(j)})_{j=1}^m)$ satisfies*

$$\Pr[\|\hat{w} - w^\star\|_\Sigma^2 \le \sigma^2/10] = 1 - o(1)$$

where the probability is over the randomness of $\mathcal{A}$ and i.i.d. samples $(X^{(j)}, y^{(j)})_{j=1}^m$ from $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$.

Then there is a polynomial-time algorithm $\mathcal{A}'$ with the following property. For any $n, m, k \in \mathbb{N}$ and $\beta \in (-1, -1 + 1/(2k)]$, if $m \ge m_{\mathsf{SLR}}(n, k) + 1600 \log(n)$, then

$$\left| \Pr_{Z \sim \mathbb{P}_{n,k,\beta,m}} [\mathcal{A}'(Z) = 1] - \Pr_{Z \sim \mathbb{Q}_{n,m}} [\mathcal{A}'(Z) = 1] \right| = 1 - o(1). \tag{6}$$

The idea behind the reduction is to check (using the sparse linear regression algorithm $\mathcal{A}$) whether any covariate in the given sparse PCA data can be explained by the other covariates better than one would expect under the null distribution. Concretely, for a sample $Z \sim N(0, I_n + \beta w w^\top)$, for any i in the support of w, it can be observed that $\mathbb{E}[Z_i \mid Z_{\sim i}]$ is a $(k-1)$-sparse linear combination of the remaining covariates, and Z_i has conditional variance

$$\sigma^2 := \mathrm{Var}(Z_i \mid Z_{\sim i}) = \frac{1 + \beta}{1 + \beta(1 - 1/k)}.$$

In the near-critical regime $\beta \in (-1, -1 + 1/(2k)]$, we have $\sigma^2 \le 1/2$. Hence, using $\mathcal{A}$, we can distinguish from the null hypothesis that our samples are drawn from $N(0, I_n)$. If $\mathcal{A}$ were as statistically efficient as Best Subset Selection, then we could also solve the distinguishing problem with only $O(k \log n)$ samples. See Section C.2 for the full proof.

Remark 12 *It remains unknown whether an analogue of Theorem 7 can be proven under the Planted Clique Hypothesis (or any other standard average-case complexity hypothesis). One could hope to achieve such a result by reducing* positive-spike *sparse PCA to sparse linear regression. But in the above reduction, if $\beta > 0$ then the conditional variance of any i in the support of the spike is within $[1, 1 + 1/\Omega(k)]$, so even using the (computationally inefficient) guarantees for best subset selection, we would need $\Omega(k^2)$ samples to distinguish from the null hypothesis (c.f. Bresler et al. (2018)). Only in the near-critical negative spike regime do we get a sufficiently large gap in conditional variance for the hardness reduction to go through.*

Informally, the reason such a reduction fails to establish hardness is that the information-theoretically optimal algorithms for positive spike sparse PCA need to optimize simultaneously over both a sparsity *and* low rank *constraint on the covariance. Surprisingly, when we have a near-critical negative spike, using Best Subset Selection (which only enforces sparsity and has no explicit notion of low-rank structure) actually achieves the information-theoretic threshold.*

Remark 13 *The negative spike sparse PCA problem can be viewed as a real-valued analogue of the celebrated sparse parities with noise (SPN) problem. See Remark 30 for explanation.*

3. Related work

3.1. Algorithms

Sparse linear regression has been widely studied throughout fields such as statistics, theoretical computer science, and signal processing, see e.g. (Candes et al., 2006; Raskutti et al., 2010; Donoho and Stark, 1989; Donoho et al., 2005; Zhang et al., 2017), and (Wainwright, 2019, Section 7.7) for additional historical context. In the random-design model (Definition 1) we consider throughout the paper, it is well-known that the Best Subset Selection estimator (Hocking and Leslie, 1967) achieves prediction error $O(\sigma^2 k \log(n)/m)$ with high probability (Foster and George, 1994). However, this requires $O(n^k)$ time, and thus is prohibitively costly even for small k.[5]

Celebrated estimators based on ℓ_1-regularization (e.g. Lasso (Tibshirani, 1996) and the Dantzig Selector (Candes et al., 2007)) and greedy variable selection (e.g. Orthogonal Matching Pursuit (Cai and Wang, 2011)) are highly practical alternatives that can be computed in polynomial time. But the analyses of these estimators all require some additional assumption on Σ in order to achieve optimal sample complexity. The archetypal guarantee is the following: if $\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma) \le \kappa$, then the Lasso program

$$\hat{w} \leftarrow \operatorname*{argmin}_{w \in \mathbb{R}^n} \frac{1}{m} \sum_{j=1}^{m} (\langle X^{(j)}, w \rangle - y^{(j)})^2 + \lambda \|w\|_1 \tag{7}$$

achieves prediction error $O(\sigma^2 k \kappa \log(n)/m)$ (Wainwright, 2019, Theorems 7.16 + 7.20) with high probability over samples $(X^{(j)}, y^{(j)})_{j=1}^m \sim \mathsf{SLR}_{\Sigma,\sigma}(w^\star)$, for an appropriate choice of $\lambda > 0$.

Significant efforts have gone into establishing guarantees for Lasso under the weakest assumptions possible, leading to more general assumptions such as the Restricted Eigenvalue Condition (Bickel et al., 2009) and the compatibility condition (Van De Geer et al., 2009); see also the submodularity condition for Orthogonal Matching Pursuit (Das and Kempe, 2011). All of these assumptions boil down to some variant of a condition number bound. This is an inherent limitation of

5. The benefits of best subset selection have motivated major integer programming efforts, see e.g. Bertsimas et al. (2016); Hastie et al. (2020); in the worst case, it likely requires $n^{(1-o(1))k}$ time (Gupte and Vaikuntanathan, 2021).

Lasso and other classical estimators, which provably fail (i.e. have poor sample complexity) in the presence of strong, sparse linear dependencies among covariates (see Section 3.2 for more details and references). Such dependencies may easily occur in the settings that we study in this paper.

Preconditioned Lasso. Most closely related to this paper are the recent works (Kelner et al., 2021, 2023), which also identify natural structural assumptions on Σ under which Lasso may fail, but a more clever algorithm succeeds. In particular, Kelner et al. (2021) studied the setting where the covariates are drawn from a Gaussian Graphical Model with low treewidth. In this setting, they showed that there is a *preconditioned Lasso* program – namely, a modification of Equation (7) where the regularization term $\|w\|_1$ is replaced by $\left\|S^\top w\right\|_1$ for some matrix S – with near-optimal statistical performance, and that, given the graphical structure (i.e. the support of the precision matrix Σ^{-1}), the preconditioner S can be efficiently computed. Setting 1 is incomparable to the low treewidth assumption, and our algorithm does not assume any knowledge about the ground truth.

Kelner et al. (2023) introduced the setting where the spectrum of Σ may have a small number of outliers (which we refer to as Setting 2 above). Their main result is a polynomial-time algorithm that achieves prediction error $(k\lambda_{d+1}/\lambda_{n-d} + k^{O(k)}d)\log(n)/m$ with high probability, where $\lambda_1 \geq \cdots \geq \lambda_n$ are the eigenvalues of Σ, and $d \in \{1, \ldots, n-1\}$. Note the exponential dependence on k, so the previous result is essentially vacuous for $k = \Omega(\log n)$. In contrast, Corollary 6 incurs only polynomial dependence on k. Additionally, unlike the prior work, our algorithm does not need to be given Σ, which is a significant advantage in applications where the sample complexity is likely sublinear in n, and thus, the empirical covariance is a poor approximation for Σ. As one caveat, we remark that our algorithm does incur an additional factor of $\log(\lambda_1/\lambda_n)$ in the time complexity; we leave it as an interesting open problem whether this dependence can be removed.

3.2. Lower bounds

There is a long line of work studying lower bounds on the sample complexity of *specific* algorithms for sparse linear regression: see e.g. Van De Geer (2018), which shows that the sample complexity of Lasso can be lower bounded in terms of the compatibility constant (in the fixed-design setting where $X^{(1)}, \ldots, X^{(m)}$ may be arbitrary), and precise high-dimensional asymptotics for the exact performance of the Lasso (e.g. Bayati and Montanari (2011); Stojnic (2013); Amelunxen et al. (2014)). There are very simple covariate distributions where Lasso (with standard normalization) fails for some sparse signal: e.g., if two covariates are approximately equal, and the remaining $n-2$ covariates are independent, then even with zero noise, Lasso provably requires $\Omega(n)$ samples to learn the difference between the first two covariates (see the "weak compatibility condition" of (Kelner et al., 2021)). Of course, in these simple examples, while the Lasso algorithm may fail, there is no inherent computational obstruction (e.g., detecting two correlated covariates is straightforward).

Efforts to prove sample complexity lower bounds have been extended to parametric algorithm families like Lasso with coordinate-wise additively separable regularization (Zhang et al., 2017), and preconditioned Lasso programs (Kelner et al., 2021, 2022). It can be shown that SLR is NP-hard if the algorithm is forced to output a k-sparse estimate of the ground truth (note: when Σ is rank degenerate, this can be a much stronger requirement than achieving zero prediction error), see e.g. Natarajan (1995); Zhang et al. (2014); Gupte and Lu (2020); Gupte and Vaikuntanathan (2021); Foster et al. (2015) — but hardness of improper learning probably cannot be based on NP-hardness (see e.g. Applebaum et al. (2008)). Evidence has also been given that Gaussian

SLR with $\Sigma = I_n$ exhibits a constant-factor gap between the information-theoretic and algorithmic sample complexities, see e.g. Gamarnik and Zadik (2017); Arpino and Venkataramanan (2023). All together, the known lower bounds seem fundamentally different from the likely *exponential* gap in random-design, general-covariance sparse linear regression. Although our lower bound (Theorem 7) does not establish the anticipated exponential gap, it introduces the first polynomial gap for sparse linear regression under a reasonable computational assumption, and pins down the correct gap in our setting of latent variable models.

Sparse PCA. In the classical, positive-spike, sparse PCA detection problem, we are given independent samples from either $N(0, I_n + \beta xx^\top)$ or $N(0, I_n)$ for a random k-sparse unit vector x and some $\beta > 0$. The goal is to distinguish between these two settings. Negative-spike sparse PCA is the variant where $\beta \in [-1, 0)$; see Section 2.2 for a formal definition. We write "near-critical" as an informal term to denote the regime where β is very close to -1, because the problem is no longer well-defined when $\beta < -1$.

There is considerable evidence that any computationally efficient algorithm requires $\Omega(k^2)$ samples to solve the the positive-spike sparse PCA detection problem, with constant signal strength $\beta = \Theta(1)$ (Berthet and Rigollet, 2013a; Krauthgamer et al., 2015; Ma and Wigderson, 2015; Gao et al., 2017; Lu et al., 2018; Brennan and Bresler, 2019; Ding et al., 2023), even though the information-theoretic limit is only $\Theta(k \log n)$ samples (Berthet and Rigollet, 2013b). More generally, it's widely believed that computationally efficient algorithms with access to m samples can only perform detection for signal strength $\beta = \Omega(\sqrt{k^2/m})$.

In the negative-spike setting, the same computational/statistical tradeoff is conjectured to hold (where signal strength is now measured by $-\beta$). But this has only been rigorously proven (under a variant of the Planted Clique conjecture) for sparsity $k = o(m^{1/6})$, or equivalently $-\beta = o(m^{-1/3})$ (Brennan and Bresler, 2020). On the one hand, there appear to be considerable technical challenges in proving reduction-based hardness of near-critical negative-spike sparse PCA (Brennan and Bresler, 2020). On the other hand, understanding the complexity of negative-spike sparse PCA seems to be an important problem in the theory of average-case hardness — besides the new reduction in this work, previous work has connected the hardness of negative-spike sparse PCA to conjectured computational-statistical gaps in phase retrieval, mixed linear regression, and in certifying the RIP property (Brennan and Bresler, 2020; Ding et al., 2021). As discussed earlier, some known hardness results for *positive-spike* sparse PCA in restricted classes of algorithms can be adapted to the negative-spike setting — see Ding et al. (2021) for previous work away from the critical threshold, discussed further below, and our new results in Section C for low-degree and SDP hardness near the critical threshold.

Related problems: planted sparse vector and certifying RIP. A long line of works have studied both upper and lower bounds for the problem of finding a planted sparse vector in a random subspace — see Ding and Hua (2023); Barak et al. (2014); Demanet and Hand (2014); Mao and Wein (2021); Hopkins et al. (2016); Qu et al. (2014). As we mentioned when we introduced Algorithm 1, the convex program we solve in the inner loop of SMARTSCALING is similar to the convex relaxations used in Demanet and Hand (2014) and Spielman et al. (2012) to search for sparse vectors in a subspace. Such a program is also similar in spirit to pseudolikelihood methods used for learning sparse graphical models, see e.g. Besag (1977); Meinshausen et al. (2006); Kelner et al. (2020). The idea of solving such a program iteratively seems new to this work.

The most relevant lower bound in this line of work to us is the low-degree hardness result of Ding et al. (2021). They phrased their lower bound as one for the problem of certifying the RIP property[6] in the average case (c.f. Wang et al. (2016)), but as they discuss in their Remark II.4, their lower bound can also be interpreted as evidence for the optimality of the Demanet and Hand (2014) linear program among computationally efficient algorithms, in the setting where the random subspace has small codimension. Their hardness result for certifying RIP is established by proving a low-degree lower bound for a version of the negatively-spiked sparse PCA problem. The crucial difference between our setting and theirs is that they use a version of negatively-spiked sparse PCA where $\beta \in (-1, 0)$ is a *fixed constant* as the sparsity k of the spike goes to infinity, whereas for us it is crucial to consider the "near-critical" regime where $\beta \to -1$ as k goes to infinity. More precisely, our reduction to SLR operates in the regime $\beta < -1 + 1/\Omega(k)$; see the discussion around Remark 12. It is necessary that we take $\beta \to -1$, because when β is a fixed constant, the SLR problems arising in our reduction are all well-conditioned, so they can actually be solved with nearly-optimal sample complexity in polynomial time using Lasso (Raskutti et al., 2010) and cannot be the basis of a hardness result. At a technical level, taking $\beta \to -1$ also leads to a difference in the low-degree analysis — Ding et al. (2021) use that when β is fixed, an i.i.d. prior for the spike will satisfy $\beta\|x\|^2 < 1$ with high probability, but in the near-critical regime this is not true (see Remark 37).

Acknowledgments. We thank the reviewers for their helpful comments.

References

Achim Ahrens, Christian B Hansen, and Mark E Schaffer. lassopack: Model selection and prediction with regularized regression in stata. *The Stata Journal*, 20(1):176–235, 2020.

Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

Animashree Anandkumar, Vincent YF Tan, Furong Huang, and Alan S Willsky. High-dimensional gaussian graphical model selection: Walk summability and local separation criterion. *Journal of Machine Learning Research*, 13:2293–2337, 2012.

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

Gabriel Arpino and Ramji Venkataramanan. Statistical-computational tradeoffs in mixed sparse linear regression. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 921–986. PMLR, 2023.

6. As a reminder, the Restricted Isometry Property (RIP) is one that is sufficient, but not necessary, for methods such as the LASSO to achieve nearly-statistically optimal performance for sparse recovery. There are many randomized constructions of sensing matrices with good RIP properties, so that efficient sparse recovery/SLR is typically possible in such an ensemble. However, *certification* that a particular sensing matrix is good may be computationally difficult and is a longstanding challenge. For the particular Gaussian ensemble considered by Ding et al. (2021), the LASSO is statistically optimal up to constants (see e.g. Raskutti et al. (2010); Candes et al. (2007)), but their lower bound gives evidence certification is computationally hard.

Afonso S Bandeira, Dmitriy Kunisky, and Alexander S Wein. Computational hardness of certifying bounds on constrained pca problems. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

Boaz Barak, Jonathan A Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 31–40, 2014.

Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on learning theory*, pages 1046–1066. PMLR, 2013a.

Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013b.

Dimitris Bertsimas, Angela King, Rahul Mazumder, et al. Best subset selection via a modern optimization lens. *Annals of statistics*, 44(2):813–852, 2016.

Dimitris Bertsimas, Martin S Copenhaver, and Rahul Mazumder. Certifiably optimal low rank factor analysis. *The Journal of Machine Learning Research*, 18(1):907–959, 2017.

Julian Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, pages 616–618, 1977.

Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

Matthew Brennan and Guy Bresler. Optimal average-case reductions to sparse pca: From weak assumptions to strong hardness. In *Conference on Learning Theory*, pages 469–470. PMLR, 2019.

Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020.

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

Rares-Darius Buhai, Jingqiu Ding, and Stefan Tiegel. Computational-statistical gaps for improper learning in sparse linear regression. *arXiv preprint arXiv:2402.14103*, 2024.

T Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information theory*, 57(7):4680–4688, 2011.

Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of statistics*, 35(6):2313–2351, 2007.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Alexandre d'Aspremont, Laurent Ghaoui, Michael Jordan, and Gert Lanckriet. A direct formulation for sparse pca using semidefinite programming. *Advances in neural information processing systems*, 17, 2004.

Laurent Demanet and Paul Hand. Scaling law for recovering the sparsest element in a subspace. *Information and Inference: A Journal of the IMA*, 3(4):295–309, 2014.

Jingqiu Ding and Yiding Hua. Sq lower bounds for random sparse planted vector problem. In *International Conference on Algorithmic Learning Theory*, pages 558–596. PMLR, 2023.

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.

Yunzi Ding, Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Subexponential-time algorithms for sparse pca. *Foundations of Computational Mathematics*, pages 1–50, 2023.

David L Donoho and Philip B Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989.

David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.

Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009.

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

Dean P Foster and Edward I George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975, 1994.

David Gamarnik and Ilias Zadik. Sparse high-dimensional linear regression. algorithmic barriers and a local search algorithm. *arXiv preprint arXiv:1711.04952*, 2017.

Chao Gao, Zongming Ma, and Harrison H Zhou. Sparse cca: Adaptive estimation and computational barriers. *The Annals of Statistics*, pages 2074–2101, 2017.

Aparna Gupte and Vinod Vaikuntanathan. The fine-grained hardness of sparse linear regression. *arXiv preprint arXiv:2106.03131*, 2021.

Aparna Ajit Gupte and Kerri Lu. Fine-grained complexity of sparse linear regression. 2020.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.

Ronald R Hocking and RN Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540, 1967.

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, Cornell University, 2018.

Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191, 2016.

Harold Hotelling. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):193–232, 1953.

Rick H Hoyle. *Structural equation modeling: Concepts, issues, and applications*. Sage, 1995.

Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer Science & Business Media, 2010.

Jonathan Kelner, Frederic Koehler, Raghu Meka, and Ankur Moitra. Learning some popular gaussian graphical models without condition number bounds. *Advances in Neural Information Processing Systems*, 33:10986–10998, 2020.

Jonathan Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. On the power of preconditioning in sparse linear regression. *62nd Annual IEEE Symposium on Foundations of Computer Science*, 2021.

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, 2022.

Jonathan Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. Feature adaptation for sparse linear regression. *arXiv preprint arXiv:2305.16892*, 2023.

Frederic Koehler and Elchanan Mossel. Reconstruction on trees and low-degree polynomials. *Advances in Neural Information Processing Systems*, 35:18942–18954, 2022.

Robert Krauthgamer, Boaz Nadler, and Dan Vilenchik. Do semidefinite relaxations solve sparse pca up to the information limit? *Annals of Statistics*, 43(3):1300–1322, 2013.

Robert Krauthgamer, Boaz Nadler, and Dan Vilenchik. Do semidefinite relaxations solve sparse pca up to the information limit? *The Annals of Statistics*, 43(3):1300–1322, 2015.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 1–50. Springer, 2019.

Hao Lu, Yuan Cao, Zhuoran Yang, Junwei Lu, Han Liu, and Zhaoran Wang. The edge density barrier: Computational-statistical tradeoffs in combinatorial inference. In *International Conference on Machine Learning*, pages 3247–3256. PMLR, 2018.

Tengyu Ma and Avi Wigderson. Sum-of-squares lower bounds for sparse pca. *Advances in Neural Information Processing Systems*, 28, 2015.

Cheng Mao and Alexander S Wein. Optimal spectral recovery of a planted vector in a subspace. *arXiv preprint arXiv:2105.15081*, 2021.

Nicolai Meinshausen, Peter Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

Sidhant Misra, Marc Vuffray, and Andrey Y Lokhov. Information theoretic optimal learning of gaussian graphical models. *arXiv preprint arXiv:1703.04886*, 2017.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Qing Qu, Ju Sun, and John Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. *Advances in Neural Information Processing Systems*, 27, 2014.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

James Saunderson, Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1395–1416, 2012.

Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, pages 37–1. JMLR Workshop and Conference Proceedings, 2012.

Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Levent Tunçel, Stephen A Vavasis, and Jingye Xu. Computational complexity of decomposing a symmetric matrix as a sum of positive semidefinite and diagonal matrices. *Foundations of Computational Mathematics*, pages 1–47, 2023.

Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 11–20. IEEE, 2012.

Sara Van De Geer. On tight bounds for the lasso. *Journal of Machine Learning Research*, 19:46, 2018.

Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Tengyao Wang, Quentin Berthet, and Yaniv Plan. Average-case hardness of rip certification. *Advances in Neural Information Processing Systems*, 29, 2016.

Yilei Wu, Yingli Qin, and Mu Zhu. High-dimensional covariance matrix estimation using a low-rank and diagonal decomposition. *Canadian Journal of Statistics*, 48(2):308–337, 2020.

Ilias Zadik and David Gamarnik. High dimensional linear regression using lattice basis reduction. *Advances in Neural Information Processing Systems*, 31, 2018.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014.

Yuchen Zhang, Martin J Wainwright, Michael I Jordan, et al. Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators. *Electronic Journal of Statistics*, 11(1):752–799, 2017.

Lijia Zhou, Frederic Koehler, Danica J Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *arXiv preprint arXiv:2112.04470*, 2021.

Appendix A. The rescaled Lasso

In this section we prove Theorem 4. We start by analyzing the procedure `SmartScaling()`.

Lemma 14 (Restatement of Lemma 8) *Let $n, m, k \in \mathbb{N}$. Let $\mathbb{X} \in \mathbb{R}^{m \times n}$. Suppose that*

$$I_n \preceq_{\mathcal{C}_n(32k)} D^{-1/2} \hat{\Sigma} D^{-1/2} \tag{8}$$

where $D \succ 0$ is a diagonal matrix, and $\hat{\Sigma} := \frac{1}{m}\mathbb{X}^\top \mathbb{X}$. Then the algorithm `SmartScaling`$(\mathbb{X}, k)$ *terminates after at most $T := n \log \max_{i \in [n]} \frac{2\hat{\Sigma}_{ii}}{D_{ii}}$ repetitions, and moreover the output $\hat{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix satisfying the following properties:*

- $\hat{D} \succeq \frac{1}{2} D$.
- $\frac{1}{m}\|\mathbb{X}v\|_2^2 > 1$ *for all* $v \in \mathbb{R}^n$ *with* $\left\|\hat{D}^{1/2} v\right\|_\infty = 1$ *and* $\left\|\hat{D}^{1/2} v\right\|_1 \leq 16k$.

Proof For notational convenience, let t_{final} be the step at which the algorithm returns the preconditioner.

Spectral lower bound. First, we show by induction that for each $1 \leq t \leq t_{\mathsf{final}}$, the intermediate result $\hat{D}^{(t)}$ satisfies $\hat{D}^{(t)} \succeq \frac{1}{2} D$. From Definition 2, note that $\mathcal{C}_n(32k)$ contains the standard basis vectors. Thus, we get from Equation (8) that $1 \leq \hat{\Sigma}_{ii}/D_{ii}$ for all $i \in [n]$. Hence, $\hat{D}^{(1)} = \mathrm{diag}(\hat{\Sigma}) \succeq D$. This proves the base case of the induction.

Now fix any $1 \leq t < t_{\mathsf{final}}$ and suppose that $\hat{D}^{(t)} \succeq \frac{1}{2} D$. We want to prove that $\hat{D}^{(t)}_{i_{\mathsf{min}} i_{\mathsf{min}}} \geq D_{i_{\mathsf{min}} i_{\mathsf{min}}}$. Suppose for the sake of contradiction that in fact

$$\hat{D}^{(t)}_{i_{\mathsf{min}} i_{\mathsf{min}}} < D_{i_{\mathsf{min}} i_{\mathsf{min}}} \tag{9}$$

Then

$$\begin{aligned}
\left\|D^{1/2} v^{(t, i_{\mathsf{min}})}\right\|_\infty &\geq D^{1/2}_{i_{\mathsf{min}} i_{\mathsf{min}}} |v^{(t, i_{\mathsf{min}})}_{i_{\mathsf{min}}}| \\
&\geq (\hat{D}^{(t)})^{1/2}_{i_{\mathsf{min}} i_{\mathsf{min}}} |v^{(t, i_{\mathsf{min}})}_{i_{\mathsf{min}}}| \\
&\geq \frac{1}{16k} \left\|(\hat{D}^{(t)})^{1/2} v^{(t, i_{\mathsf{min}})}\right\|_1 \\
&\geq \frac{1}{32k} \left\|D^{1/2} v^{(t, i_{\mathsf{min}})}\right\|_1
\end{aligned}$$

where the second inequality uses the assumption Equation (9), the third inequality uses the constraints in the program that defines $v^{(t, i_{\mathsf{min}})}$ (Equation (4)), and the fourth inequality uses the induction hypothesis. It follows that $D^{1/2} v^{(t, i_{\mathsf{min}})} \in \mathcal{C}_n(32k)$. So by Equation (8), we get that $\left\|v^{(t, i_{\mathsf{min}})}\right\|_D^2 \leq \frac{1}{m}\left\|\mathbb{X} v^{(t, i_{\mathsf{min}})}\right\|_2^2$. Moreover since $t \neq t_{\mathsf{final}}$, we have that $\frac{1}{m}\left\|\mathbb{X} v^{(t, i_{\mathsf{min}})}\right\|_2^2 \leq 1$. Thus,

$$\hat{D}^{(t)}_{i_{\mathsf{min}} i_{\mathsf{min}}} \cdot (v^{(t, i_{\mathsf{min}})}_{i_{\mathsf{min}}})^2 = 1 \geq \frac{1}{m}\left\|\mathbb{X} v^{(t, i_{\mathsf{min}})}\right\|_2^2 \geq \left\|v^{(t, i_{\mathsf{min}})}\right\|_D^2 \geq D_{i_{\mathsf{min}} i_{\mathsf{min}}} \cdot (v^{(t, i_{\mathsf{min}})}_{i_{\mathsf{min}}})^2.$$

By Equation (4) we know that $v^{(t, i_{\mathsf{min}})}_{i_{\mathsf{min}}} \neq 0$. Simplifying the above display therefore gives that $\hat{D}^{(t)}_{i_{\mathsf{min}} i_{\mathsf{min}}} \geq D_{i_{\mathsf{min}} i_{\mathsf{min}}}$. This contradicts the assumption that $\hat{D}^{(t)}_{i_{\mathsf{min}} i_{\mathsf{min}}} < D_{i_{\mathsf{min}} i_{\mathsf{min}}}$, so in fact $\hat{D}^{(t)}_{i_{\mathsf{min}} i_{\mathsf{min}}} \geq D_{i_{\mathsf{min}} i_{\mathsf{min}}}$ holds unconditionally. By definition of $\hat{D}^{(t+1)}$ and the induction hypothesis, we get $\hat{D}^{(t+1)} \succeq \frac{1}{2} D$, which completes the induction.

Repetition bound. Note that by definition we have $\det(\hat{D}^{(t)}) = 2^{1-t}\det(\hat{D}^{(1)})$ for all $1 \le t \le t_{\mathsf{final}}$. Suppose that the algorithm requires more than T repetitions, i.e. $t_{\mathsf{final}} \ge T+1$. Then

$$\det(\hat{D}^{(T+1)}) = 2^{-T}\det(\hat{D}^{(1)}) < \left(\min_{i\in[n]} \frac{D_{ii}}{2\hat{\Sigma}_{ii}}\right)^n \det(\hat{D}^{(1)})$$

by definition of T. But we have already seen that $\hat{D}^{(T+1)} \succeq \frac{1}{2}D$, so on the other hand

$$\frac{\det(\hat{D}^{(T+1)})}{\det(\hat{D}^{(1)})} \ge 2^{-n}\prod_{i=1}^n \frac{D_{ii}}{\hat{\Sigma}_{ii}} \ge \left(\min_{i\in[n]} \frac{D_{ii}}{2\hat{\Sigma}_{ii}}\right)^n.$$

This is a contradiction, so in fact the algorithm terminates after at most T repetitions.

Restricted eigenvalue bound. The output of the algorithm is $\hat{D} := \hat{D}^{(t_{\mathsf{final}})}$. Fix any $v \in \mathbb{R}^n$ with $\left\|\hat{D}^{1/2}v\right\|_\infty = 1$ and $\left\|\hat{D}^{1/2}v\right\|_1 \le 16k$. Since t_{final} is the final step of the algorithm, by the termination condition it must be that $\frac{1}{m}\left\|\mathbb{X}v^{(t_{\mathsf{final}},i_{\mathsf{min}})}\right\|_2^2 > 1$ and thus, since v is feasible for Equation (4) for some $i \in [n]$, $\frac{1}{m}\|\mathbb{X}v\|_2^2 > 1$ as well. ■

Notation. For the remainder of Section A, we fix the following notation. Let $n, m, k, h \in \mathbb{N}$ and $\alpha > 0$. Let $\Sigma \in \mathbb{R}^{n\times n}$ be a positive semi-definite matrix. We make the assumption that Σ is (α, h)-rescalable at sparsity k (Definition 3), i.e.

$$I_n \preceq_{\mathcal{C}_n(32k)} D^{-1/2}\Sigma D^{-1/2} \preceq \alpha I_n + L \tag{10}$$

where $D \succ 0$ is some diagonal matrix, and $L \succeq 0$ is some rank-h matrix.

Let $\sigma > 0$, and let $w^\star \in \mathbb{R}^n$ be k-sparse with support S. Let $(X^{(j)}, y^{(j)})_{j=1}^m$ be m independent samples from $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$. Let $\mathbb{X}$ be the $m \times n$ matrix with rows $X^{(1)}, \dots, X^{(m)}$, and let $\hat{\Sigma} := \frac{1}{m}\mathbb{X}^\top\mathbb{X}$.

A.1. The good event: generalization, concentration, and noise bounds

The following definition states the event $\mathcal{E}_{15}(\delta)$ under which we will show that `RescaledLasso()` (deterministically) has low prediction error. The event consists of three conditions, of which the first states that the covariates have accurate sample variances, the second is a uniform generalization bound, and the third bounds the bias of the noise term.

Definition 15 *Let $\delta > 0$. Let $\mathcal{E}_{15}(\delta)$ be the event (over the samples $(X^{(j)}, y^{(j)})_{j=1}^m$) that the following properties hold:*

- *For all $i \in [n]$,*

$$\frac{1}{2}\Sigma_{ii} \le \hat{\Sigma}_{ii} \le 2\Sigma_{ii}. \tag{11a}$$

- *For all $w \in \mathbb{R}^n$,*

$$\|w\|_\Sigma^2 \le 16\|w\|_{\hat{\Sigma}}^2 + \frac{16\alpha\log(48n/\delta)}{m}\left\|D^{1/2}w\right\|_1^2. \tag{11b}$$

- *For all* $w \in \mathbb{R}^n$,

$$\frac{1}{m}\langle w, \mathbb{X}^\top(y - \mathbb{X}w^\star)\rangle \le \frac{\sigma}{\sqrt{m}}\left(2\left\|D^{1/2}w\right\|_1\sqrt{\alpha\log(24n/\delta)} + \|w\|_\Sigma\sqrt{2C_{17}h\log(24/\delta)}\right). \tag{11c}$$

Our first step is to show that $\mathcal{E}_{15}(\delta)$ holds with probability at least $1-\delta$. The first property (11a) is standard and requires no additional assumptions on Σ, but the second and third properties both crucially use the spectral upper bound $D^{-1/2}\Sigma D^{-1/2} \preceq \alpha I_n + L$ guaranteed by rescalability. In particular, this upper bound implies (by Weyl's inequality and the fact that L is low-rank) that $D^{-1/2}\Sigma D^{-1/2}$ has at most h eigenvalues larger than α. We can bound the corresponding eigenspaces separately to obtain the desired generalization bounds, a technique also applied in Kelner et al. (2023) to deal with large outlier eigenvalues (and previously in other contexts – see the discussion in Zhou et al. (2021) on "covariance splitting").

Concretely, the following theorem due to Zhou et al. (2021) shows that to prove a uniform generalization bound (e.g. of the form (11b)), it suffices to provide a uniform high-probability bound on $\sup_{w\in\mathbb{R}^n}\langle w, x\rangle$ for $X \sim N(0,\Sigma)$. In Lemma 17, we use the technique of covariance splitting to derive such a bound for the rescalable setting. In Lemma 18, we then invoke Theorem 16 to prove (11b).

Theorem 16 (Theorem 1 in Zhou et al. (2021)) *Let* $n, m \in \mathbb{N}$ *and* $\epsilon, \delta > 0$. *Let* $\Sigma \in \mathbb{R}^{n\times n}$ *be a positive semi-definite matrix. Let* $\mathbb{X} \in \mathbb{R}^{m\times n}$ *have i.i.d. rows* $X_1, \dots, X_m \sim N(0,\Sigma)$. *Let* $F : \mathbb{R}^n \to [0,\infty]$ *be a continuous function such that*

$$\Pr_{x\sim N(0,\Sigma)}[\sup_{w\in\mathbb{R}^n}\langle w, x\rangle - F(w) > 0] \le \delta.$$

If $m \ge 196\epsilon^{-2}\log(12/\delta)$, *then with probability at least* $1-4\delta$ *it holds that for all* $w \in \mathbb{R}^n$,

$$\|w\|_\Sigma^2 \le \frac{1+\epsilon}{m}\left(\|Xw\|_2 + F(w)\right)^2.$$

Lemma 17 *There is a constant* C_{17} *with the following property. Let* $n, h \in \mathbb{N}$ *and* $\alpha > 0$, *and let* $\Sigma \in \mathbb{R}^{n\times n}$ *be a positive semi-definite matrix. Suppose that* $\Sigma \preceq \alpha D + L$ *for some* $D, L \in \mathbb{R}^{n\times n}$ *where* $D \succ 0$ *is diagonal and* $L \succeq 0$ *has rank at most* h. *Then*

$$\Pr_{G\sim N(0,\Sigma)}\left[\forall w \in \mathbb{R}^n : \langle w, G\rangle \le \left\|D^{1/2}w\right\|_1\sqrt{2\alpha\log(16n/\delta)} + \|w\|_\Sigma\sqrt{Ch\log(16/\delta)}\right] \ge 1-\delta.$$

Proof Define the eigendecomposition of $D^{-1/2}\Sigma D^{-1/2}$ as $\sum_{i=1}^n \lambda_i u_i u_i^\top$ where $\lambda_1 \ge \cdots \ge \lambda_n$. By assumption we have $D^{-1/2}\Sigma D^{-1/2} \preceq \alpha I_n + D^{-1/2}LD^{-1/2}$, and thus $D^{-1/2}(\Sigma - L)D^{-1/2} \preceq \alpha I_n$. Since $\operatorname{rank}(D^{-1/2}LD^{-1/2}) \le h$, Weyl's inequality (Lemma 54) gives $\lambda_{h+1} \le \alpha$.

Define the orthogonal projection matrix $P := \sum_{i=h+1}^n u_i u_i^\top$. Let $G \sim N(0,\Sigma)$. By the Gaussian maximal inequality, with probability at least $1-\delta/2$ over the draw G, we have

$$\begin{aligned}\left\|PD^{-1/2}G\right\|_\infty &\le \sqrt{\lambda_{\max}(PD^{-1/2}\Sigma D^{-1/2}P)\cdot 2\log(4n/\delta)} \\ &\le \sqrt{2\lambda_{h+1}\log(4n/\delta)}\end{aligned}$$

$$\leq \sqrt{2\alpha \log(4n/\delta)}$$

where the first inequality follows from the definition of P. Next, by Corollary 53, if C_{17} is chosen to be a sufficiently large constant, then with probability at least $1 - \delta/2$,

$$\begin{aligned}\left\|\Sigma^{-1/2} D^{1/2} P^{\perp} D^{-1/2} G\right\|_2 &\leq \sqrt{\mathrm{tr}(\Sigma^{-1/2} D^{1/2} P^{\perp} D^{-1/2} \Sigma D^{-1/2} P^{\perp} D^{1/2} \Sigma^{-1/2}) \cdot C_{17} \log(4/\delta)} \\ &= \sqrt{\mathrm{tr}(\Sigma^{1/2} D^{-1/2} P^{\perp} D^{1/2} \Sigma^{-1/2}) \cdot C_{17} \log(4/\delta)} \\ &= \sqrt{\mathrm{tr}(P^{\perp}) \cdot C_{17} \log(4/\delta)} \\ &= \sqrt{C_{17} h \log(4/\delta)}\end{aligned}$$

where the first equality uses the fact that $P^{\perp} = I_n - P$ commutes with $D^{-1/2} \Sigma D^{-1/2}$, and the second equality uses the cyclic property of trace. Consider the event (which occurs with probability at least $1 - \delta$ over the draw G) that both of the above bounds hold. Then for any $w \in \mathbb{R}^n$, we have

$$\begin{aligned}\langle w, G \rangle &= \langle D^{1/2} w, D^{-1/2} G \rangle \\ &= \langle D^{1/2} w, P D^{-1/2} G \rangle + \langle D^{1/2} w, P^{\perp} D^{-1/2} G \rangle \\ &= \langle D^{1/2} w, P D^{-1/2} G \rangle + \langle \Sigma^{1/2} w, \Sigma^{-1/2} D^{1/2} P^{\perp} D^{-1/2} G \rangle \\ &\leq \left\|D^{1/2} w\right\|_1 \sqrt{2\alpha \log(4n/\delta)} + \|w\|_{\Sigma} \sqrt{C_{17} h \log(4/\delta)}.\end{aligned}$$

as needed. ■

Lemma 18 *Let $\Sigma \in \mathbb{R}^{n \times n}$ satisfy $\Sigma \preceq \alpha D + L$ for some diagonal matrix $D \succ 0$ and rank-h matrix L. Let $\mathbb{X} \in \mathbb{R}^{m \times n}$ have i.i.d. rows $X_1, \ldots, X_m \sim N(0, \Sigma)$. If $m \geq C(h+1)\log(96/\delta)$ for a sufficiently large constant C, then with probability at least $1 - \delta$, it holds that for all $w \in \mathbb{R}^n$,*

$$\|w\|_{\Sigma}^2 \leq \frac{16}{m} \left(\|\mathbb{X} w\|_2^2 + \alpha \left\|D^{1/2} w\right\|_1^2 \log(16n/\delta) \right).$$

Proof By Lemma 17 and the fact that $m \geq 196 \log(48/\delta)$, the hypothesis of Theorem 16 is satisfied $\epsilon := 1$, error probability $\delta/4$, and the functional F defined as

$$F(w) := \left\|D^{1/2} w\right\|_1 \sqrt{2\alpha \log(16n/\delta)} + \|w\|_{\Sigma} \sqrt{C_{17} h \log(16/\delta)}.$$

The conclusion of Theorem 16 gives that with probability at least $1 - \delta$, the following holds. For all $w \in \mathbb{R}^n$,

$$\begin{aligned}\|w\|_{\Sigma}^2 &\leq \frac{2}{m} \left(\|\mathbb{X} w\|_2 + \left\|D^{1/2} w\right\|_1 \sqrt{2\alpha \log(16n/\delta)} + \|w\|_{\Sigma} \sqrt{C_{17} h \log(16/\delta)} \right)^2 \\ &\leq \frac{8}{m} \left(\|\mathbb{X} w\|_2^2 + \alpha \left\|D^{1/2} w\right\|_1^2 \log(16n/\delta) + \|w\|_{\Sigma}^2 \cdot C_{17} h \log(16/\delta) \right) \\ &\leq \frac{16}{m} \left(\|\mathbb{X} w\|_2^2 + \alpha \left\|D^{1/2} w\right\|_1^2 \log(16n/\delta) \right)\end{aligned}$$

where the last inequality holds by rearranging terms and using the fact that $m \geq 16C_{17}h\log(16/\delta)$ (so long as C is chosen sufficiently large). ■

We now prove that $\mathcal{E}_{15}(\delta)$ holds with probability at least $1-\delta$, observing that Lemma 17 is exactly what is needed to prove the third property (11c).

Lemma 19 *There is a constant C_{19} so that the following holds. Let $\delta > 0$, and suppose that $m \geq C_{19}(h+1)\log(288n/\delta)$. Then $\Pr[\mathcal{E}_{15}(\delta)] \geq 1-\delta$.*

Proof Since $m \geq 32\log(6n/\delta)$, we have by Lemma 51 and a union bound that Equation (11a) holds for all $i \in [n]$, with probability at least $1-\delta/3$. By Lemma 18, so long as C_{19} is a sufficiently large constant, we have that Equation (11b) holds for all $w \in \mathbb{R}^n$, with probability at least $1-\delta/3$.

It remains to prove Equation (11c). Define the random variable $\xi := y - \mathbb{X}w^\star$. Since $\|\xi\|_2^2 \sim \sigma^2\chi^2_m$, and $m \geq 8\log(12/\delta)$, it holds with probability at least $1-\delta/6$ that $\frac{1}{\sqrt{m}}\|\xi\|_2 \leq \sigma\sqrt{2}$. Condition on ξ and suppose that this event holds. Since ξ is, by construction, independent of $\mathbb{X}$, the random variable $\mathbb{X}^\top\xi$ has distribution $N(0, \|\xi\|_2^2\,\Sigma)$. Thus, by Lemma 17, we have with probability at least $1-\delta/6$ over the randomness of $\mathbb{X}$ that for all $w \in \mathbb{R}^n$,

$$\left\langle w, \frac{\mathbb{X}^t\xi}{\|\xi\|_2}\right\rangle \leq \left\|D^{1/2}w\right\|_1\sqrt{2\alpha\log(24n/\delta)} + \|w\|_\Sigma\sqrt{C_{17}h\log(24/\delta)}.$$

Substituting in the bound on $\|\xi\|_2$, we get that with probability at least $1-\delta/3$ (over $\mathbb{X}$ and ξ), for all $w \in \mathbb{R}^n$,

$$\frac{1}{m}\langle w, \mathbb{X}^\top\xi\rangle \leq 2\sigma\left\|D^{1/2}w\right\|_1\sqrt{\frac{\alpha\log(24n/\delta)}{m}} + \sigma\|w\|_\Sigma\sqrt{\frac{2C_{17}h\log(24/\delta)}{m}} \tag{12}$$

which proves Equation (11c). By the union bound, $\mathcal{E}_{15}(\delta)$ holds with probability at least $1-\delta$. ■

A.2. Bounding the ℓ_1 norm of the error

Next, we assume that $\mathcal{E}_{15}(\delta)$ holds, and that the rescaling matrix $\hat{D}$ is (approximately) lower bounded by the oracle rescaling matrix D. Under these conditions, we derive a win-win (Lemma 21) where either $\hat{D}^{1/2}(\hat{w}-w^\star)$ has small ℓ_1 norm (where $\hat{w}$ is the rescaled Lasso solution) or $\hat{w}-w^\star$ is a violation to the second guarantee of Lemma 14.

Why is this a win-win? Ultimately, we need to bound the population variance $\|e\|_\Sigma$ of the error $e := \hat{w}-w^\star$ in terms of the empirical variance $\|e\|_{\hat{\Sigma}}$, and also bound the empirical variance (which would be trivially zero in the noiseless setting, but not in general). The first goal is partially accomplished by the generalization bound Equation (11b), but it remains to bound $\|\hat{D}^{1/2}e\|_1$. The second goal also requires bounding $\|\hat{D}^{1/2}e\|_1$ in terms of $\|e\|_{\hat{\Sigma}}$ (to bound the bias induced by the regularization term of the Lasso program). Lemma 21 will allow us to achieve both of these goals.

We first need the following technical lemma, which should be thought of as a *cone condition* for the error e (compare to the noiseless case, where the analogous inequality is $\|\hat{D}^{1/2}e\|_1 \leq 2\|\hat{D}^{1/2}e_S\|_1 \leq 2k\|\hat{D}^{1/2}e\|_\infty$):

Lemma 20 *Let $\delta, \lambda > 0$. Let $\hat{D} \in \mathbb{R}^{n\times n}$ be a positive-definite diagonal matrix, and let $\hat{w}$ be a solution to the modified Lasso program*

$$\hat{w} \in \operatorname*{argmin}_{w\in\mathbb{R}^n} \frac{1}{m}\|\mathbb{X}w - y\|_2^2 + \lambda \left\|\hat{D}^{1/2}w\right\|_1. \tag{13}$$

If $\lambda \geq 64\sigma\sqrt{\frac{\alpha\log(24n/\delta)}{m}}$, event $\mathcal{E}_{15}(\delta)$ holds, and $D \preceq 64\hat{D}$, then $e := \hat{w} - w^\star$ satisfies

$$\lambda\left\|\hat{D}^{1/2}e\right\|_1 + \frac{2}{m}\|\mathbb{X}e\|_2^2 \leq 4\lambda\left\|\hat{D}^{1/2}e_S\right\|_1 + 4\sigma\|e\|_\Sigma\sqrt{\frac{2C_{17}h\log(24/\delta)}{m}}.$$

Proof First observe that

$$\begin{aligned}
\frac{1}{m}\|\mathbb{X}\hat{w} - y\|_2^2 + \lambda\left\|\hat{D}^{1/2}(\hat{w} - w^\star)_{S^c}\right\|_1 &= \frac{1}{m}\|\mathbb{X}\hat{w} - y\|_2^2 + \lambda\left\|\hat{D}^{1/2}\hat{w}_{S^c}\right\|_1 \\
&\leq \frac{1}{m}\|\mathbb{X}w^\star - y\|_2^2 + \lambda\left\|\hat{D}^{1/2}w^\star\right\|_1 - \lambda\left\|\hat{D}^{1/2}\hat{w}\right\|_1 + \lambda\left\|\hat{D}^{1/2}\hat{w}_{S^c}\right\|_1 \\
&= \frac{1}{m}\|\mathbb{X}w^\star - y\|_2^2 + \lambda\left\|\hat{D}^{1/2}w^\star\right\|_1 - \lambda\left\|\hat{D}^{1/2}\hat{w}_S\right\|_1 \\
&\leq \frac{1}{m}\|\mathbb{X}w^\star - y\|_2^2 + \lambda\left\|\hat{D}^{1/2}(\hat{w} - w^\star)_S\right\|_1 \qquad (14)
\end{aligned}$$

where the first inequality is by optimality of $\hat{w}$ in Equation (13) and the second inequality is by reverse triangle inequality. It follows that

$$\begin{aligned}
\lambda\left\|\hat{D}^{1/2}e\right\|_1 &= \lambda\left\|\hat{D}^{1/2}e_S\right\|_1 + \lambda\left\|\hat{D}^{1/2}e_{S^c}\right\|_1 \\
&\leq \frac{1}{m}\|\mathbb{X}w^\star - y\|_2^2 - \frac{1}{m}\|\mathbb{X}\hat{w} - y\|_2^2 + 2\lambda\left\|\hat{D}^{1/2}e_S\right\|_1 \\
&= -\frac{1}{m}\|\mathbb{X}e\|_2^2 - \frac{2}{m}\langle\mathbb{X}e, \mathbb{X}w^\star - y\rangle + 2\lambda\left\|\hat{D}^{1/2}e_S\right\|_1
\end{aligned}$$

where the first inequality uses Equation (14), and the second equality expands $\|\mathbb{X}\hat{w} - y\|_2^2 = \|\mathbb{X}e + (\mathbb{X}w^\star - y)\|_2^2$. Now since the event $\mathcal{E}_{15}(\delta)$ was assumed to hold, we can apply Equation (11c) with vector e to get that

$$\begin{aligned}
-\frac{2}{m}\langle\mathbb{X}e, \mathbb{X}w^\star - y\rangle &\leq 4\sigma\left\|D^{1/2}e\right\|_1\sqrt{\frac{\alpha\log(24n/\delta)}{m}} + 2\sigma\|e\|_\Sigma\sqrt{\frac{2C_{17}h\log(24/\delta)}{m}} \\
&\leq \frac{\lambda}{2}\left\|\hat{D}^{1/2}e\right\|_1 + 2\sigma\|e\|_\Sigma\sqrt{\frac{2C_{17}h\log(24/\delta)}{m}}
\end{aligned}$$

where the second inequality uses the lemma assumptions that $\lambda \geq 64\sigma\sqrt{\frac{\alpha\log(24n/\delta)}{m}}$ and $D \preceq 64\hat{D}$ (and D, $\hat{D}$ are diagonal). Substituting into the previous display and rearranging terms gives that

$$\frac{\lambda}{2}\left\|\hat{D}^{1/2}e\right\|_1 + \frac{1}{m}\|\mathbb{X}e\|_2^2 \leq 2\lambda\left\|\hat{D}^{1/2}e_S\right\|_1 + 2\sigma\|e\|_\Sigma\sqrt{\frac{2C_{17}h\log(24/\delta)}{m}}$$

which completes the proof. ■

Lemma 21 *Let $\delta, \lambda > 0$. Let $\hat{D} \in \mathbb{R}^{n\times n}$ be a positive-definite diagonal matrix, and let $\hat{w}$ be a solution to the modified Lasso program Equation (13). Suppose that event $\mathcal{E}_{19}(\delta)$ holds, that $D \preceq 64\hat{D}$, that $m \geq 128C_{17}h\log(24/\delta)$, and that $\lambda \geq \max\left(64\sigma\sqrt{\frac{\alpha\log(48n/\delta)}{m}}, \frac{32\sigma\sqrt{2C_{17}h\log(24/\delta)}}{k\sqrt{m}}\right)$. Then $e := \hat{w} - w^\star$ satisfies at least one of the following properties:*

(a) $\left\|\hat{D}^{1/2}e\right\|_1 \leq \frac{20k}{\sqrt{m}}\left\|\mathbb{X}e\right\|_2$, *or*

(b) $\frac{1}{m}\left\|\mathbb{X}e\right\|_2^2 \leq \left\|\hat{D}^{1/2}e\right\|_\infty^2$ *and* $\hat{D}^{1/2}e \in \mathcal{C}_n(9k)$.

Proof By Lemma 20 (dropping the second term on the left-hand side), we have

$$\begin{aligned}
\lambda\left\|\hat{D}^{1/2}e\right\|_1 &\leq 4\lambda\left\|\hat{D}^{1/2}e_S\right\|_1 + 4\sigma\left\|e\right\|_\Sigma\sqrt{\frac{2C_{17}h\log(24/\delta)}{m}} \\
&\leq 4k\lambda\left\|\hat{D}^{1/2}e\right\|_\infty + \frac{16\sigma\sqrt{2C_{17}h\log(24/\delta)}}{m}\left(\left\|\mathbb{X}e\right\|_2 + \left\|D^{1/2}e\right\|_1\sqrt{\alpha\log(48n/\delta)}\right) \\
&\leq 4k\lambda\left\|\hat{D}^{1/2}e\right\|_\infty + \frac{16\sigma\sqrt{2C_{17}h\log(24/\delta)}}{m}\left\|\mathbb{X}e\right\|_2 + \frac{\lambda}{32}\left\|D^{1/2}e\right\|_1
\end{aligned}$$

where the second inequality uses k-sparsity of e_S to bound $\left\|\hat{D}^{1/2}e_S\right\|_1$, and Equation (11b) (along with the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$) to bound $\left\|e\right\|_\Sigma$; and the third inequality uses the lemma assumptions that $\lambda \geq 64\sigma\sqrt{\frac{\alpha\log(48n/\delta)}{m}}$ and $m \geq 128C_{17}h\log(24/\delta)$. Since $D \preceq 64\hat{D}$ and D, $\hat{D}$ are diagonal, we have $\left\|D^{1/2}e\right\|_1 \leq 8\left\|\hat{D}^{1/2}e\right\|_1$. Substituting this bound above and rearranging terms gives

$$\begin{aligned}
\lambda\left\|\hat{D}^{1/2}e\right\|_1 &\leq 8k\lambda\left\|\hat{D}^{1/2}e\right\|_\infty + \frac{32\sigma\sqrt{2C_{17}h\log(24/\delta)}}{m}\left\|\mathbb{X}e\right\|_2 \\
&\leq 8k\lambda\left\|\hat{D}^{1/2}e\right\|_\infty + \frac{k\lambda}{\sqrt{m}}\left\|\mathbb{X}e\right\|_2 \qquad (15)
\end{aligned}$$

where the second inequality is by the lemma assumption that $\lambda \geq \frac{32\sigma\sqrt{2C_{17}h\log(24/\delta)}}{k\sqrt{m}}$. Now suppose that property (a) of the lemma statement fails to hold, i.e.

$$\frac{1}{\sqrt{m}}\left\|\mathbb{X}e\right\|_2 < \frac{1}{20k}\left\|\hat{D}^{1/2}e\right\|_1.$$

Substituting into the right-hand side of Equation (15) and rearranging terms gives $\lambda\left\|\hat{D}^{1/2}e\right\|_1 \leq 9k\lambda\left\|\hat{D}^{1/2}e\right\|_\infty$. As a result, we have $\hat{D}^{1/2}e \in \mathcal{C}_n(9k)$ and also

$$\frac{1}{m}\left\|\mathbb{X}e\right\|_2^2 \leq \frac{1}{400k^2}\left\|\hat{D}^{1/2}e\right\|_1^2 \leq \left\|\hat{D}^{1/2}e\right\|_\infty^2.$$

Thus, property (b) of the lemma statement holds. ■

A.3. Putting everything together

The following lemma shows that under the good event $\mathcal{E}_{15}(\delta)$ (which, as we've shown, occurs with high probability), the assumed spectral lower bound on Σ in Equation (10) transfers to a lower bound on the empirical covariance $\hat{\Sigma}$.

Lemma 22 *Let $\delta > 0$ and suppose that $m \geq 2^{15}\alpha k^2 \log(48n/\delta)$. In the event $\mathcal{E}_{15}(\delta)$, we have that*

$$I_n \preceq_{\mathcal{C}_n(32k)} 32D^{-1/2}\hat{\Sigma}D^{-1/2}. \tag{16}$$

Proof Fix $v \in \mathcal{C}_n(32k)$. Applying Equation (11b) to the vector $w := D^{-1/2}v$, we have

$$\begin{aligned}
\left\|D^{-1/2}v\right\|_{\Sigma}^2 &\leq 16\left\|D^{-1/2}v\right\|_{\hat{\Sigma}}^2 + \frac{16\alpha\log(48n/\delta)}{m}\|v\|_1^2 \\
&\leq 16\left\|D^{-1/2}v\right\|_{\hat{\Sigma}}^2 + \frac{2^{14}\alpha k^2\log(48n/\delta)}{m}\|v\|_2^2 \\
&\leq 16\left\|D^{-1/2}v\right\|_{\hat{\Sigma}}^2 + \frac{1}{2}\left\|D^{-1/2}v\right\|_{\Sigma}^2
\end{aligned}$$

where the second inequality uses that $v \in \mathcal{C}_n(32k)$ (along with $\|\cdot\|_\infty \leq \|\cdot\|_2$), and the third inequality uses the assumption on m together with Equation (10). Simplifying and once again using Equation (10), we get

$$\|v\|_2^2 \leq \left\|D^{-1/2}v\right\|_{\Sigma}^2 \leq 32\left\|D^{-1/2}v\right\|_{\hat{\Sigma}}^2$$

as needed. ■

Note that (16) is exactly the condition needed to apply Lemma 14. We now have all the pieces needed to bound the prediction error of `RescaledLasso()` under the rescalability assumption. We restate the notation for completeness.

Theorem 23 (Restatement of Theorem 4) *Let $n, m, k, h \in \mathbb{N}$ and $\alpha, \sigma, \delta, \lambda > 0$. Let $\Sigma \in \mathbb{R}^{n\times n}$ be a positive definite matrix that is (α, h)-rescalable at sparsity k (Definition 3), and let $w^\star \in \mathbb{R}^n$ be a k-sparse vector. Let $(X^{(j)}, y^{(j)})_{j=1}^m$ be m independent samples from $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$, and define the estimator $\hat{w} :=$ `RescaledLasso`$((X^{(j)}, y^{(j)})_{j=1}^m, k, \lambda)$. Suppose that*

$$\lambda \geq \max\left(64\sigma\sqrt{\frac{\alpha\log(48n/\delta)}{m}}, \frac{32\sigma\sqrt{2C_{17}h\log(24/\delta)}}{k\sqrt{m}}\right)$$

and $m \geq C_{23}(h + \alpha k^2)\log(288n/\delta)$ for a sufficiently-large constant C_{23}. Then

$$\Pr[\|\hat{w} - w^\star\|_\Sigma \leq 318k\lambda] \geq 1 - \delta.$$

Moreover, the time complexity of `RescaledLasso()` *is $\mathrm{poly}(n, \log\max_i \frac{\Sigma_{ii}}{D_{ii}})$, where $D \succ 0$ is the diagonal matrix guaranteed by Definition 3.*

Proof As before, let $\mathbb{X}$ be the $m \times n$ random matrix with rows $X^{(1)}, \ldots, X^{(m)}$, and let $\hat{\Sigma} := \frac{1}{m}\sum_{j=1}^m X^{(j)}(X^{(j)})^\top$ be the empirical covariance matrix. By Definition 3, there is a diagonal

matrix $D \succ 0$ and a rank-h matrix $L \succeq 0$ such that $I_n \preceq_{\mathcal{C}_n(32k)} D^{-1/2}\Sigma D^{-1/2} \preceq \alpha I_n + L$. By Lemma 19 and choice of m, the event $\mathcal{E}_{15}(\delta)$ occurs with probability at least $1-\delta$. From now on, let us condition on this event. By Lemma 22 and choice of m, we have that

$$I_n \preceq_{\mathcal{C}_n(32k)} 32 D^{-1/2}\hat{\Sigma} D^{-1/2}.$$

Thus, we can apply Lemma 14 with the sample matrix $\mathbb{X}$, sparsity k, and rescaling matrix $\frac{1}{32}D$. We get that the time complexity of `SmartScaling(`$\mathbb{X}$`,`k`)` is $\mathrm{poly}(n, \log \max_i \frac{\Sigma_{ii}}{D_{ii}})$ (using that $\hat{\Sigma}_{ii} \le \Sigma_{ii}$ for all $i \in [n]$ under event $\mathcal{E}_{15}(\delta)$); this implies the claimed time complexity bound for `RescaledLasso()`. Moreover, the output $\hat{D}$ satisfies $D \preceq 64\hat{D}$ and also $\frac{1}{m}\|\mathbb{X}v\|_2^2 > \left\|\hat{D}^{1/2}v\right\|_\infty^2$ for all $v \in \mathbb{R}^n$ with $\hat{D}^{1/2}v \in \mathcal{C}_n(16k)$.

We now apply Lemma 21 with this choice of $\hat{D}$, using the bound $D \preceq 64\hat{D}$ as well as the assumptions on m and λ. Note that property (b) of Lemma 21 contradicts our previously-derived guarantee on $\hat{D}$. Thus, property (a) must hold, i.e. $\left\|\hat{D}^{1/2}e\right\|_1 \le \frac{20k}{\sqrt{m}}\|\mathbb{X}e\|_2$, where $e := \hat{w} - w^\star$ is the error of the output of the algorithm. We now apply Equation (11b) to the vector e, which yields

$$\begin{aligned}
\|e\|_\Sigma^2 &\le \frac{16}{m}\|\mathbb{X}e\|_2^2 + \frac{16\alpha\log(48n/\delta)}{m}\left\|D^{1/2}e\right\|_1^2 \\
&\le \frac{16}{m}\|\mathbb{X}e\|_2^2 + \frac{2^{10}\alpha\log(48n/\delta)}{m}\left\|\hat{D}^{1/2}e\right\|_1^2 \\
&\le \frac{16}{m}\|\mathbb{X}e\|_2^2 + \frac{2^{10}\alpha\log(48n/\delta)}{m}\cdot\frac{400k^2}{m}\|\mathbb{X}e\|_2^2 \\
&\le \frac{32}{m}\|\mathbb{X}e\|_2^2 \qquad (17)
\end{aligned}$$

where the second inequality uses the bound $D \preceq 64\hat{D}$ (together with the fact that D, $\hat{D}$ are diagonal), and the final inequality is by choice of m. But by Lemma 20 (this time, dropping the first term of the left-hand side), we can conversely bound $\frac{2}{m}\|\mathbb{X}e\|_2^2$ as

$$\begin{aligned}
\frac{2}{m}\|\mathbb{X}e\|_2^2 &\le 4\lambda\left\|\hat{D}^{1/2}e_S\right\|_1 + 4\sigma\|e\|_\Sigma\sqrt{\frac{2C_{17}h\log(24/\delta)}{m}} \\
&\le \frac{80k\lambda}{\sqrt{m}}\|\mathbb{X}e\|_2 + \frac{32\sigma\sqrt{C_{17}h\log(24/\delta)}}{m}\|\mathbb{X}e\|_2 \\
&\le \frac{112k\lambda}{\sqrt{m}}\|\mathbb{X}e\|_2,
\end{aligned}$$

where the second inequality again uses the bound $\left\|\hat{D}^{1/2}e\right\|_1 \le \frac{20k}{\sqrt{m}}\|\mathbb{X}e\|_2$, as well as Equation (17), and the third inequality uses the assumption that $\lambda \ge \sigma\sqrt{\frac{C_{17}h\log(24/\delta)}{k^2m}}$. Hence, dividing out by $\frac{2}{\sqrt{m}}\|\mathbb{X}e\|_2$ and combining with Equation (17), we see that

$$\|e\|_\Sigma \le \frac{4\sqrt{2}}{\sqrt{m}}\|\mathbb{X}e\|_2 \le 318k\lambda$$

as claimed. ■

Appendix B. Rescalability of covariance with few outlier eigenvalues

In this section we prove Corollary 6, restated (in slightly greater generality) below. It asserts that `RescaledLasso()` is sample-efficient for sparse linear regression when the covariance matrix has few outlier eigenvalues (and runs in polynomial time under the mild assumption that the condition number of Σ is at most exponential in n).

Theorem 24 (Restatement of Corollary 6) *There is a constant C_{24} so that the following holds. Let $n, m, k, d_{\mathsf{low}}, d_{\mathsf{high}} \in \mathbb{N}$ with $d_{\mathsf{low}} + d_{\mathsf{high}} < n$. Let $\sigma, \delta, \lambda > 0$. Let $\Sigma \in \mathbb{R}^{n \times n}$ be a positive definite matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, and let $w^\star \in \mathbb{R}^n$ be a k-sparse vector. Let $(X^{(j)}, y^{(j)})_{j=1}^m$ be m independent samples from $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$, and define the estimator $\hat{w} := \texttt{RescaledLasso}((X^{(j)}, y^{(j)})_{j=1}^m, k, \lambda)$. Suppose that*

$$k\lambda \geq \frac{C_{24}\sigma}{\sqrt{m}} \left(k^2 \sqrt{\frac{\lambda_{d_{\mathsf{high}}+1}}{\lambda_{n-d_{\mathsf{low}}}}} + k\sqrt{d_{\mathsf{low}}} + \sqrt{d_{\mathsf{high}}} \right) \sqrt{\log(48n/\delta)}$$

and $m \geq C_{24}(k^4 \frac{\lambda_{d_{\mathsf{high}}+1}}{\lambda_{n-d_{\mathsf{low}}}} + k^2 d_{\mathsf{low}} + d_{\mathsf{high}}) \log(288n/\delta)$. Then

$$\Pr[\|\hat{w} - w^\star\|_\Sigma \leq 318k\lambda] \geq 1 - \delta.$$

Moreover, the time complexity of `RescaledLasso()` *is* $\mathrm{poly}(n, \log(\lambda_n/\lambda_1))$.

Note that d_{low} quantifies the number of outlier eigenvalues at the "low" end of the eigenspectrum, and d_{high} quantifies the number of outliers at the "high" end; moreover, `RescaledLasso()` is agnostic to these parameters, meaning that it automatically achieves the statistical accuracy guaranteed by the optimal choice of d_{low} and d_{high} above (so long as λ is chosen appropriately, which is also important for the standard Lasso). For simplicity, we stated Corollary 6 only for the special case $d_{\mathsf{low}} = d_{\mathsf{high}} = d$ (that nonetheless captures the essence of the result).

This theorem will follow immediately from our generic analysis of `RescaledLasso()` (Theorem 23) once we prove that any covariance matrix with few outlier eigenvalues is (α, h)-rescalable (Definition 3) with appropriate parameters α and h. Towards this end, the following lemma is key. Essentially, it states that if Σ is a covariance matrix with at most d "small" eigenvalues, then there is a rescaling $\tilde{D}$ that touches only $O(dk^2)$ coordinates, but makes every approximate linear dependency among the covariates $\Omega(k)$-quantitatively dense. The matrix D needed for Definition 3 will then be an appropriate scalar multiple of $\tilde{D}$, and the existence of a low-rank matrix L satisfying the upper bound in Definition 3 will use the fact that most entries of $\tilde{D}$ are equal to one.

Lemma 25 *Let $n, k \in \mathbb{N}$, and let d be an integer with $0 \leq d < n$. Let $\Sigma \in \mathbb{R}^{n \times n}$ be a positive-definite matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. Then there is a diagonal matrix $\tilde{D} \in \mathbb{R}^{n \times n}$ with $0 \prec \tilde{D} \preceq I_n$ satisfying the following properties:*

- $|\{i : \tilde{D}_{ii} \neq 1\}| \leq 128dk^2$
- *For every $v \in \mathbb{R}^n$ with $\left\|\tilde{D}^{1/2}v\right\|_2 \leq (k/2)\left\|\tilde{D}^{1/2}v\right\|_\infty$, it holds that $\left\|\tilde{D}^{1/2}v\right\|_2 \leq 4k\lambda_{n-d}^{-1/2}\|v\|_\Sigma$.*

Remark 26 (Comparison to Kelner et al. (2023)) *It is useful to compare Lemma 25 with (Kelner et al., 2023, Lemma 2.4), the main structural lemma underlying the prior algorithm for sparse linear regression in Setting 2. Their lemma may be interpreted as constructing a rescaling matrix with* binary *entries, such that all but $k^{O(k)}d$ entries are equal to one, and there are no sparse approximate linear dependencies. Lemma 25 relaxes the binary assumption, and thereby achieves a much stronger guarantee both quantitatively – in the number of entries not equal to one – and qualitatively – in that it rules out quantitatively sparse dependencies, not just algebraically sparse dependencies. The first improvement is the source of our improved sample complexity, but the second improvement is even more crucial: rescalability requires a spectral lower bound that holds for all quantitatively sparse vectors. Thus, (Kelner et al., 2023, Lemma 2.4) does not have any direct implication for rescalability.*

The proof of Lemma 25 is constructive, and looks quite similar to an "oracle" version of `SmartScaling()` where the procedure is given access to Σ, and thus can compute an eigendecomposition (but of course, this is only within the analysis). The intuition is as follows. Drawing on the proof technique of Kelner et al. (2023), one might hope that there is a good *binary* rescaling matrix $\tilde{D}$ (and that it can be constructed by iteratively zeroing out "bad" coordinates). Unfortunately, the following example shows that this is false:

Example 3 *Let $v := (1, 1/2, 1/4, \dots, 1/2^{n-1}) \in \mathbb{R}^n$, and let $\Sigma := I_n - (1-\epsilon)\frac{vv^\top}{\|v\|_2^2}$, so that Σ has a single small eigenvalue, with eigenvector v. Note that v is $O(1)$-quantitatively sparse, and $\|v\|_\Sigma \ll \|v\|_2$. Thus, $\tilde{D} := I_n$ does not satisfy the conditions of Lemma 25. Moreover, if any t diagonal entries of $\tilde{D}$ are set to zero, the vector $\tilde{D}^{1/2}v$ will still be $O(1)$-quantitatively sparse, and $\|\tilde{D}^{1/2}v\|_2 \geq 2^{-t}$ whereas $\|v\|_\Sigma = \epsilon$. So the conditions can't be satisfied unless $t \geq O(\log(1/\epsilon))$.*

In particular, Example 3 shows that if $\tilde{D}$ is required to be binary, then we cannot avoid dependence on $\log(\lambda_1/\lambda_n)$ in the bound on $|\{i : \tilde{D}_{ii} \neq 1\}|$, which will show up in the final sample complexity bound for `RescaledLasso()`. The workaround for Example 3 is to set $\tilde{D}_{ii} = 2^{i-k}$ for each $1 \leq i \leq k$: at this point, $\tilde{D}^{1/2}v$ is $\Omega(k)$-quantitatively dense, so v no longer violates the guarantee of Lemma 25. More generally, the idea is to iteratively rescale large coordinates (of each approximate dependency vector, like v above) by a constant factor, rather than zeroing them out. This discourages quantitative sparsity, as formalized in Lemma 27: after enough iterations, any vector will either be quantitatively dense or have small ℓ_2 norm (with respect to the rescaling). We bound the number of rescaled coordinates using Lemma 28.

Having discussed the high-level plan, we now proceed to the formal proof of Lemma 25.

Proof of Lemma 25. Let $\Sigma = \sum_{i=1}^n \lambda_i u_i u_i^\top$ be an eigendecomposition of Σ with $\lambda_1 \geq \cdots \geq \lambda_n$, and let $P = \sum_{i=1}^{n-d} u_i u_i^\top$. Set $T = \lceil \log_2(\lambda_1/\lambda_n) \rceil$. We will iteratively define diagonal matrices $D^{(1)}, D^{(2)}, \dots, D^{(T+1)}$ and set $\tilde{D} := D^{(T+1)}$. In particular, set $D^{(1)} := I_n$. For each $1 \leq t \leq T$, define the set

$$\mathcal{S}^{(t)} := \left\{ i \in [n] : \sup_{x \in \ker(P) \setminus \{0\}} \frac{(D^{(t)})_{ii}^{1/2} x_i}{\left\|(D^{(t)})^{1/2} x\right\|_2} > \frac{1}{8k} \right\}. \tag{18}$$

Then we let $D^{(t+1)}$ be the $n \times n$ diagonal matrix defined by

$$D_{ii}^{(t+1)} := \begin{cases} D_{ii}^{(t)}/2 & \text{if } i \in \mathcal{S}^{(t)} \\ D_{ii}^{(t)} & \text{otherwise} \end{cases}.$$

This defines $\tilde{D} := D^{(T+1)}$. From the definition it is clear that $0 \prec \tilde{D} \preceq I_n$. We start by bounding $|\{i : \tilde{D}_{ii} \neq 1\}|$, which is exactly $|\bigcup_{t=1}^{T} \mathcal{S}^{(t)}|$. By applying Lemma 28 to each set $\mathcal{S}^{(t)}$ individually, we could get a straightforward bound of $O(dk^2T)$. But we would like to avoid the factor of T, which we can do as follows. Define $V := \{\tilde{D}^{1/2}x : x \in \ker(P)\}$ and define

$$\mathcal{S} := \left\{ i \in [n] : \sup_{y \in V \setminus \{0\}} \frac{y_i}{\|y\|_2} > \frac{1}{8\sqrt{2}k} \right\}.$$

We claim that $\bigcup_{t=1}^{T} \mathcal{S}^{(t)} \subseteq \mathcal{S}$. Indeed, for any $i \in \bigcup_{t=1}^{T} \mathcal{S}^{(t)}$, let $f(i) := \arg\max\{t \in [T] : i \in \mathcal{S}^{(t)}\}$. Then $\tilde{D}_{ii} = D_{ii}^{(f(i))}/2$, and $\tilde{D}_{jj} \leq D_{jj}^{(f(i))}$ for all $j \in [n]$. Also, since $i \in \mathcal{S}^{(f(i))}$, there is some $x \in \ker(P) \setminus \{0\}$ such that $(D^{(f(i))})_{ii}^{1/2} x_i > \left\|(D^{(f(i))})^{1/2}x\right\|_2 / (8k)$. It follows that

$$\tilde{D}_{ii}^{1/2} x_i = \frac{1}{\sqrt{2}} (D^{f(i)})_{ii}^{1/2} x_i > \frac{\left\|(D^{(f(i))})^{1/2}x\right\|_2}{8\sqrt{2}k} \geq \frac{\left\|\tilde{D}^{1/2}x\right\|_2}{8\sqrt{2}k}$$

and thus $y = \tilde{D}^{1/2}x \in V \setminus \{0\}$ satisfies $y_i / \|y\|_2 > 1/(8\sqrt{2}k)$, so $i \in \mathcal{S}$. As claimed, we get $\bigcup_{t=1}^{T} \mathcal{S}^{(t)} \subseteq \mathcal{S}$. But now since V is a d-dimensional subspace, Lemma 28 gives that $|\mathcal{S}| \leq 128dk^2$. This proves the first part of the lemma.

Next, fix any $v \in \mathbb{R}^n$ with $\left\|\tilde{D}^{1/2}v\right\|_2 > 4k\lambda_{n-d}^{-1/2} \|v\|_\Sigma$. Since $\lambda_{n-d}P \preceq \Sigma$, it follows that $\left\|\tilde{D}^{1/2}v\right\|_2 > 4k \|v\|_P$. For any $t \in [T]$, we have $D^{(t)} \succeq \tilde{D}$ and thus

$$\left\|(D^{(t)})^{1/2}v\right\|_2 > 4k \|v\|_P . \tag{19}$$

We claim that the vectors $(D^{(1)})^{1/2}v, \ldots, (D^{(T+1)})^{1/2}v$ evolve according to the procedure described in Lemma 27. Indeed, fix $t \in [T]$ and suppose that $\left\|(D^{(t)})^{1/2}v\right\|_2 \leq k \left\|(D^{(t)})^{1/2}v\right\|_\infty$. Pick any $i \in [n]$ such that $|(D^{(t)})_{ii}^{1/2} v_i| \geq \left\|(D^{(t)})^{1/2}v\right\|_\infty / 2$. We need to show that $i \in \mathcal{S}^{(t)}$.

Define $v = a + b$ where $b \in \ker(P)$ and $a \in \operatorname{span}(P)$, so that $\|a\|_2 = \|v\|_P$. Then

$$\begin{aligned}
|(D^{(t)})_{ii}^{1/2} b_i| &\geq |(D^{(t)})_{ii}^{1/2} v_i| - |(D^{(t)})_{ii}^{1/2} a_i| \\
&\geq \frac{1}{2} \left\|(D^{(t)})^{1/2}v\right\|_\infty - \left\|(D^{(t)})^{1/2}a\right\|_\infty \\
&\geq \frac{1}{2k} \left\|(D^{(t)})^{1/2}v\right\|_2 - \left\|(D^{(t)})^{1/2}a\right\|_2 \\
&\geq \frac{1}{2k} \left\|(D^{(t)})^{1/2}v\right\|_2 - \|a\|_2 \\
&\geq \frac{1}{4k} \left\|(D^{(t)})^{1/2}v\right\|_2
\end{aligned}$$

where the first inequality is by reverse triangle inequality, the second inequality uses the choice of i, the fourth inequality uses that $D^{(t)} \preceq D^{(1)} = I_n$, and the fifth inequality uses Equation (19) together with the fact that $\|a\|_2 = \|v\|_P$. But now

$$\left\|(D^{(t)})^{1/2}b\right\|_2 \leq \left\|(D^{(t)})^{1/2}v\right\|_2 + \left\|(D^{(t)})^{1/2}a\right\|_2 \leq \left\|(D^{(t)})^{1/2}v\right\|_2 + \|a\|_2 \leq 2\left\|(D^{(t)})^{1/2}v\right\|_2$$

again using the triangle inequality, the fact that $D^{(t)} \preceq I_n$, and Equation (19). Combining the above two displays, we get that $x = (D^{(t)})^{1/2} b$ satisfies $|x_i| \geq \|x\|_2 / (8k)$. Since $b \in \ker(P)$ (and $b \neq 0$) we conclude by definition of $\mathcal{S}^{(t)}$ that $i \in \mathcal{S}^{(t)}$. Thus, the vectors $(D^{(1)})^{1/2} v, \ldots, (D^{(T+1)})^{1/2} v$ indeed evolve according to the procedure described in Lemma 27. Since $D^{(T+1)} = \tilde{D}$ and $D^{(1)} = I_n$, it follows from that lemma that at least one of the following occurs:

- $\left\|\tilde{D}^{1/2} v\right\|_2 > (k/2) \left\|\tilde{D}^{1/2} v\right\|_\infty$, or
- $\left\|\tilde{D}^{1/2} v\right\|_2 \leq 2^{-T} k \left\|(D^{(1)})^{1/2} v\right\|_\infty \leq 2^{-T} k \|v\|_2$.

In the former case, we are done. In the latter case, by choice of T and the fact that $\Sigma \succeq \lambda_n I_n$, we have $\left\|\tilde{D}^{1/2} v\right\|_2 \leq \sqrt{\frac{\lambda_n}{\lambda_1}} k \|v\|_2 \leq \frac{k}{\sqrt{\lambda_1}} \|v\|_\Sigma \leq \frac{k}{\sqrt{\lambda_{n-d}}} \|v\|_\Sigma$ which contradicts the assumption we made about v. We conclude that for any $v \in \mathbb{R}^n$, either $\left\|\tilde{D}^{1/2} v\right\|_2 \leq 4k \lambda_{n-d}^{-1/2} \|v\|_\Sigma$ or $\left\|\tilde{D}^{1/2} v\right\|_2 > (k/2) \left\|\tilde{D}^{1/2} v\right\|_\infty$. ■

Lemma 27 *Let $v \in \mathbb{R}^n$ be an arbitrary vector with ℓ_∞ norm at most 1. Consider the following procedure that we repeat T times:*

(a) If $\|v\|_2 \leq k \|v\|_\infty$, then let $S \subseteq [n]$ be a set of indices containing $\{i \in [n] : |v_i| \geq \|v\|_\infty / 2\}$. Halve v_i for all $i \in S$.

(b) Otherwise, let $S \subseteq [n]$ be arbitrary. Halve v_i for all $i \in S$.

After this procedure, the final value of v satisfies either $\|v\|_2 > (k/2) \|v\|_\infty$ or $\|v\|_2 \leq 2^{-T} k$.

Proof At any step where case (a) occurs, note that $\|v\|_2$ decreases by a factor of at most 2, whereas $\|v\|_\infty$ decreases by a factor of at least 2, so the ratio $\|v\|_2 / \|v\|_\infty$ cannot decrease. Moreover, when case (b) occurs, $\|v\|_2$ decreases by a factor of at most 2, and $\|v\|_\infty$ is non-increasing, so the ratio $\|v\|_2 / \|v\|_\infty$ can decrease by at most a factor of 2. Thus, if at any step we have $\|v\|_2 > k \|v\|_\infty$, inductively we have at all subsequent steps (and in particular after the final step) that $\|v\|_2 > (k/2) \|v\|_\infty$.

It remains to consider the case that $\|v\|_2 \leq k \|v\|_\infty$ at all steps. Then $\|v\|_\infty$ decreases by a factor of at least 2 at every step. Since initially we had $\|v\|_\infty \leq 1$, at the end we must have $\|v\|_\infty \leq 2^{-T}$ and thus $\|v\|_2 \leq 2^{-T} k$. ■

Above, we used the following simple bound on the number of basis vectors that can be correlated with a low-dimensional subspace.

Lemma 28 (Kelner et al. (2023)) *Let $V \subseteq \mathbb{R}^n$ be a subspace with $d := \dim V$. For some $\alpha > 0$ define*

$$S = \left\{ i \in [n] : \sup_{x \in V \setminus \{0\}} \frac{x_i}{\|x\|_2} \geq \alpha \right\}.$$

Then $|S| \leq d/\alpha^2$.

We now use Lemma 25 to prove that any covariance matrix Σ with few outlier eigenvalues is rescalable, in the following quantitative sense.

Lemma 29 *There is a constant C_{29} with the following property. Let $n, k, d_{\mathsf{low}}, d_{\mathsf{high}} \in \mathbb{N}$ and let $\Sigma \in \mathbb{R}^{n\times n}$ be a positive semi-definite matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. Then Σ is (α, h)-rescalable at sparsity k, where $\alpha := C_{29}k^2\frac{\lambda_{d_{\mathsf{high}}+1}}{\lambda_{n-d_{\mathsf{low}}}}$ and $h := C_{29}d_{\mathsf{low}}k^2 + d_{\mathsf{high}}$. Moreover, the diagonal matrix D realizing the rescaling satisfies $\max_i \frac{\Sigma_{ii}}{D_{ii}} \leq C_{29}k^2\frac{\lambda_n^2}{\lambda_1^2}$.*

As previously discussed, the diagonal matrix D witnessing the rescalability is an appropriate scalar multiple of the matrix $\tilde{D}$ constructed in Lemma 25. The spectral lower bound needed for rescalability follows from the second guarantee of Lemma 25. Proving the spectral upper bound requires choosing the low-rank matrix L to handle both the coordinates i for which $\tilde{D}_{ii} \neq 1$ (of which there are not many, by the first guarantee of Lemma 25), as well as the "high" end of the eigenspectrum of Σ.

Proof Let $\Sigma = \sum_{i=1}^n \lambda_i u_i u_i^\top$ be an eigendecomposition of Σ. Let $\tilde{D}$ be the matrix guaranteed by Lemma 25 with parameters d_{low} and $64k$, and define $D := \frac{\lambda_{n-d_{\mathsf{low}}}}{2^{16}k^2}\tilde{D}$. By construction it is clear that $\min_i \tilde{D}_{ii} \geq \lambda_n/(2\lambda_1)$, so $\max_i \frac{\Sigma_{ii}}{D_{ii}} \leq \frac{2^{17}k^2\lambda_1^2}{\lambda_n^2}$. Also define

$$L := C_{29}k^2\frac{\lambda_{d_{\mathsf{high}}+1}}{\lambda_{n-d_{\mathsf{low}}}} \cdot (\tilde{D}^{-1} - I_n) + \sum_{i=1}^{d_{\mathsf{high}}} \lambda_i u_i u_i^\top.$$

By the first guarantee of Lemma 25 we have $|\{i : \tilde{D}_{ii} \neq 1\}| \leq O(d_{\mathsf{low}}k^2)$, and thus $\mathrm{rank}(L) \leq O(d_{\mathsf{low}}k^2) + d_{\mathsf{high}}$. It remains to check that

$$I_n \preceq_{\mathcal{C}_n(32k)} D^{-1/2}\Sigma D^{-1/2} \preceq C_{29}k^2\frac{\lambda_{n-d_{\mathsf{high}}}}{\lambda_{d_{\mathsf{low}}+1}}I_n + L$$

when C_{29} is a sufficiently large constant. Pick any $w \in \mathcal{C}_n(32k)$ and set $v := \tilde{D}^{-1/2}w$. We know that $\left\|\tilde{D}^{1/2}v\right\|_2 \leq \left\|\tilde{D}^{1/2}v\right\|_1 \leq 32k\left\|\tilde{D}^{1/2}v\right\|_\infty$. So by the second guarantee of Lemma 25, we have $\left\|\tilde{D}^{1/2}v\right\|_2 \leq 256k\lambda_{n-d_{\mathsf{low}}}^{-1/2}\|v\|_\Sigma$ (recall that we are taking the parameter k in Lemma 25 to be $64k$). Hence,

$$\|w\|_2^2 = \left\|\tilde{D}^{1/2}v\right\|_2^2 \leq 2^{16}k^2\lambda_{n-d_{\mathsf{low}}}^{-1} \cdot w^\top\tilde{D}^{-1/2}\Sigma\tilde{D}^{-1/2}w = w^\top D^{-1/2}\Sigma D^{-1/2}w.$$

This proves the first inequality. To prove the second inequality, note that

$$\begin{aligned} C_{29}k^2\frac{\lambda_{d_{\mathsf{high}}+1}}{\lambda_{n-d_{\mathsf{low}}}}D + D^{1/2}LD^{1/2} &= C_{29}k^2\frac{\lambda_{d_{\mathsf{high}}+1}}{\lambda_{n-d_{\mathsf{low}}}}D^{1/2}\tilde{D}^{-1}D^{1/2} + \sum_{i=1}^{d_{\mathsf{high}}}\lambda_i u_i u_i^\top \\ &\succeq \lambda_{d_{\mathsf{high}}+1}I_n + \sum_{i=1}^{d_{\mathsf{high}}}\lambda_i u_i u_i^\top \\ &\succeq \Sigma \end{aligned}$$

so long as $C_{29} \geq 2^{16}$. Applying $D^{-1/2}$ on the left and right yields the second claimed inequality. ■

Proof of Theorem 24. Immediate from Theorem 23 and Lemma 29. ■

Appendix C. Hardness evidence via sparse PCA with a near-critical negative spike

In this section we prove Theorem 7, which asserts (under Conjecture 33, defined below) that no polynomial-time algorithm for sparse linear regression can achieve prediction error $\sigma^2/10$ with significantly less than $O(k^2 \log n)$ samples, even when Σ is $(1,k)$-rescalable. Since `RescaledLasso()` has sample complexity $O(k^2 \log n)$ in Setting 1 with up to $O(k^2)$ latent variables, it follows under the same conjecture that `RescaledLasso()` is essentially optimal in that setting.

Section outline. In Section C.1, we introduce Conjecture 33 and other necessary background. In Section C.2, we formally prove Theorem 11, which states that negative-spike k-sparse PCA can be reduced to k-sparse linear regression with $(1,k)$-rescalable covariance. In Section C.3, we show that low-degree polynomials cannot solve near-critical negative spike sparse PCA with $o(k^2)$ samples, adapting an argument from Bandeira et al. (2020). Combined with Theorem 11, this yields Theorem 7. In Section C.4, we give additional evidence of a statistical/computational tradeoff for negative spike sparse PCA, by showing that a natural semidefinite program also requires $\tilde{\Omega}(k^2)$ samples.

Remark 30 *The negative spiked sparse PCA testing problem can be thought of as a real-valued analogue of the foundational Sparse Parities with Noise (SPN) problem (Feldman et al., 2009), where the task is to distinguish between m independent samples from the null distribution $\mathrm{Unif}\{\pm 1\}^n$, and m samples from a random planted distribution on $\{\pm 1\}^n$ defined as follows. First, a set $S \subseteq [n]$ of size k is sampled uniformly at random. Then, the constraint $x_S = 0$ is noisily "planted" in each of the m samples. Formally, conditioned on S, each of the m samples is i.i.d. with distribution*

$$X \sim (1-\epsilon)\,\mathrm{Unif}\left\{x : \sum_{i\in S} x_i = 0 \pmod 2\right\} + \epsilon\,\mathrm{Unif}\left\{x : \sum_{i\in S} x_i = 1 \pmod 2\right\}.$$

This problem is solvable via Gaussian elimination when $\epsilon = 0$ but is conjectured to require $n^{\Omega(k)}$ time when $\epsilon > 0$ is fixed. See e.g. Valiant (2012) and references within for a history of this well-known problem. When ϵ is close to zero, we can informally view the planted measure as the result of conditioning on $\sum_{i\in S} X_i \approx 0$ (where the approximation is in the sense of L^2).

The negative spike sparse PCA problem can similarly be viewed as testing between $N(0,I)$ and a planted distribution where for a random set S (corresponding to the support of the spike vector w in Definition 10), we condition a standard Gaussian vector X on the event that $\sum_{i\in S} X_i$ has smaller variance than in the null measure. In the near-critical case, we are essentially conditioning on $\sum_{i\in S} X_i \approx 0$.

In both cases, the information-theoretic limit is at $\tilde{O}(k)$ samples. Of course, the computational limits are somewhat different, since SPN is believed to be computationally intractable even with $\mathrm{poly}(n)$ samples.

C.1. Preliminaries

Recall the definition of the sparse spike prior $\mathcal{W}_{n,k}$ (Definition 9), the sparse spiked Wishart distribution $\mathbb{P}_{n,k,\beta,m}$, and the null distribution $\mathbb{Q}_{n,m}$ (Definition 10) from Section 2.2. We formally define the testing problem associated with these distributions, and the induced likelihood ratio.

Definition 31 *Fix functions $k, m : \mathbb{N} \to \mathbb{N}$ with $k(n) \leq n$, and $\beta : \mathbb{N} \to (-1, \infty)$. An algorithm $\mathcal{A}$ solves the* strong detection *problem for the k-sparse spiked Wishart model (with parameter functions k, m, and β) if it distinguishes $\mathbb{P}_{n,k(n),\beta(n),m(n)}$ from $\mathbb{Q}_{n,m(n)}$ with probability $1 - o(1)$, i.e.*

$$\left| \Pr_{Z \sim \mathbb{P}_{n,k(n),\beta(n),m(n)}} [\mathcal{A}(Z) = 1] - \Pr_{Z \sim \mathbb{Q}_{n,m(n)}} [\mathcal{A}(Z) = 1] \right| = 1 - o(1).$$

Definition 32 *Let $n, k, m \in \mathbb{N}$ with $k \leq n$ and $\beta \in (-1, \infty)$. We define the* likelihood ratio *as $L_{n,k,\beta,m} := \frac{d\mathbb{P}_{n,k,\beta,m}}{d\mathbb{Q}_{n,m}}$.*

For a planted distribution $\mathbb{P}$ and a null distribution $\mathbb{Q}$, the (degree-D) *low-degree likelihood ratio* (LDLR) $\left\| L^{\leq D} \right\|_{L^2(\mathbb{Q})}$ is the norm under $L^2(\mathbb{Q})$ of the likelihood ratio $L = d\mathbb{P}/d\mathbb{Q}$ after orthogonal projection onto the space of multivariate polynomials of degree at most D. For a family of planted distributions $(\mathbb{P}_n)_{n \in \mathbb{N}}$ and null distributions $(\mathbb{Q}_n)_{n \in \mathbb{N}}$, if the low-degree likelihood ratio between $\mathbb{P}_n$ and $\mathbb{Q}_n$ can be bounded above by a constant as $n \to \infty$, then it can be seen that no degree-D polynomial can distinguish $\mathbb{P}_n$ from $\mathbb{Q}_n$ with error $o(1)$ as $n \to \infty$ (see e.g. Proposition 1.15 in Kunisky et al. (2019) and references).

Moreover, it has been conjectured that for any "natural" statistical hypothesis testing problem, the best degree-$\log^{1+c} n$ polynomial (for any constant $c > 0$) is at least as good a distinguisher as the best polynomial-time algorithm. Informally, this is known as the *Low Degree Conjecture*. There are concrete, formal statements of this conjecture (see e.g. Hopkins (2018)) for broad classes of statistical problems, although these do not specifically capture the spiked Wishart model. See also Holmgren and Wein (2020); Koehler and Mossel (2022) for further discussion about the settings where low-degree polynomials are good proxies for polynomial time algorithms. Below we formalize the precise conjecture that needs to hold for our purposes (i.e. to prove non-existence of an efficient algorithm for negative-spike sparse PCA via a low-degree likelihood ratio bound).

Conjecture 33 (Hardness thresholds for spiked Wishart match Low-Degree) *Fix functions $k, m : \mathbb{N} \to \mathbb{N}$ with $k(n) \leq n$, and $\beta : \mathbb{N} \to (-1, 0)$ with $1 + \beta(n) \geq 1/\operatorname{poly}(n)$. If there exists some $D : \mathbb{N} \to \mathbb{N}$ with $D(n) = \log^{1+\Omega(1)} n$ and $\left\| L^{\leq D(n)}_{n,k(n),\beta(n),m(n)} \right\|_{L^2(\mathbb{Q}_{n,m(n)})} = O(1)$, then there is no randomized polynomial-time algorithm $\mathcal{A}$ that solves strong detection (Definition 31) for the spiked Wishart model with parameter functions k, m, and β.*

Remark 34 *In this conjecture, we do not allow β to equal -1 (or to converge to -1 more than polynomially fast). In part, this is because low-degree hardness is only expected to be a good heuristic when there is at least a small amount of noise in the underlying problem. If the underlying signal is binary valued and there is extremely little noise, algebraic methods like the LLL algorithm can sometimes be used to solve regression tasks with very few samples, see Zadik and Gamarnik (2018).*

C.2. Reduction from negative-spike sparse PCA to sparse linear regression

We start by reducing (near-critical) negative-spike sparse PCA to sparse linear regression – with a covariance matrix that satisfies $(1, k)$-rescalability. As was explained in Section 2.2, the idea is to check whether any covariate in the given sparse PCA data can be explained by the other covariates better than one would expect under the null distribution.

Theorem 35 (Restatement of Theorem 11) *Let $m_{\mathsf{SLR}}(n,k)$ be a function. Suppose that there is a polynomial-time sparse linear regression algorithm $\mathcal{A}$ with the following property. For any $n,k \in \mathbb{N}$, $\sigma > 0$, positive semi-definite $(1,k)$-rescalable matrix $\Sigma \in \mathbb{R}^{n\times n}$, k-sparse vector $w^\star \in \mathbb{R}^n$, and $m \geq m_{\mathsf{SLR}}(n,k)$, the output $\hat{w} \leftarrow \mathcal{A}((X^{(j)}, y^{(j)})_{j=1}^m)$ satisfies*

$$\Pr[\|\hat{w} - w^\star\|_\Sigma^2 \leq \sigma^2/10] \geq 1 - o(1)$$

where the probability is over the randomness of $\mathcal{A}$ and m independent samples $(X^{(j)}, y^{(j)})_{j=1}^m$ from $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$.

Then there is a polynomial-time algorithm $\mathcal{A}'$ with the following property. For any $n, m, k \in \mathbb{N}$ and $\beta \in (-1, -1 + 1/(2k)]$, if $m \geq m_{\mathsf{SLR}}(n,k) + 1600\log(n)$, then

$$\left| \Pr_{Z\sim\mathbb{P}_{n,k,\beta,m}}[\mathcal{A}'(Z) = 1] - \Pr_{Z\sim\mathbb{Q}_{n,m}}[\mathcal{A}'(Z) = 1] \right| = 1 - o(1). \tag{20}$$

To be precise, the asymptotics here (as elsewhere in this section) are in terms of n. That is, each term $o(1)$ represents a function of n (that of course depends on the algorithm $\mathcal{A}$, but not the other parameters) that goes to 0 as $n \to \infty$.

Proof The algorithm $\mathcal{A}'$ on input $Z = (Z^{(j)})_{j=1}^m$ has the following behavior. Let $m' = m - 1600\log(n)$. For each $i \in [n]$, compute

$$\hat{w}^{(i)} := \mathcal{A}((Z_{-i}^{(j)}, Z_i^{(j)})_{j=1}^{m'})$$

and

$$\eta^{(i)} := \frac{1}{1600\log n} \sum_{j=m'+1}^{m} \left(Z_i^{(j)} - \langle Z_{-i}^{(j)}, \hat{w}^{(i)} \rangle \right)^2$$

The output of $\mathcal{A}$ is then

$$\mathbb{1}[\exists i \in [n] : \eta^{(i)} < 9/10].$$

Analysis. It's clear that the time complexity of $\mathcal{A}'$ is dominated by the time complexity of $\mathcal{A}$ (multiplied by n), which is by assumption polynomial in n. It remains to check Equation (20).

First, suppose that $Z \sim \mathbb{P}_{n,k,\beta,m}$. Recall that Z is sampled by first drawing a spike vector from $\mathcal{W}_{n,k}$. Let us condition on this vector being some $w \in \mathbb{R}^n$, which by definition of $\mathcal{W}_{n,k}$ (Definition 9) is k-sparse. Fix any $i \in \arg\max_{j\in[n]} |w_j|$. We will show that $\eta^{(i)} < 9/10$ with high probability. By definition, the random variables $Z^{(1)}, \ldots, Z^{(m)}$ are independent and identically distributed (after conditioning on w). Fix any $j \in [m]$. Then $Z^{(j)} \sim N(0, I_n + \beta w w^\top)$. Thus, the marginal distribution of $Z_{-i}^{(j)}$ is $N(0, I_{n-1} + \beta w_{-i} w_{-i}^\top)$, and for any $x \in \mathbb{R}^n$, $Z_i^{(j)} | Z_{-i}^{(j)} = x$ has distribution

$$N\left(\left\langle x, \frac{\beta w_i}{1 + \beta(1 - w_i^2)} w_{-i} \right\rangle, \frac{1+\beta}{1 + \beta(1 - w_i^2)} \right).$$

Define $\Sigma := I_{n-1} + \beta w_{-i} w_{-i}^\top$ and $\theta := \frac{\beta w_i}{1+\beta(1-w_i^2)} w_{-i}$ and $\sigma^2 := \frac{1+\beta}{1+\beta(1-w_i^2)}$. Then the tuple $(Z_{-i}^{(j)}, Z_i^{(j)})$ is distributed according to $\mathsf{SLR}_{\Sigma,\sigma}(\theta)$. Since w_{-i} is $(k-1)$-sparse, we get that θ is $(k-1)$-sparse.

Let $S := \mathrm{supp}(w_{-i}) \subseteq [n-1]$ and let $D \in \mathbb{R}^{n-1\times n-1}$ be the diagonal matrix defined by $D_{aa} := \mathbb{1}[a \notin \mathrm{supp}(w)]$ for $a \in [n-1]$. Then for any $v \in \mathbb{R}^{n-1}$,

$$\|v\|_\Sigma^2 = \|v\|_2^2 + \beta\langle v, w_{-i}\rangle^2 \geq \|v\|_2^2 - \left\|v_{[n-1]\setminus S}\right\|_2^2 = \|v\|_D^2$$

where the inequality is by Cauchy-Schwarz, the assumption $\beta > -1$, and the bound $\|w_{-i}\|_2 \le \|w\|_2 = 1$. Thus, $D \preceq \Sigma \preceq I_n$. Since $I_n - D$ has rank at most k, it follows that Σ is $(1,k)$-rescalable.

We can now apply the theorem hypothesis. Since $m' \ge m_{\mathsf{SLR}}(n,k)$, it holds with probability $1-o(1)$ that $\hat{w}^{(i)} \leftarrow \mathcal{A}((Z_{-i}^{(j)}, Z_i^{(j)})_{i=1}^{m'})$ satisfies

$$\left\|\hat{w}^{(i)} - \theta\right\|_\Sigma^2 \le \sigma^2/10 \le 1/20. \tag{21}$$

where the last inequality uses that

$$\sigma^2 = \frac{1+\beta}{1+\beta(1-w_i^2)} \le \frac{1+\beta}{1+\beta-\beta/k} \le \frac{1+\beta}{1+\beta+1/(2k)} \le 1/2$$

(since $|w_i| \ge 1/\sqrt{k}$ and $1+\beta \le 1/(2k)$). Condition on the event that Equation (21) holds. Then for any $m' < j \le m$, since $Z^{(j)}$ is independent of $\hat{w}^{(i)}$, we have

$$\mathbb{E}\left(Z_i^{(j)} - \langle Z_{-i}^{(j)}, \hat{w}^{(i)}\rangle\right)^2 = \sigma^2 + \left\|\hat{w}^{(i)} - \theta\right\|_\Sigma^2 \le 3/5.$$

By concentration of χ^2 random variables, it follows that $\eta^{(i)} < 9/10$ (and hence $\mathcal{A}$ outputs 1) with probability $1-o(1)$ over $Z \sim \mathbb{P}_{n,k,\beta,m}$.

On the other hand, suppose that $Z \sim \mathbb{Q}_{n,m}$. Again, $(Z^{(j)})_{j=1}^m$ are independent. Fix $i \in [n]$ and condition on the first m' samples, which fixes $\hat{w}^{(i)}$. For any $m' < j \le m$, we know that $Z_i^{(j)}$ is independent of $\langle Z_{-i}^{(j)}, \hat{w}^{(i)}\rangle$, so

$$\mathbb{E}\left(Z_i^{(j)} - \langle Z_{-i}^{(j)}, \hat{w}^{(i)}\rangle\right)^2 \ge \mathbb{E}(Z_i^{(j)})^2 = 1.$$

Concentration of χ^2 random variables gives that $\eta^{(i)} \ge 9/10$ with probability at least $1-2n^{-2}$. A union bound over $i \in [n]$ implies that $\mathcal{A}'$ outputs 0 with probability $1-o(1)$. This completes the proof of Equation (20). ■

C.3. Statistical efficiency of low-degree polynomials

We next show that in the low-sample regime $m = O(k^2/D)$, the degree-D likelihood ratio for the spiked Wishart model with parameter functions k, β, and m (see Section C.1) is indeed bounded. Together with Theorem 35, this proves a tight statistical/computational tradeoff for sparse linear regression with $(1,k)$-rescalable covariance, conditional on Conjecture 33.

Our starting point for bounding the low-degree likelihood ratio is the following instantiation of a general calculation due to Bandeira et al. (2020):

Lemma 36 *Let $n, k, m, D \in \mathbb{N}$ with $k \le n$ and $\beta \in (-1, \infty)$. Then*

$$\left\|L_{n,k,\beta,m}^{\le D}\right\|_{L^2(\mathbb{Q}_{n,m})}^2 = \mathop{\mathbb{E}}_{w_1,w_2\sim\mathcal{W}_{n,k}} \sum_{d=0}^{\lfloor D/2\rfloor} \left(\frac{\beta^2\langle w_1,w_2\rangle}{4}\right)^d \sum_{\substack{d_1,\dots,d_m \\ \sum d_i = d}} \prod_{i=1}^m \binom{2d_i}{d_i}.$$

Proof We apply Lemma 5.9 from Bandeira et al. (2020). We only need to check that $\mathcal{W}_{n,k}$ is a β-good normalized spike prior (see Definitions 2.9 and 2.11 in Bandeira et al. (2020)), but this is immediate from the fact that $\beta > -1$ and $\Pr_{w\sim\mathcal{W}_{n,k}}[\|w\|_2 \le 1] = 1$. ■

We now roughly follow the proof of (Ding et al., 2023, Theorem 2.14(b)), which gives a low-degree bound in the closely related setting where $\beta \ge 0$ and the sparse spike prior has independent entries.

Remark 37 (Inapplicability of previous bounds) *Note that the above expression is an even function of β, so for any given well-defined spike prior, the low-degree computation in the negative-spike case is identical to that in the positive-spike case. Unfortunately, the spike prior with independent entries is only well-defined in the positive-spike case, since it's possible for the spike vector to have norm larger than one. If this degeneracy occurred only with vanishing probability, then one could hope to perform a truncation argument (Bandeira et al., 2020; Ding et al., 2021), but in the near-critical regime that we care about (i.e. $\beta \in (-1, -1+1/(2k))$), a spike prior with i.i.d. entries and expected sparsity k will have norm exceeding one with constant probability.*

Thus, the fixed-size sparse spike prior $\mathcal{W}_{n,k}$ (Definition 9) seems crucial to the proof. We are not aware of a previous low-degree analysis with this prior, so we have to do it ourselves.

We start by bounding the moments of $\langle w_1, w_2\rangle$ for independent $w_1, w_2 \sim \mathcal{W}_{n,k}$.

Lemma 38 *Let $n, k, d \in \mathbb{N}$ with $k \le \sqrt{n/(4e)}$. Then*

$$A_{n,k,d} := k^{2d}\,\mathbb{E}\langle w_1, w_2\rangle^{2d} \le 2\cdot(2d)^{2d}$$

where the expectation is over independent draws $w_1, w_2 \sim \mathcal{W}_{n,k}$.

Remark 39 *Note that the trivial bound (from Cauchy-Schwarz) is $A_{n,k,d} \le k^{2d}$. However, for $d \ll k$ this is very loose. We improve it by using the fact that the supports of w_1, w_2 are unlikely to have large overlap.*

Proof Define the (random) set

$$S := \mathrm{supp}(w_1) \cap \mathrm{supp}(w_2) \subseteq [n].$$

Let $\mathrm{Rad}(1/2)$ denote the Rademacher distribution $\mathrm{Unif}(\{-1, 1\})$. Observe that after conditioning on any realization of S with $|S| = \ell$, the random variable $\langle w_1, w_2\rangle$ has the distribution of $\frac{1}{k}\sum_{i=1}^{\ell} a_i$ where $a_1, \dots, a_\ell \sim \mathrm{Rad}(1/2)$ are independent. Thus,

$$\begin{aligned}
A_{n,k,d} &= \sum_{\ell=0}^{k} \Pr[|S| = \ell] \cdot \mathop{\mathbb{E}}_{a_1,\dots,a_\ell\sim\mathrm{Rad}(1/2)} \left(\sum_{i=1}^{\ell} a_i\right)^{2d} \\
&\le \sum_{\ell=0}^{k} \Pr[|S| = \ell] \cdot \ell^{2d} \\
&\le (2d)^{2d} + \sum_{\ell=2d+1}^{k} \Pr[|S| = \ell] \cdot \ell^{2d}.
\end{aligned}$$

Define $g(\ell) := \Pr[|S| = \ell] \cdot \ell^{2d}$. Then for any $2d \le \ell < k$, we have

$$\begin{aligned} g(\ell+1) &= \frac{\binom{k}{\ell+1}\binom{n-k}{k-\ell-1}}{\binom{n}{k}}(\ell+1)^{2d} \\ &= \frac{k-\ell}{\ell+1} \cdot \frac{k-\ell}{n-2k+\ell+1} \cdot \left(1+\frac{1}{\ell}\right)^{2d} \cdot g(\ell) \\ &\le \frac{2k^2}{n} e^{2d/\ell} g(\ell) \\ &\le \frac{g(\ell)}{2}. \end{aligned}$$

Since $g(2d) \le (2d)^{2d}$, it follows that

$$A_{n,k,d} \le (2d)^{2d} + \sum_{\ell=2d+1}^{k} (2d)^{2d}(1/2)^{\ell-2d} \le 2 \cdot (2d)^{2d}$$

as claimed. ∎

We also use the following bound from Ding et al. (2023):

Lemma 40 (Lemma 4.7 in Ding et al. (2023)) *There are constants $c_1, c_2 > 0$ with the following property. Let $m, D : \mathbb{N} \to \mathbb{N}$ be functions with $D = o(m)$. Then for all sufficiently large $n \in \mathbb{N}$, for all $1 \le d \le D(n)$, it holds that*

$$\sum_{\substack{d_1,\dots,d_m \ge 0 \\ \sum d_i = d}} \prod_{i=1}^{m(n)} \binom{2d_i}{d_i} \le c_1 d^{3/2} e^{c_2 d^2/m(n)} \frac{(2m(n))^d}{d!}.$$

Combining the above pieces, we get the following bound on the low-degree likelihood ratio.

Theorem 41 *Let $k, m, D : \mathbb{N} \to \mathbb{N}$ be functions with $2e\sqrt{m(n)D(n)} \le k(n) \le \sqrt{n/(4e)}$ for sufficiently large $n \in \mathbb{N}$, and $D(n) = o(m(n))$. Let $\beta : \mathbb{N} \to (-1, 1)$. Then $\left\| L^{\le D(n)}_{n,k(n),\beta(n),m(n)} \right\|_{L^2(\mathbb{Q}_{n,m(n)})} \le O(1)$.*

Proof For notational simplicity, we write $k = k(n)$, $D = D(n)$, and $m = m(n)$. Applying Lemma 36 and the definition of $A_{n,k,d}$ (Lemma 38), we have that for all sufficiently large n,

$$\begin{aligned} \left\| L^{\le D}_{n,k,\beta,m} \right\|^2_{L^2(\mathbb{Q}_{n,m})} &= 1 + \sum_{d=0}^{\lfloor D/2 \rfloor} \left(\frac{\beta^2}{4k^2}\right)^d A_{n,k,d} \sum_{\substack{d_1,\dots,d_m \\ \sum d_i = d}} \prod_{i=1}^{m} \binom{2d_i}{d_i} \\ &\le 1 + 2c_1 \sum_{d=1}^{\lfloor D/2 \rfloor} \left(\frac{\beta^2}{4k^2}\right)^d (2d)^{2d} d^{3/2} e^{c_2 d^2/m} \frac{(2m)^d}{d!} \\ &\le 1 + 2c_1 \sum_{d=1}^{\lfloor D/2 \rfloor} d^{3/2} \left(\frac{8ed^2 m}{4k^2 (d!)^{1/d}}\right)^d \end{aligned}$$

$$\leq 1 + 2c_1 \sum_{d=1}^{\lfloor D/2 \rfloor} d^{3/2} \left(\frac{8e^2 dm}{4k^2}\right)^d$$
$$\leq 1 + 2c_1 \sum_{d=1}^{\lfloor D/2 \rfloor} d^{3/2} (1/2)^d$$

where the first inequality applies Lemma 38 (using that $k \leq \sqrt{n/(4e)}$) and Lemma 40 (using that $D = o(m)$); the second inequality uses the bounds $\beta^2 \leq 1$ and $e^{c_2 d^2/m} \leq e^{c_2 dD/m} \leq e^d$ (which holds for sufficiently large n, since $D = o(m)$); the third inequality uses Stirling's approximation; and the fourth inequality uses the assumption that $k^2 \geq 2e^2 mD$. It's clear that the final summation is upper bounded by an absolute constant, which completes the proof. ■

Corollary 42 (Restatement of Theorem 7) *Let $\epsilon, C > 0$ with $\epsilon \leq 2$. Suppose that there is a polynomial-time algorithm $\mathcal{A}$ satisfying the following property. For any $n, k \in \mathbb{N}$, $\sigma > 0$, positive semi-definite, $(1,k)$-rescalable matrix $\Sigma \in \mathbb{R}^{n \times n}$, k-sparse vector $w^\star \in \mathbb{R}^n$, and $m \geq Ck^{2-\epsilon} \log n$, the output $\hat{w} \leftarrow \mathcal{A}((X^{(j)}, y^{(j)})_{j=1}^m)$ satisfies*

$$\Pr[\|\hat{w} - w^\star\|_\Sigma^2 \leq \sigma^2/10] \geq 1 - o(1)$$

where the probability is over the randomness of $\mathcal{A}$ and m independent samples $(X^{(j)}, y^{(j)})_{j=1}^m$ from $\mathsf{SLR}_{\Sigma,\sigma}(w^\star)$. Then Conjecture 33 is false.

Proof Define functions $m, k, D : \mathbb{N} \to \mathbb{N}$ and $\beta : \mathbb{N} \to (-1, 0)$ by $k(n) := \log^{10/\epsilon} n$, $m(n) := Ck(n)^{2-\epsilon} \log^3 n + 1600 \log(n)$, $D(n) := (2e)^{-2} \log^2 n$, and $\beta(n) = -1 + 1/(2k(n))$. By Theorem 35, there is a polynomial-time algorithm $\mathcal{A}'$ that solves strong detection for the spiked Wishart model with parameter functions k, m, and β. But now

$$2e\sqrt{m(n)D(n)} \leq \sqrt{(C + 1600)k(n)^{2-\epsilon} \log^5 n} \leq \sqrt{\frac{(C + 1600)k(n)^2}{\log^5 n}} \leq k(n)$$

for sufficiently large $n \in \mathbb{N}$. Also, $k(n) = \log^{10/\epsilon} n \leq \sqrt{n/(4e)}$ for sufficiently large n. Finally, $m(n) = \Omega(\log^3 n)$, so $D(n) = o(m(n))$. We can therefore apply Theorem 41, which gives that $\left\| L^{\leq D(n)}_{n,k(n),\beta(n),m(n)} \right\|_{L^2(\mathbb{Q}_{n,m(n)})} \leq O(1)$. Together with the guarantee on $\mathcal{A}'$, this contradicts Conjecture 33. ■

Remark 43 *Theorem 41 also implies a computational-statistical gap for learning Gaussian Graphical Models, under Conjecture 33. In particular, for any $\beta \in (-1, -1 + 1/k)$ and any k-sparse $w \in \mathbb{R}^n$, define $\Sigma := I_n + \beta w w^\top$. Suppose that w lies in the support of $\mathcal{W}_{n,k}$, so that every nonzero entry of w lies in $\{-1/\sqrt{k}, 1/\sqrt{k}\}$. Then it can be checked that the Gaussian Graphical Model with distribution $N(0, \Sigma)$ is κ-nondegenerate for some $\kappa = \Omega(1)$, i.e. the precision matrix $\Theta := \Sigma^{-1}$ satisfies $|\Theta_{ij}| \geq \Theta(1) \cdot \sqrt{\Theta_{ii}\Theta_{jj}}$ for all $i, j \in [n]$ with $\Theta_{ij} \neq 0$.*

For any κ-nondegenerate Gaussian Graphical Model $N(0, \Sigma)$ with maximum degree k, the Markov structure (i.e. the support of Σ^{-1}) can be information-theoretically learned with $O(k \log(n)/\kappa^2)$

samples (Misra et al., 2017). But for any $\epsilon > 0$, if there were a computationally efficient algorithm for learning such models with $O(k^{2-\epsilon}\log(n)/\kappa^2)$ samples, then by taking κ to be a constant, there would be a computationally efficient algorithm for distinguishing the spiked Wishart distribution $\mathbb{P}_{n,k,\beta,m}$ from the null distribution $\mathbb{Q}_{n,m}$ for some $m = O(k^{2-\epsilon}\log(n))$, so long as $\beta \in (-1, -1 + 1/k)$: simply learn the underlying structure and check if it is the empty graph or not. But such an algorithm would contradict Theorem 41, assuming that Conjecture 33 holds.

Finally, we observe that this lower bound actually holds for the natural testing problem of distinguishing between an empty GGM and a sparse GGM with at least one nondegenerate edge. We establish upper bound with matching dependence on k later in Section D.

C.4. Statistical efficiency of semi-definite programming

We now give a second piece of evidence that negative-spike sparse PCA may exhibits a statistical/computational tradeoff, based on the failure of a natural semi-definite program for solving the detection problem. Given samples $Z^{(1)}, \dots, Z^{(m)} \in \mathbb{R}^n$, one can solve the following program in polynomial time:

$$V_k(\hat{\Sigma}) := \min_{A \in \mathbb{R}^{n\times n} : A \succeq 0} \langle \hat{\Sigma}, A \rangle \quad \text{s.t. } \operatorname{tr}(A) = 1 \text{ and } \sum_{i,j=1}^{n} |A_{ij}| \le k \tag{22}$$

where $\hat{\Sigma} = \frac{1}{m}\sum_{j=1}^{m} Z^{(j)}(Z^{(j)})^\top$ is the empirical covariance of the samples $(Z^{(j)})_{j=1}^m$. This is the same as the standard semi-definite programming relaxation suggested for positive spike sparse PCA d'Aspremont et al. (2004); Krauthgamer et al. (2013) except that the program minimizes rather than maximizing.[7]

Since the matrix $A := ww^\top$ is feasible for (22), the value of the program is at most $O(1+\beta)$ with high probability over $(X^{(j)})_{j=1}^m \sim \mathbb{P}_{n,k,\beta,m}$ (so long as $m = \Omega(k \log n)$). One would hope that under the null hypothesis $(Z^{(j)})_{j=1}^m \sim \mathbb{Q}_{n,m}$, the program value $V_k(\hat{\Sigma})$ is with high probability concentrated near one, in which case (22) would solve the strong detection problem. However, we show that this is not the case when $m \ll k^2$: in this regime, the value of the program is in fact *zero* with high probability under the null hypothesis (Theorem 45). Since the value is always nonnegative, it follows that the natural, computationally efficient test based on this program – reject the null hypothesis if $V_k(\hat{\Sigma})$ is below some threshold – fails to solve the strong detection problem if the number of samples m is significantly less than $O(k^2)$.

We prove Theorem 45 by taking A in the above program to be (an appropriate rescaling of) the orthogonal projection onto $\ker(\hat{\Sigma})$. It's clear that $\langle \hat{\Sigma}, A \rangle = 0$, so it remains to check that A is feasible for (22) with high probability, when rescaled to have trace one.

Quantitatively, if Q is an orthogonal projection matrix onto a random $(n-m)$-dimensional subspace of $\mathbb{R}^n$ (as $\operatorname{span}(\hat{\Sigma})$ indeed is), we would like to bound $\sum_{i,j=1}^n |Q_{ij}|$ by roughly $n\sqrt{m}$. Since $I_n - Q$ has the same sum of entries up to $O(n)$, we can equivalently consider an orthogonal projection matrix P onto a random m-dimensional subspace. The following lemma gives the desired bound on the sum of entries of P (essentially as an application of Johnson-Lindenstrauss concentration).

7. This difference means that we need a different construction than they use to establish the SDP lower bound — ours is based on looking at the kernel of the empirical covariance.

Lemma 44 *Let $n, m \in \mathbb{N}$ with $m \leq n$, and let $V \subseteq \mathbb{R}^n$ be a uniformly random m-dimensional subspace. Let $P \in \mathbb{R}^{n \times n}$ be the orthogonal projection matrix onto V. Then for any $\delta > 0$,*

$$\Pr\left[\sum_{i,j=1}^{n} |P_{ij}| > 11n\sqrt{m \log(n/\delta)}\right] \leq \delta.$$

Proof For any fixed unit vector $v \in \mathbb{R}^n$, it is known (Dasgupta and Gupta, 2003, Lemma 2.2) that

$$\Pr\left[\left|\|v\|_P^2 - \frac{m}{n}\right| > \frac{\epsilon m}{n}\right] \leq \exp(-k\epsilon^2/12).$$

Define $S := \{(e_i + e_j)/\sqrt{2} : i, j \in [n] \wedge i \neq j\} \cup \{e_i : i \in [n]\}$ and $\epsilon := \sqrt{12\log(n^2/\delta)/m}$. By a union bound, we get that $|\,\|v\|_P^2 - m/n| \leq \epsilon m/n$ for all $v \in S$, with probability at least $1 - \delta$. Henceforth we condition on this event.

First, $\sum_{i=1}^n |P_{ii}| = \operatorname{tr}(P) = m$. Next, fix any $i, j \in [n]$ with $i \neq j$. Then

$$P_{ij} = e_i^\top P e_j = \frac{1}{2}\left(\|e_i + e_j\|_P^2 - \|e_i\|_P^2 - \|e_j\|_P^2\right).$$

Thus,

$$|P_{ij}| \leq \frac{1}{2}\left|\|e_i + e_j\|_P^2 - \frac{2m}{n}\right| + \frac{1}{2}\left|\|e_i\|_P^2 - \frac{m}{n}\right| + \frac{1}{2}\left|\|e_j\|_P^2 - \frac{m}{n}\right| \leq \frac{2\epsilon m}{n}.$$

Substituting in the choice of ϵ, we get that

$$\sum_{i,j=1}^{n} |P_{ij}| \leq n + 2\epsilon mn \leq n + 2n\sqrt{12m\log(n^2/\delta)} \leq 11n\sqrt{m\log(n/\delta)}$$

as claimed. ■

We can now prove the claimed result that (22) has value zero with high probability under the null hypothesis.

Theorem 45 *There is a constant $c > 0$ with the following property. Let $n, m, k \in \mathbb{N}$, and let $(Z^{(j)})_{j=1}^m \sim \mathbb{Q}_{n,m}$. Let $\hat{\Sigma} := \frac{1}{m}\sum_{j=1}^m Z^{(j)}(Z^{(j)})^\top$. For any $\delta > 0$, if $m \leq \min(ck^2/\log(n/\delta), n/2)$, then*

$$\Pr[V_k(\hat{\Sigma}) > 0] \leq \delta.$$

Proof Let $P \in \mathbb{R}^{n \times n}$ be the orthogonal projection onto $\operatorname{span}(\hat{\Sigma})$ and let $A = \frac{I_n - P}{n-m}$. Note that $\operatorname{tr}(P) = \operatorname{rank}(\hat{\Sigma}) = m$ almost surely, so $\operatorname{tr}(A) = 1$ almost surely. Moreover $I_n - P = \sum_{i=1}^{n-m} w_i w_i^\top$ where $w_1, \ldots, w_{n-m}$ form an orthonormal basis for $\ker(\hat{\Sigma})$. Hence, $\langle \hat{\Sigma}, A\rangle = \frac{1}{n-m}\sum_{i=1}^{n-m} w_i^\top \hat{\Sigma} w_i = 0$. Thus, if $\sum_{i,j=1}^n |A_{ij}| \leq k$, then $V_k(\hat{\Sigma}) = 0$. It only remains to bound the probability that $\sum_{i,j=1}^n |A_{ij}| > k$. By Lemma 44 and the fact that $\operatorname{span}(\hat{\Sigma})$ is a uniformly random m-dimensional subspace of $\mathbb{R}^n$, with probability at least $1 - \delta$ we have $\sum_{i,j=1}^n |P_{ij}| \leq 11n\sqrt{m\log(n/\delta)}$. In this event,

$$\sum_{i,j=1}^{n} |A_{ij}| \leq \frac{n}{n-m} + \frac{1}{n-m}\sum_{i,j=1}^{n} |P_{ij}|$$

$$\leq \frac{12n\sqrt{m\log(n/\delta)}}{n-m}$$
$$\leq 24\sqrt{m\log(n/\delta)}$$
$$\leq k$$

where the third inequality uses the assumption that $m \leq n/2$, and the fourth inequality holds so long as $c \geq 576$. ■

Appendix D. Testing between an empty and non-empty GGM

In this section, we give a polynomial-time algorithm for testing between an empty and sparse non-empty Gaussian Graphical Model (GGM). Interestingly, this is possible to do with polynomial dependence on the sparsity and strength of the strongest edge in the graphical model, even though it is not known how to *learn* the entire graphical structure in the same setting in polynomial time (see e.g. discussion in Anandkumar et al. (2012); Misra et al. (2017); Kelner et al. (2020)). The sample complexity of this algorithm is suboptimal information-theoretically. In particular, it has a quadratic dependence on the sparsity k even though information-theoretically, it is possible to learn the entire model with sample complexity only linear in k. However, our lower bound based on negatively spiked sparse PCA (see Section C and Remark 43 for the connection with GGMs) suggests that it is optimal among polynomial-time algorithms.

The test that we use is very simple — we simply estimate all of the correlation coefficients between the variables in our model and check if any of them are significantly different from zero. The fact that such a test is possible to construct was alluded to in Remark 10 of Kelner et al. (2020) without a proof or precise statement of the sample complexity, both of which we provide here. In particular, we show here that the test obtains the conjecturally sharp quadratic dependence on k among efficient algorithms.

Lemma 46 *Let $\Sigma \in \mathbb{R}^{n\times n}$ be positive-definite and let $\Theta := \Sigma^{-1}$ be the corresponding precision matrix. Let $X \sim N(0,\Sigma)$. For any indices $i \neq j$, we have*

$$\Theta_{ii}\operatorname{Var}(X_i \mid X_{\sim i,j}) = \frac{1}{1-\Theta_{ij}^2/\Theta_{ii}\Theta_{jj}}.$$

Proof Define $S = \{i,j\}$. Conditional on any fixing of $X_{\sim S} = x_{\sim S}$, the conditional density of X_S is proportional to $\exp(-\langle x_S, \Theta_{SS}x_S\rangle/2 + \langle h, x_S\rangle)$ where h, the coefficient of the linear term, is determined by $x_{\sim S}$. This is the pdf of a Gaussian distribution with precision matrix Θ_{SS}, so using the explicit formula for 2×2 matrix inversion yields

$$\operatorname{Var}(X_i \mid X_{\sim S}) = (\Theta_{SS})^{-1}_{ii} = \frac{\Theta_{jj}}{\Theta_{ii}\Theta_{jj}-\Theta_{ij}^2}$$

which is equivalent to the claim. ■

The following crucial lemma says that if we invert a sparse matrix containing a nonnegligible off-diagonal entry, then its inverse contains a nonnegligible off-diagonal entry as well.

Lemma 47 *Let $\Theta \in \mathbb{R}^{n\times n}$ be a positive-definite matrix with at most $k+1$ nonzero entries in each row, and define $\Sigma := \Theta^{-1}$. Then there exist indices $i, \ell \in [n]$ such that $i \neq \ell$ and*

$$\frac{\Sigma_{i\ell}}{\sqrt{\Sigma_{ii}\Sigma_{\ell\ell}}} \geq \frac{1}{2k} \max_{a\neq b} \frac{\Theta_{ab}^2}{\Theta_{aa}\Theta_{bb}}.$$

Proof Since the statement of the lemma is invariant to rescaling (i.e. replacing Θ by $D\Theta D$ for any positive-definite diagonal matrix D), we may assume without loss of generality that $\Theta_{ii} = 1$ for all $i \in [n]$. Since Θ is positive definite, it follows that $\Theta_{ij}^2 < \Theta_{ii}\Theta_{jj} = 1$ for any indices $i \neq j$. Furthermore, by Lemma 46 and the law of total variance (where we define $X \sim N(0, \Sigma)$), we have

$$\Sigma_{ii} \geq \mathbb{E}\,\mathrm{Var}(X_i \mid X_{\sim i,j}) = \frac{1}{1-\Theta_{ij}^2}. \tag{23}$$

Since $I_n = \Sigma\Theta$, we have for any coordinate $i \in [n]$ that

$$1 = \sum_{\ell=1}^{n} \Theta_{i\ell}\Sigma_{i\ell} = \Sigma_{ii} + \sum_{\ell\neq i} \Theta_{i\ell}\Sigma_{i\ell},$$

and thus

$$\Sigma_{ii} - 1 = -\sum_{\ell\neq i} \Theta_{i\ell}\Sigma_{i\ell} \leq \|\Theta_{i,\sim i}\|_1 \max_{\ell} |\Sigma_{i\ell}| \leq k \max_{\ell\neq i} |\Sigma_{i\ell}|$$

where the final inequality uses that $\Theta_{i,\sim i}$ has at most k nonzero entries, each with magnitude at most one. If we define $\ell^\star(i) := \arg\max_{\ell\neq i} |\Sigma_{i\ell}|$, then the above display implies that

$$|\Sigma_{i\ell^\star(i)}| \geq \frac{\Sigma_{ii}-1}{k}. \tag{24}$$

We now consider two cases:

1. If $\max_i \Sigma_{ii} > 2$, define $i^\star := \arg\max \Sigma_{ii}$. Then by (24) and choice of $i^\star$, we have

$$|\Sigma_{i^\star\ell^\star(i^\star)}| \geq \frac{\Sigma_{i^\star i^\star}}{2k} = \frac{\max\{\Sigma_{i^\star i^\star}, \Sigma_{\ell^\star(i^\star)\ell^\star(i^\star)}\}}{2k}.$$

2. Otherwise we have $\max_i \Sigma_{ii} \leq 2$, so by (24) and (23) we have for any $i, j \in [n]$ with $i \neq j$ that

$$|\Sigma_{i\ell^\star(i)}| \geq \frac{\Sigma_{ii}-1}{k} \geq \frac{1/(1-\Theta_{ij}^2)-1}{k} \geq \frac{1/(1-\Theta_{ij}^2)-1}{2k} \max\{\Sigma_{ii}, \Sigma_{\ell^\star\ell^\star}\}.$$

   In particular, this bound holds when i and j are chosen to maximize $|\Theta_{ij}|$.

Therefore in either case, there exists some $i \in [n]$ such that

$$\frac{\Sigma_{i\ell^\star(i)}}{\max\{\Sigma_{ii}, \Sigma_{\ell^\star(i)\ell^\star(i)}\}} \geq \frac{1}{2k} \min\left\{1, \frac{1}{1-\max_{a\neq b}\Theta_{ab}^2} - 1\right\}.$$

Finally, using the inequality $1/(1-x)-1 \geq x$ which is valid for all $x < 1$, we obtain

$$\frac{\Sigma_{i\ell^\star(i)}}{\max\{\Sigma_{ii}, \Sigma_{\ell^\star(i)\ell^\star(i)}\}} \geq \frac{1}{2k} \min\left\{1, \max_{a\neq b}\Theta_{ab}^2\right\} = \frac{1}{2k} \max_{a\neq b}\Theta_{ab}^2.$$

Using the fact that $\max\{\Sigma_{ii}, \Sigma_{\ell^\star(i)\ell^\star(i)}\} \geq \sqrt{\Sigma_{ii}\Sigma_{\ell^\star(i)\ell^\star(i)}}$ proves the result. ■

This structural result yields a tester because we can directly estimate correlations from data. Very precise results about the sample correlation coefficient were obtained by Hotelling (1953). Below, we give a simple argument which yields easy-to-use nonasymptotic bounds.

We recall the following basic fact about Gaussians. See e.g. Lemma 2 of Kelner et al. (2020) for an explicit proof.

Lemma 48 (classical) *If X, Y are jointly Gaussian random variables with $\mathrm{Var}(Y) > 0$, then*

$$\mathrm{Var}(Y) - \mathrm{Var}(Y \mid X) = \frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Var}(Y)}.$$

For convenience, we use the following lemma. It is a special case of a general statement about testing for changes in conditional variance, which is closely related to classical results about non-central F-statistics and Wishart matrices (see e.g. Keener (2010)).

Lemma 49 (Special case of Lemma 12 of Kelner et al. (2020)) *For jointly Gaussian random variables X, Y with $\mathrm{Var}(Y|X) > 0$, define*

$$\gamma(Y; X) := \frac{\mathrm{Var}(Y) - \mathrm{Var}(Y \mid X)}{\mathrm{Var}(Y \mid X)}.$$

There exists an efficiently computable (i.e. polynomial time) statistic $\hat{\gamma}$ of m i.i.d. copies $(X_1, Y_1), \ldots, (X_m, Y_m)$ of (X, Y) such that

$$\left|\sqrt{\hat{\gamma}} - \sqrt{\gamma}\right| \leq \sqrt{\frac{4\log(4/\delta)}{m}} + \sqrt{\gamma/64}.$$

Theorem 50 *Suppose that $X \sim N(0, \Sigma)$, $\kappa \geq 0, \delta > 0$ and consider the following two hypotheses. The null hypothesis H_0 is that Σ is a diagonal matrix. The alternative hypothesis H_1 is that $\Sigma = \Theta^{-1}$ where Θ has $(k+1)$-sparse rows, and where*

$$\kappa \leq \max_{a \neq b} \frac{|\Theta_{ab}|}{\sqrt{\Theta_{aa}\Theta_{bb}}},$$

i.e. the maximal partial correlation coefficient is at least κ. Then provided $m = \Omega(k^2 \log(n/\delta)/\kappa^4)$ i.i.d. copies of X, we can distinguish in polynomial time between H_0 and H_1 with sum of probability of type I and type II errors at most δ.

Proof We test the maximal correlation of X_i and X_j over all $i, j \in [n]$ with $i \neq j$. Under the null hypothesis the true correlation $\gamma(X_i; X_j)$ equals zero no matter the choice of i, j. Under the alternative hypothesis, let i, j be the indices given by Lemma 47 and without loss of generality suppose $\mathrm{Var}(X_i) \leq \mathrm{Var}(X_j)$. Then $\gamma = \gamma(X_i; X_j)$ can be lower bounded as

$$\gamma = \frac{\mathrm{Var}(X_i) - \mathrm{Var}(X_i \mid X_j)}{\mathrm{Var}(X_i|X_j)} \geq \frac{\mathrm{Cov}(X_i, X_j)^2}{\mathrm{Var}(X_i)^2} \geq \frac{\mathrm{Cov}(X_i, X_j)^2}{\mathrm{Var}(X_i)\,\mathrm{Var}(X_j)} \geq \left(\frac{1}{2k} \max_{a \neq b} \frac{\Theta_{ab}^2}{\Theta_{aa}\Theta_{bb}}\right)^2$$

where the first inequality uses Lemma 48 and the fact that $\mathrm{Var}(X_i|X_j) \leq \mathrm{Var}(X_i)$. Thus, by the theorem assumption, it holds that $\gamma(X_i; X_j) \geq \kappa^4/(4k^2)$. By Lemma 49 and a union bound over all choices of i, j, given

$$m = \Omega(k^2 \log(4n/\delta)/\kappa^4)$$

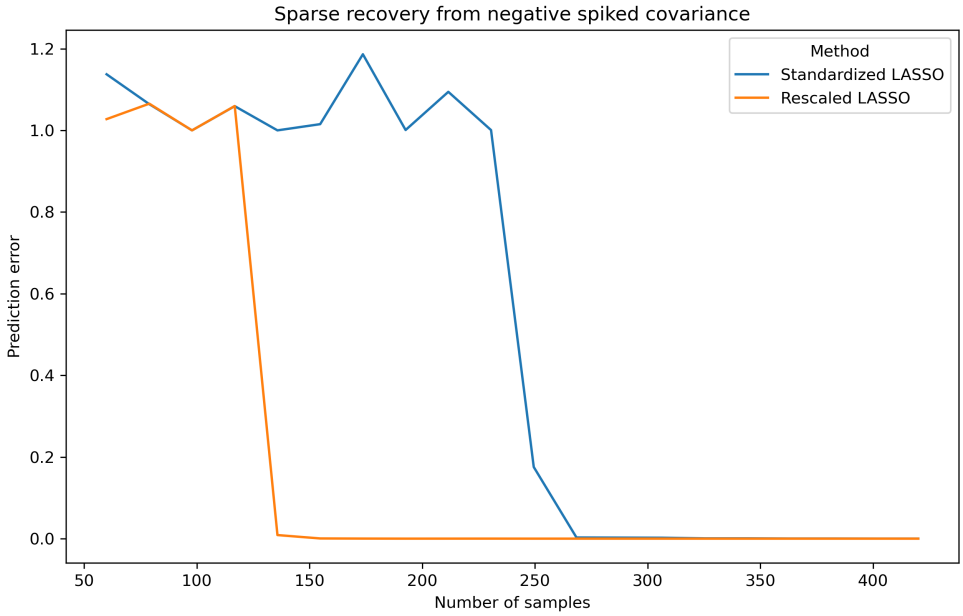

Figure 2: Standardized Lasso vs `RescaledLasso` in a simple example with varying number of samples. For each datapoint, the covariates were drawn i.i.d. from the negatively spiked sparse PCA model with ambient dimension $n = 300$ and with $\beta = -0.99$. For covariate vector X, the ground truth response Y is generated as $Y = \frac{1}{\sqrt{(1+\beta)k}}\langle 1_S, X\rangle$ where S is the set of coordinates of size $k = 5$ where the spike is supported. As we expect from the theory, `RescaledLasso` recovers the signal from fewer samples than Lasso applied with the usual standardization/normalization of covariates.

samples we can distinguish between the two hypotheses with the sum of probability of type I and type II error at most δ. ■

Appendix E. Simulation

As a simple example, we ran `RescaledLasso()` in a simple simulation on synthetic data where covariates follow negatively spiked sparse PCA (Figure 2) and verified that it indeed had improved prediction error compared to usual standardization of covariates. We used the glmnet package in R and optimized the regularization hyperparameter using a validation set.

When we ran the algorithm in the simulation, we changed the value of the hyperparameters DIV to 1.1 and of B to $2k$ — generally speaking, we expect setting the value of DIV closer to 1 will not significantly hurt the statistical performance (and may help a bit in some cases) although it may make the algorithm run a bit slower, and the proof does not depend on the particular value of DIV chosen in the original pseudocode (which was only chosen for mathematical convenience). Also, instead of updating the scale only of i_{min}, in each pass we we update the scale of all indices i such

that $\frac{1}{m}\|\mathbb{X}v^{(t,i)}\|_2^2 \leq 1$ — this significantly reduces the number of iterations and thus the runtime of the algorithm (it can be checked that the theoretical guarantees also hold with this modification).

Appendix F. Technical lemmas

Lemma 51 *Let $\Sigma \in \mathbb{R}^{n\times n}$ be a positive semi-definite matrix and fix $w \in \mathbb{R}^n$. Let $\mathbb{X} \in \mathbb{R}^{m\times n}$ have i.i.d. rows $X_1, \ldots, X_m \sim N(0, \Sigma)$. If $m \geq 32\log(2/\delta)$, then*

$$\Pr\left[\frac{1}{2}\|w\|_\Sigma^2 \leq \frac{1}{m}\|Xw\|_2^2 \leq 2\|w\|_\Sigma^2\right] \geq 1-\delta.$$

Proof By a change of variables, it suffices to consider the case $\Sigma = I_n$. But then $\|\mathbb{X}w\|_2^2 / \|w\|_2^2$ is distributed as a χ-squared random variable with m degrees of freedom. The statement follows from concentration of χ-squared random variables. ■

Lemma 52 (Hanson-Wright inequality (Rudelson and Vershynin, 2013)) *Let $\Sigma \in \mathbb{R}^{n\times n}$ be a positive semi-definite matrix. Let $X \sim N(0, \Sigma)$. Then for any $t > 0$,*

$$\Pr[|\,\|X\|_2^2 - \operatorname{tr}(\Sigma)| > t] \leq 2\exp\left(-c\min\left(\frac{t^2}{\|\Sigma\|_F^2}, \frac{t}{\|\Sigma\|_{\mathsf{op}}}\right)\right)$$

where $c > 0$ is a universal constant.

In particular we will use the following simplification of the Hanson-Wright inequality.

Corollary 53 *Let $\Sigma \in \mathbb{R}^{n\times n}$ be a positive semi-definite matrix. Let $X \sim N(0, \Sigma)$. Let $\delta \in (0, 1/4)$. Then*

$$\Pr[\|X\|_2^2 > C\operatorname{tr}(\Sigma)\log(2/\delta)] \leq \delta$$

where $C > 0$ is a universal constant.

Proof Observe that $\|\Sigma\|_{\mathsf{op}} \leq \|\Sigma\|_F \leq \operatorname{tr}(\Sigma)$ (since $\|\Sigma\|_{\mathsf{op}}$ is the ℓ_∞ norm of the eigenvalues of Σ, whereas $\|\Sigma\|_F$ is the ℓ_2 norm and $\operatorname{tr}(\Sigma)$ is the ℓ_1 norm). Thus, Lemma 52 gives that for any $t > 0$,

$$\Pr[|\,\|X\|_2^2 - \operatorname{tr}(\Sigma)| > t] \leq 2\exp\left(-c\min\left(\frac{t^2}{\operatorname{tr}(\Sigma)^2}, \frac{t}{\operatorname{tr}(\Sigma)}\right)\right).$$

Taking $t := \max(1, 1/c)\operatorname{tr}(\Sigma)\log(2/\delta)$ gives the claimed result. ■

For a symmetric matrix A, let $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ denote the eigenvalues of A. The following inequality is well-known.

Lemma 54 (Weyl's inequality) *Let $N, R \in \mathbb{R}^{n\times n}$ be symmetric matrices. Suppose that $\operatorname{rank}(R) = r$. Then*

$$\lambda_{r+1}(N+R) \leq \lambda_1(N).$$