Attribute-efficient learning of monomials over highly-correlated variables

Alexandr Andoni ANDONI@CS.COLUMBIA.EDU
Computer Science Department and Data Science Institute
Columbia University
New York City, NY 10027, USA

Rishabh Dudeja RD2714@COLUMBIA.EDU
Department of Statistics
Columbia University
New York City, NY 10027, USA

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Computer Science Department and Data Science Institute
Columbia University
New York City, NY 10027, USA

Kiran Vodrahalli KIRAN.VODRAHALLI@COLUMBIA.EDU
Computer Science Department
Columbia University
New York City, NY 10027, USA

Editors: Aurélien Garivier and Satyen Kale

Abstract

We study the problem of learning a real-valued function of correlated variables. Solving this problem is of interest since many classical learning results apply only in the case of learning functions of random variables that are independent. We show how to recover a high-dimensional, sparse monomial model from Gaussian examples with sample complexity that is poly-logarithmic in the total number of variables and polynomial in the number of relevant variables. Our algorithm is based on a transformation of the variables—taking their logarithm—followed by a sparse linear regression procedure, which is statistically and computationally efficient. While this transformation is commonly used in applied non-linear regression, its statistical guarantees have never been rigorously analyzed. We prove that the sparse regression procedure succeeds even in cases where the original features are highly correlated and fail to satisfy the standard assumptions required for sparse linear regression.

Keywords: attribute-efficient, computationally efficient, statistics, monomial, learning, restricted eigenvalue condition, lasso, log-transform, dependent features

1. Introduction

Consider the following canonical problem in learning theory. We observe n features-response pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ drawn i.i.d. from the following model:

$$x_i \sim \mathcal{D}_x, \quad y_i = f(x_i).$$

Here, $\mathcal{D}_x$ is some distribution on $\mathbb{R}^p$. The goal is to design an algorithm to accurately estimate the unknown function f with small sample complexity (n) and small run-time. Moreover, the unknown function f may depend on only k out of the p features, with $k \ll p$. This models the problem of feature selection in machine learning and statistics. In this situation, a reasonable goal is to design algorithms that are *attribute-efficient*—that is, require $n = \mathrm{poly}(\log(p), k)$ samples and $\mathrm{poly}(n, p, k)$ run-time. While there is a long line of work studying this problem, most existing work has one or more of the following limitations:

1. Many existing results provide algorithms and hardness results when the features are Boolean, i.e., $\mathcal{D}_x$ is supported $\{0, 1\}^p$ or $\{-1, +1\}^p$. These results, however, do not necessarily reflect the difficulty or qualities of the learning problem when the features are real-valued, which is common in many practical settings.
2. A long line of work in compressed sensing and high-dimensional statistics assumes f is a (sparse) linear function, but does not extend to non-linear functions.
3. To the best of our knowledge, all existing work for real-valued attributes and non-linear functions f assumes that $\mathcal{D}_x$ is a product measure, for example a standard normal $\mathcal{D}_x = \mathcal{N}(0, I_p)$ (Andoni et al., 2014).

In particular, the question of attribute-efficient learning is not well understood even for simple classes of non-linear functions and some canonical non-product measures. In this work, we address this gap by considering the problem of learning sparse monomials in the noiseless setting under the Gaussian measure. In particular, we assume:

$$\mathcal{D}_x = \mathcal{N}(0, \Phi), \qquad f(x) = \prod_{i \in S} x_i^{\beta_i}.$$

For simplicity, we assume that covariance matrix Φ satisfies $\Phi_{i,i} = 1$ for all i, since we can rescale the features to have unit variance. The $\beta_i \in \mathbb{N} \cup \{0\}$ are degrees of each of the relevant variables $S \subseteq \{1, \ldots, p\}$, and $|S| = k$. Even in this simple setup, a number of standard approaches fail to give an algorithm that runs in $\mathrm{poly}(n, p, k)$ time and has $\mathrm{poly}(\log(p), k)$ sample complexity.

1. One natural approach is to expand the feature space by constructing all possible monomials of degree $\leq d$ and consisting of at most k variables (there are at least $\Omega(p^k)$ such monomials) and using Empirical Risk Minimization. One expects this procedure to work with sample size $n = O(\log(p^k)) = O(k \log(p))$, but the approach is computationally inefficient. Sparse regression (e.g., Tibshirani, 1996)) in the expanded feature space has similar sample complexity and run-time (and may require additional assumptions on the expanded design matrix). Negahban and Shah (2012) analyze this approach when $\mathcal{D}_x$ is the uniform distribution on $\{-1, 1\}^p$ and f is a sum of s monomial terms and obtain a sample complexity of $O(ps^2)$ and a run-time of $O(2^p)$.
2. One can avoid explicit feature expansion by using the kernel trick. Kernel ridge regression is equivalent to ℓ_2-penalized least squares in the expanded feature space, and can be solved in $\mathrm{poly}(n, p, k)$ time. Standard analyses of kernel ridge regression imply that the sample complexity of this approach is proportional to the Rademacher complexity of linear classes with bounded ℓ_2 norm in the expanded space (e.g., Bartlett and Mendelson, 2002, Lemma 22). Unfortunately, the latter quantity depends on the average squared norm of the feature vector in the expanded

space, which in the Gaussian case scales like $\Omega(p^k)$. We also refer the reader to Theorem 2 of Quang (2006) for a precise analysis of the L_2 risk bound which makes the $p^{\Omega(d)}$ dependence explicit for kernel ridge regression when f is a degree d polynomial and $\mathcal{D}_x$ is supported on the unit sphere.

3. Andoni et al. (2014) describe an algorithm that learns a degree-k polynomial with at most s monomial terms under a product measure on $\mathbb{R}^p$, achieving a run-time and sample complexity of $\mathrm{poly}(p, 2^k, s)$. There is a natural reduction of our problem to their setting: learn the matrix Φ and then apply a whitening transformation $\Phi^{-1/2}$ to the feature vectors. However, this reduction may convert a degree-k monomial over the original features into a dense polynomial with $s = \Omega(p^k)$ terms over the new features.

Our contributions. We design an attribute-efficient algorithm for learning the function $f(x) = \prod_{i \in S} x_i^{\beta_i}$, where $x \sim \mathcal{D}_x = \mathcal{N}(0, \Phi)$, that uses sample size $n = O(k^2 \cdot \mathrm{poly}(\log(p), \log(k)))$ and runs in $\mathrm{poly}(n, p, k)$ time. In particular, the algorithm exactly recovers the set S and exponents β_i with high probability. The algorithm does not have access to Φ, and indeed, the sample size may be too small to learn it accurately.

Our algorithm provably succeeds as long as $\max_{i \neq j} |\Phi_{i,j}| < 1$. This is, in a sense, the minimal assumption on Φ: if violated, this model is not even *identifiable*. To put this into context, it is instructive to contrast to the case when f is a sparse *linear* function, under the same input distribution $x \sim \mathcal{N}(0, \Phi)$. For the latter problem, there is no known computationally efficient and attribute-efficient algorithm to estimate the set S under similarly-weak assumptions on Φ.

The key algorithmic technique is to apply a log-transform to the features and response, and reduce the problem to a sparse linear regression problem. While this is a commonly-used technique in applied statistics, to the best of our knowledge, it has not been rigorously analyzed before. We show that this log-transform is precisely what allows us to provably learn f when it is a monomial. Specifically, we analyze how the covariance matrix changes after the log-transform, showing that the log-transform eliminates linear dependencies between two or more features. To again contrast with the case of learning sparse *linear* functions, the linear dependencies are precisely the obstacle for designing computationally-efficient and attribute-efficient algorithms.

2. Preliminaries

This section presents the formal learning problem, and introduces technical tools and notations used in our algorithm and analysis.

2.1. Problem statement

We observe n i.i.d. feature-response pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ from the following model:

$$x_i \sim \mathcal{N}(0, \Phi), \quad y_i = \prod_{j \in S} x_{i,j}^{\beta_j}, \tag{1}$$

where $S \subseteq [p] := \{1, \ldots, p\}$ is the set of relevant variables, and $\beta \in (\mathbb{N} \cup \{0\})^p$ is the vector of degrees (with $\beta_j \neq 0$ iff $j \in S$). The total degree of the monomial is $\|\beta\|_1 = \sum_{j \in S} \beta_j$. We say the monomial is k-sparse when $|S| = k$. (Our results also permit $\beta_j < 0$, but such a model would not be a monomial.)

The attribute-efficient learning goal is to recover, with high probability, both S and β with sample size $n = \text{poly}(\log(p), k)$ and run-time $\text{poly}(n, p, k)$.

For simplicity, we assume that the features are standardized, so the feature variances satisfy $\Phi_{i,i} = 1$ for all $i \in [p]$. We also assume the cross-correlations satisfy

$$|\Phi_{i,j}| \leq 1 - \epsilon, \quad \forall i \neq j$$

for some $\epsilon > 0$. This latter assumption is necessary so that β is identifiable. Indeed, if there are two perfectly correlated features, then it is impossible to distinguish them, in which case β cannot be uniquely determined. These assumptions are not restrictive and still permit highly correlated features. In particular, the covariance matrix is permitted to be rank deficient, so some features can be linear combinations of others.

2.2. Concepts and results from compressed sensing

Our main algorithm is based on a reduction to sparse linear regression / compressed sensing.

The main problem in compressed sensing is to recover an s-sparse vector $w \in \mathbb{R}^p$ from observations of the form $Aw + \eta = b$ where $A \in \mathbb{R}^{n \times p}$ is the sensing matrix, $w \in \mathbb{R}^p$ is the signal vector, $\eta \in \mathbb{R}^n$ is the measurement noise, and $b \in \mathbb{R}^n$ is the observation vector. A commonly-used estimator is the Lasso (Tibshirani, 1996): for $\vartheta > 0$, the Lasso estimator is $\hat{w}_{\text{Lasso}}(\vartheta) := \arg\min_{u \in \mathbb{R}^p} \frac{1}{2n} \|Au - b\|_2^2 + \vartheta \|u\|_1$. This estimator succeeds in recovering w under certain conditions on A. One such condition is the restricted eigenvalue condition introduced by Bickel et al. (2009) which we review here.

Definition 1 *For $T \subset [p]$ and $q_0 > 0$, define $\mathcal{C}(q_0, T) := \{v \in \mathbb{R}^p : \|v\|_2 = 1, \|v_{T^c}\|_1 \leq q_0 \|v_T\|_1\}$. T is commonly taken to be the non-zero support S of the sparse vector to recover. We say the (q_0, T, A)-restricted eigenvalue condition* (REC) *is satisfied by matrix $A \in \mathbb{R}^{n \times p}$ if $\tilde{\lambda}(q_0, T, A) := \min_{v \in \mathcal{C}(q_0, T)} \frac{1}{n} \|Av\|_2^2 > 0$. When q_0 and T are apparent from context and $|T| = s$, we will simply write $\tilde{\lambda}(s, A)$.*

The following well-known result about the performance of the estimator $\hat{w}_{\text{Lasso}}(\vartheta)$ is due to Bickel et al. (2009); the specific form we state is taken from Hastie et al. (2015).

Theorem 2 *Consider the model $Aw + \eta = b$, and suppose the support S of $w \in \mathbb{R}^p$ has size k, and the measurement matrix $A \in \mathbb{R}^{n \times p}$ satisfies (q_0, S, A)-REC with $q_0 = 3$. For any $\vartheta > 0$ such that $\vartheta \geq (2/n) \|A^T \eta\|_\infty$, the Lasso estimate $\hat{w}_{\text{Lasso}}(\vartheta)$ satisfies*

$$\|w - \hat{w}_{\text{Lasso}}(\vartheta)\|_2 \leq \frac{3\vartheta\sqrt{k}}{\tilde{\lambda}(k, 3, S, A)}.$$

2.3. Additional notations

Let $X = [x_1 | \cdots | x_n]^T \in \mathbb{R}^{n \times p}$ be the data matrix, and let $y = [y_1 | \cdots | y_n]^T \in \mathbb{R}^n$ be the vector of responses. Throughout, $\log$ denotes the natural logarithm, and applying $\log$ or absolute value to a matrix or vector means these operations are taken element-wise. For any matrix M, we write $M^{(l)}$ to denote its l-th Hadamard power, so $M_{i,j}^{(l)} = M_{i,j}^l$.

3. Learning sparse monomials

In this section, we present our learning algorithm and its performance guarantees.

3.1. Algorithm

Our proposed attribute-efficient learning algorithm, given as Algorithm 1, is based on a log-transformation of the data, followed by sparse linear regression. For concreteness, we use Lasso (Tibshirani, 1996) for the second step, although other sparse regression methods could also be used.

Algorithm 1 Learn Sparse Monomial

```
Require: data matrix $X \in \mathbb{R}^{n \times p}$, responses $y \in \mathbb{R}^n$, regularization parameter $\vartheta > 0$
1: Apply $\log\big(|\cdot|\big)$ transformation to data and responses, element-wise: $\hat{X} \leftarrow \log\big(|X|\big)$ and $\hat{y} \leftarrow \log\big(|y|\big)$.
2: Solve Lasso optimization problem: $\hat{\beta} \leftarrow \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|\hat{X}\beta - y\|_2^2 + \vartheta \|\beta\|_1$.
3: Select variables: $\hat{S} \leftarrow \{j \in [p] : \hat{\beta}_j \neq 0\}$.
4: return $\hat{S}$ and $\hat{\beta}$.
```

The logarithm transformation is a folklore technique in applied statistics (see, e.g., Keene, 1995) but, to the best of our knowledge, has not received a non-trivial theoretical analysis in a setting similar to ours. We compose the log-transform with absolute value in Algorithm 1 to ensure non-negativity.

We make two observations about the $\log\big(|\cdot|\big)$-transformation. First, it converts the monomial model in Eq. (1) to the following:

$$\log\big(|y_i|\big) = \sum_{j \in S} \beta_j \log\big(|x_{i,j}|\big) \tag{2}$$

Second, the transformation is only applicable to non-zero entries in the data matrix X and response vector y. For the Gaussian data in our problem setup, all entries are non-zero almost surely.

So, after the transformation, the problem reduces to a linear sparse recovery problem, which can be efficiently solved using well-known techniques from compressed sensing under appropriate conditions on the design matrix (e.g., restricted eigenvalues).

The following simple proposition formalizes the reduction.

Proposition 3 *A unique solution $\hat{\beta}$ to the transformed model in Eq. (2) is the unique solution to the original model in Eq. (1).*

Proof See Appendix A. ■

3.2. Performance guarantees

Our approach to analyzing Algorithm 1 is based on applying the performance guarantee for Lasso from Theorem 2. Because we apply Lasso to data from the $\log\big(|\cdot|\big)$-transformed model in Eq. (2), we need to prove that REC is satisfied by $\hat{X} = \log\big(|X|\big)$. As noted before, it is sufficient to lower-bound $\tilde{\lambda}(k, 3, S, \hat{X}/\sqrt{n})$. This is the content of the following theorem.

Theorem 4 *Let $\delta \in (0,1)$ be an arbitrary confidence parameter. Suppose the covariance matrix Φ satisfies $\Phi_{i,i} = 1$, $\forall i \in [p]$ and $\max_{i \neq j} |\Phi_{i,j}| < 1 - \epsilon$. Then, the $\log\big(|\cdot|\big)$-transformed design matrix $\hat{X} = \log\big(|X|\big)$ for X taken from the model in Eq. (1) with true support $|S| = k$ satisfies*

$$\tilde{\lambda}\left(k, \frac{1}{\sqrt{n}}\hat{X}\right) \geq \frac{1}{5}\sqrt{\frac{\epsilon}{\log(16k) + 2}},$$

with probability $1 - \delta$, provided that

$$n \geq C \cdot \frac{k^2 \log(2k)}{\epsilon} \cdot \log^2\left(\frac{2p}{\delta}\right) \cdot \log^2\left(\frac{k\log(k)}{\epsilon} \log\left(\frac{2p}{\delta}\right)\right). \tag{3}$$

In the above display, C is a universal constant.

Therefore, applying Theorem 2, we immediately get as a corollary the following performance guarantee for Algorithm 1.

Corollary 5 *Let $\delta \in (0,1)$ be an arbitrary confidence parameter and ϑ be the regularization parameter. Suppose the covariance matrix Φ satisfies $\Phi_{i,i} = 1$ for all $i \in [p]$ and $\max_{i \neq j} |\Phi_{i,j}| < 1 - \epsilon$, and that the sample size n satisfies the inequality in Eq. (3). For X and y taken from the model in Eq. (1) with $|S| = k$, Algorithm 1 returns $\hat{\beta}$ such that, with probability at least $1 - \delta$,*

$$\|\hat{\beta} - \beta\|_2 \leq 15\vartheta\sqrt{\frac{k(\log(16k) + 2)}{\epsilon}}.$$

Remark 6 *We note that, as $\vartheta \to 0$, $\|\hat{\beta} - \beta\|_2 \to 0$, and hence, Algorithm 1 recovers β exactly. Furthermore, in the limit $\vartheta \to 0$, Algorithm 1 is equivalent to the Basis Pursuit estimator (Chen and Donoho, 1994) defined as:*

$$\hat{\beta}^{BP} = \underset{v \in \mathbb{R}^p}{\arg\min} \|v\|_1 \textit{ subject to } \hat{X}v = \hat{y}.$$

In particular, this means that under the conditions of Corollary 5, the Basis Pursuit estimator satisfies

$$\hat{\beta}^{BP}_j = 0 \quad \forall j \notin S, \quad \hat{\beta}^{BP}_j = \beta_j \quad \forall j \in S.$$

Remark 7 *Suppressing logarithmic factors in p and k, the above result shows that Algorithm 1 succeeds in recovering the monomial with high probability with $\tilde{O}(k^2/\epsilon)$ samples.*

Remark 8 *If we observe data with multiplicative noise, that is,*

$$y_i = e^{\eta_i} \cdot \prod_{j \in S} x_{i,j}^{\beta_j} \tag{4}$$

where $\eta_i \in \mathbb{R}$ is a zero-mean sub-gaussian noise (e.g., $\eta_i \sim \mathcal{N}(0, \sigma^2)$), then the $\log|\cdot|$ transform reduces our problem to a noisy compressed sensing problem. Hence we can still apply Theorems 2 and 4, as long as we set the parameter ϑ according to the noise level. If the sample size is large enough relative to the noise level, we can exactly recover the degrees by rounding $\hat{\beta}$ to nearest integers. The details are straightforward and omitted.

4. Restricted eigenvalues for the $\log\big(|\cdot|\big)$-transformed data

In this section, we present the main technical results used in the proof of Theorem 4. We define the following notations:

$$z := \log(|x|), \qquad \Sigma := \mathbb{E}_z[zz^T], \qquad \hat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} z^{(i)} {z^{(i)}}^T.$$

where $\log(|\cdot|)$ is applied elementwise and $z^{(i)}$ denotes the i^{th} empirical data point. We also use z_i to denote the i^{th} feature of z. The proof of Theorem 4 involves three steps:

1. We first determine an explicit formula for the population covariance matrix Σ given in Lemma 9.
2. We leverage this explicit formula to prove a lower bound on $\lambda_{\min}(\Sigma)$ and $\tilde{\lambda}(k, \Sigma^{1/2})$ in Theorem 10.
3. Finally, we show that $\tilde{\lambda}(k, \hat{\Sigma}^{1/2})$ concentrates around $\tilde{\lambda}(k, \Sigma^{1/2})$ in Lemma 13.

One of our main technical contributions is a lower bound on $\tilde{\lambda}_{\min}(k, \Sigma^{1/2})$ under very weak assumptions about the covariance matrix Φ of the original features, namely, $|\Phi_{i,j}| < 1 - \epsilon$ for any $i \neq j$. In particular, this holds even in cases where Φ is low-rank or $\Phi^{1/2}$ doesn't satisfy REC. Intuitively, this result holds because the logarithm, a highly non-linear operation, destroys the linear dependence structure of a low-rank matrix as long as no two features are perfectly correlated (which is anyway necessary for identifiability).

4.1. Properties of $\log\big(|\cdot|\big)$-transform

The following key lemma provides several useful properties of the $\log|\cdot|$ transform, culminating in an explicit and convenient expression for Σ in terms of Φ.

Lemma 9 *Let $x \sim \mathcal{N}(0, \Phi)$ where $\Phi_{i,i} = 1$ for all $i \in [p]$. Define $z = \log\big(|x|\big)$. Then:*

1. *The random variable z_i has bounded variance, in particular,* $\mathrm{var}(z_i) = \pi^2/8$.
2. *The function $a \mapsto \log(|a|)$ admits the following expansion in the Hermite polynomial basis $\{H_l\}_{l\geq 0}$:*

$$\log(|a|) = \sum_{l=0}^{\infty} c_{2l} H_{2l}(a), \quad c_{2l} = \frac{(-1)^{l-1} 2^{l-1} (l-1)!}{\sqrt{(2l)!}}.$$

3. $\mathbb{E}[z_i z_j] = \sum_{l=0}^{\infty} c_{2l}^2 \Phi_{i,j}^{2l}$.
4. $\Sigma = c_0^2 \boldsymbol{1}_{p\times p} + \sum_{l=1}^{\infty} c_{2l}^2 \Phi^{(2l)}$, *where $\boldsymbol{1}_{p\times p}$ is the $p \times p$ matrix of all 1's.*

Proof [Proof sketch]

1. The challenge in calculating the variance of z_i is that integrals involving log moments and the Gaussian measure are not analytically easy to work with. To get around this, we leverage the fact that for any non-negative random variable a and any $m \in \mathbb{N}$,

$$\mathbb{E}_a[\log^m a] = \lim_{\nu\to 0} \tfrac{\mathrm{d}^m}{\mathrm{d}\nu^m} \mathbb{E}_a[a^\nu].$$

When $a = |x_i|$, the RHS of the above expression is available in closed-form. (This is the "Replica Trick" from statistical physics (Edwards and Anderson, 1975).)

2. Since the Hermite polynomials form a complete orthonormal basis for $L_2(\mathcal{N}(0,1))$, we can compute c_l by the integral:

$$c_l = \int_{-\infty}^{\infty} \log\left(|a|\right) \cdot H_l(a) \cdot \frac{e^{-a^2/2}}{\sqrt{2\pi}}\,\mathrm{d}a.$$

   We calculate the above integral by-parts and by leveraging the recursive structure of Hermite Polynomials.

3. The rationale behind expanding the $\log(|a|)$ in the Hermite polynomial basis is that there is a clean formula between the correlation of Hermite polynomials applied to two correlated Gaussian random variables (see, e.g., O'Donnell, 2014):

$$\mathbb{E}[H_l(x_i)H_m(x_j)] = \Phi_{i,j}^l \mathbb{1}_{\{l=m\}}.$$

   Using this fact and the expansion of $\log|\cdot|$ gives us the expression for $\mathbb{E}[z_i z_j]$.

4. The formula for Σ immediately follows given the general expression for $\Sigma_{i,j} = \mathbb{E}[z_i z_j]$.

See Appendix B for a detailed proof. ■

4.2. Restricted eigenvalues of population covariance matrices

Theorem 10 *Let Φ be any covariance matrix with $\Phi_{i,i} = 1$ for all i and $|\Phi_{i,j}| < 1 - \epsilon$ for $i \neq j$, and let $\Sigma = \mathbb{E}_z[zz^T]$ for $z \sim \mathcal{N}(0, \Phi)$. The following inequalities hold.*

1. $\lambda_{\min}(\Sigma) \geq \frac{\pi^2}{8}\lambda_{\min}(\Phi).$

2. $\tilde{\lambda}\left(k, \Sigma^{1/2}\right) \geq \sum_{\ell=1}^{\frac{1}{2}\left\lfloor \frac{\log(16k))}{\log(\frac{1}{1-\epsilon})} \right\rfloor} \frac{\tilde{\lambda}\left(k, [\Phi^{(2\ell)}]^{1/2}\right)}{5\ell^{3/2}} + \frac{2}{5}\sqrt{\frac{2\log((1-\epsilon)^{-1})}{\log\left(\frac{16k}{1-\epsilon}\right) + \max\{2, \log((1-\epsilon)^{-1})\}}}.$

Remark 11 *If Φ already has a positive minimum eigenvalue, we automatically have a constant multiplicative factor improvement after applying the* $\log\left(|\cdot|\right)$*-transformation. But even if $\lambda_{\min}(\Phi) = 0$, we still obtain a positive lower bound on $\tilde{\lambda}\left(k, \Sigma\right)$.*

Remark 12 *In Appendix C (specifically Theorem 30), we also prove a simpler minimum eigenvalue lower bound of $\lambda_{\min}(\Sigma) \geq \Omega(\sqrt{\epsilon/\log(p)})$, which is similar to the lower bound on $\tilde{\lambda}(k, \Sigma^{1/2})$ except with* $\log(k)$ *replaced by* $\log(p)$*. The improvement in Theorem 10, which has no explicit dependence on the ambient dimension p, uses a restricted form of Gershgorin's Circle Theorem (Lemma 32). Using either lower bound is sufficient to obtain the sample complexity guarantees in Theorem 4, but the improved bound highlights the power of the* $\log\left(|\cdot|\right)$*-transformation and may be of independent interest.*

Proof [Proof sketch] We recall the explicit expression for Σ from Lemma 9:

$$\Sigma = c_0^2 \mathbf{1}_{p\times p} + \sum_{l=1}^{\infty} c_l^2 \Phi^{(l)}.$$

The definitions of $\lambda_{\min}(\cdot)$ and $\tilde{\lambda}(k,\cdot)$ imply both are superadditive:

$$\tilde{\lambda}(k,\Sigma^{1/2}) \geq \sum_{l=1}^{\infty} c_l^2 \tilde{\lambda}(k,[\Phi^{(l)}]^{1/2}).$$

We obtained the bound on $\lambda_{\min}(\Sigma)$ by applying a linear algebraic result from Bapat and Sunder (1985) which implies that $\lambda_{\min}(\Phi^{(l)}) \geq \lambda_{\min}(\Phi)$. As for the second expression, we split the infinite sum into two parts and apply a restricted version of the Gershgorin Circle Theorem to the second part (see Lemma 32 in Appendix C.2.2). We then analyze how fast the coefficients c_l of the remaining terms decay to 0. We refer the reader to Appendix C for a complete proof. ■

4.3. Analysis of the empirical covariance matrix

The last piece required to complete the proof Theorem 4 is a concentration result about $|\tilde{\lambda}(k,\Sigma^{1/2}) - \tilde{\lambda}(k,\hat{\Sigma}^{1/2})|$. This is stated in the following lemma.

Lemma 13 *Let $\delta \in (0,1)$ be an arbitrary confidence parameter. With probability $1-\delta$,*

$$|\tilde{\lambda}(k,\Sigma^{1/2}) - \tilde{\lambda}(k,\hat{\Sigma}^{1/2})| \leq Ck\left(\sqrt{\frac{(\log(3/\delta)+2\log(p))}{n}} + \frac{\log^2(n)(\log(3/\delta)+2\log(p))^2}{n}\right).$$

In the above display C is a universal constant.

Proof [Proof sketch] We apply Theorem 4.2 of (Kuchibhotla and Chakrabortty, 2018) after verifying that the log-transformed features z_i are entry-wise sub-exponential. See Appendix D for a detailed proof. ■

5. Simulations

We conducted a simple simulation to evaluate the robustness of our procedure to small *additive* noise (which our analysis does not cover). The $p = 512$-dimensional feature vectors are $x_i \sim \mathcal{N}(0,\Phi)$ for a rank-$p/2$ covariance matrix Φ given by

$$\Phi := \begin{bmatrix} I & \sqrt{\frac{2}{p}}H \\ \sqrt{\frac{2}{p}}H & I \end{bmatrix}.$$

Above, I is the $(p/2) \times (p/2)$ identity matrix, and H is the $(p/2) \times (p/2)$ Hadamard matrix. The responses are $y_i = \prod_{j\in S} x_{i,j} + \eta_i$ for independent $\eta_i \sim \mathcal{N}(0,\sigma^2)$, where $\sigma = 10^{-3}$, and $S = \{1,\dots,k/2, p/2+1,\dots,p/2+k/2\}$.

Algorithm 1 was implemented with a setting of $\vartheta = \vartheta(n,p,\sigma)$ as suggested by Bickel et al. (2009). For different values of the cardinality $k = |S|$ and sample size n, we estimated the probability of exact recovery of S on 100 independent trials:

Estimated probability of exact recovery				
	$k=2$	$k=4$	$k=6$	$k=8$
$n=128$	0.99	0.76	0.01	0.00
$n=384$	1.00	1.00	0.97	0.22
$n=640$	1.00	1.00	1.00	0.88

The results suggest that our procedure tolerates some level of additive noise, but that the sample size may need to increase significantly with the sparsity level k. This is reasonable, as the signal-to-noise ratio decreases exponentially with k.

6. Related work

There are a large number of results on attribute-efficient learning under different assumptions on $\mathcal{D}_x$ and the target function f. We discuss representative results from each category.

6.1. Learning with Boolean features

When $\mathcal{D}_x$ is supported on $\{0,1\}^p$, learning monomials with positive integral degrees is the same as learning conjunctions. This class was shown to be PAC learnable by Valiant (1984). Furthermore, there also exists a computationally efficient and attribute-efficient learner due to Littlestone (1988). When $\mathcal{D}_x$ is supported on $\{-1,+1\}^p$, then learning monomials with positive integral degrees corresponds to learning parities. Parity functions are PAC learnable using Gaussian elimination over $\mathbb{F}_2$ in time $O(n^3)$ (Helmbold et al., 1992).

When a parity function involves only k variables, a brute force search over all size-k subsets of variables PAC learns k-sparse parities with an attribute-efficient sample complexity of $\mathrm{poly}(\log(p),k)$ but has a run-time of $O(p^k)$. Finding an attribute-efficient algorithm with $\mathrm{poly}(n,p,k)$ run-time is a long-standing open problem of Blum (1998). Some notable improvements over the brute-force run-time include an attribute-efficient algorithm with run-time $O(p^{k/2})$ due to Dan Spielman (Klivans and Servedio, 2006), and an attribute-inefficient improper learner with sample complexity $n = O(p^{1-1/k})$ and run-time $O(p^4)$ for the noiseless case with an arbitrary distribution over $\{-1,+1\}^p$ due to Klivans and Servedio (2006). Finally, an $O(p^{0.8k}\,\mathrm{poly}\,\big(1/(1-2\eta)\big)$-time (but attribute-inefficient) algorithm of Valiant (2015) learns parities in the noisy setting (where labels are flipped with probability η) under the uniform distribution.

6.2. Average case analysis for learning parities

The key bottleneck in avoiding the $p^{O(k)}$ dependence in run-time while learning k-sparse parities over the uniform distribution on $\{-1,+1\}^p$ in an attribute-efficient manner is that it is not clear how to decide if a feature is relevant or not without considering its interaction with every possible set of $k-1$ features. In light of this, Kalai et al. (2009) study the problem when f is a DNF with s terms (k-sparse parities are DNFs of size $s=2^k$) and show that a natural greedy feature selection algorithm can learn such f in time and sample complexity $\mathrm{poly}(s,p)$ under a product distribution whose parameters are adversarially chosen and then randomly perturbed by a small amount. Similarly, Kocaoglu et al. (2014) identify a property of the function f called the unique sign property (USP) that facilitates learning. For functions f defined on $\{-1,+1\}^p$ satisfies USP and depends on just k features, their algorithm learns f under the uniform distribution with run-time and sample complexity $\mathrm{poly}(p,2^k)$. In the spirit of smoothed analysis, they show the USP is satisfied when an adversarially chosen k-sparse function f is perturbed by a small amount of random noise.

6.3. Learning with real-valued features

When $\mathcal{D}_x$ is a product measure and the features are real-valued (for example, the uniform measure on $[-1, 1]^p$ or the standard Gaussian measure on $\mathbb{R}^p$), Andoni et al. (2014) consider the problem of learning sparse polynomials of degree d that contain at most s monomial terms with additive noise. They show a surprising result that in contrast to learning sparse parities with noise, it is possible to avoid a p^d dependence in run-time. They design an algorithm with $\mathrm{poly}(p, 2^d, s)$ sample complexity and run-time. At the heart of their approach are linear-time correlation tests that detect if a feature participates in the highest degree (lexicographically) monomial. Once they detect all features participating in the highest degree monomial, they remove it, and recurse on the residual polynomial. An interesting property of their algorithm is that it never looks at the signs of either the responses or the features. This highlights the fact that in the real-valued case the magnitudes of the observations contain valuable information (which was not present in the case of parities) that can be leveraged to design algorithms with sub-$O(p^d)$ run-time. The algorithm we propose has the same property. While the class of functions we can handle is smaller (1-sparse polynomials), we are able to handle extremely large correlations between features. In this highly-correlated setting, it is not immediately clear how to analyze the correlation tests proposed by Andoni et al. (2014). Hence, we rely on a completely different technique: computing a log-transform of the responses and using sparse linear regression.

7. Summary and open problems

In this paper, we studied the problem of learning sparse monomials of highly-correlated features. Our work provides the first attribute-efficient analysis (handling arbitrarily high correlations) in a non-product distribution setting, which has been a major challenge in the prior work (e.g., Kalai et al. (2009); Andoni et al. (2014)). By leveraging a folklore technique from applied statistics, namely applying the $\log(|\cdot|)$ transform to the features and responses, we reduced this problem to a sparse linear regression problem. By analyzing how the covariance matrix changes after the $\log(|\cdot|)$ transform, we show that our procedure works under the minimal conditions required for the model to be identifiable.

We summarize the conceptual contributions of the paper as follows.

1. Learning degree-k sparse polynomial functions with $\mathrm{poly}(\log(p), k)$ samples in $p^{o(k)}$-time under non-product distributions is a challenging problem. Our work gives a new algorithmic line-of-attack for this problem, namely transforming both the response and the features such that each relevant variable participates in an $O(1)$-degree interaction in the transformed model, reducing the computational burden of searching for relevant variables from $p^{\Omega(k)}$ to $p^{O(1)}$. Although we study this general principle in a specialized setting, we believe our techniques (Lemma 9) can be useful for analyzing other instances of this algorithmic idea. The fact that no existing approach gives attribute-efficient algorithms with $p^{o(k)}$ run-time for the comparatively simple sparse monomial problem underscores the promise of this approach.

2. Our analysis uncovers a *blessing of non-linearity*. Specifically, the assumptions on the correlation structure needed to learn a class of sparse non-linear functions are less restrictive than

those needed to learn sparse linear functions. We require only minimal assumptions on the dependence structure to ensure identifiability, a significant departure from previous results.

3. We demonstrate the minimum eigenvalue of the log-transformed data covariance matrix is strictly positive with high probability, regardless of the initial rank. Thus, nonlinear data transformations can destroy low-rank covariance structure, a principle which may be useful for other estimation problems.

We conclude with a few open problems. The most immediate is to find an efficient algorithm for learning sparse monomials in the presence of additive noise, which our simulations suggest should be possible. Beyond this extension, we would like to relax the Gaussian distribution assumption (e.g., to rotations of general product distributions), and to also handle larger families of sparse polynomials over highly-correlated features.

Acknowledgments

We acknowledge support from the Simons Foundation (#491119 to Alexandr Andoni), a Sloan Research Fellowship (Daniel Hsu), a Google Research Award (Alexandr Andoni), NSF awards CCF-1617955 and CCF-1740833, and an NSF Graduate Research Fellowship (Kiran Vodrahalli).

References

Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions*, chapter 6, sec. 4, pages 260–261. Dover Publications, New York, 1964. ISBN 978-0-486-61272-0.

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning sparse polynomial functions. In *Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 500–510, Philadelphia, PA, USA, 2014. Society for Industrial and Applied Mathematics. ISBN 978-1-611973-38-9.

Ravindra B. Bapat and Vaikalathur S. Sunder. On majorization and schur products. *Linear Algebra and its Applications*, 72:107 – 117, 1985. ISSN 0024-3795.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

Avrim Blum. On-line algorithms in machine learning. In *Online algorithms*, pages 306–325. Springer, 1998.

Shaobing Chen and David Donoho. Basis pursuit. In *28th Asilomar conf. Signals, Systems Computers*, 1994.

Sam F. Edwards and Phillip W. Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996. ISBN 0-8018-5414-8.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015. ISBN 1498712169, 9781498712163.

David Helmbold, Robert Sloan, and Manfred K Warmuth. Learning integer lattices. *SIAM Journal on Computing*, 21(2):240–266, 1992.

Adam Tauman Kalai, Alex Samorodnitsky, and Shang-Hua Teng. Learning and smoothed analysis. In *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*, pages 395–404. IEEE, 2009.

Oliver N Keene. The log transformation is special. *Statistics in medicine*, 14(8):811–819, 1995.

Adam R Klivans and Rocco A Servedio. Toward attribute efficient learning of decision lists and parities. *Journal of Machine Learning Research*, 7(Apr):587–602, 2006.

Murat Kocaoglu, Karthikeyan Shanmugam, Alexandros G Dimakis, and Adam Klivans. Sparse polynomial learning and graph sketching. In *Advances in Neural Information Processing Systems*, pages 3122–3130, 2014.

Arun K. Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *ArXiv e-prints*, 2018.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Sahand Negahban and Devavrat Shah. Learning sparse boolean polynomials. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 2032–2036. IEEE, 2012.

Ryan O'Donnell. *Analysis of Boolean Functions*, chapter 11 sec. 2, 8, pages 334–338, 368–385. Cambridge University Press, New York, NY, USA, 2014. ISBN 1107038324, 9781107038325.

Minh Ha Quang. *Reproducing Kernel Hilbert Spaces in Learning Theory*. PhD thesis, Brown University, 2006.

Herbert Robbins. A remark on stirling's formula. *The American Mathematical Monthly*, 62(1): 26–29, 1955. ISSN 00029890, 19300972.

Jonathan Sondow. An antisymmetric formula for euler's constant. *Mathematics Magazine*, 71: 219–220, June 1998.

Elias M. Stein and Rami Shakarchi. *Complex Analysis*, chapter 6, 7, pages 159–204. Princeton University Press, Princeton, New Jersey, 2003. ISBN 9781400831159.

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

Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem. *Journal of the ACM (JACM)*, 62(2):13, 2015.

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

Andreas Winkelbauer. Moments and absolute moments of the normal distribution. *ArXiv e-prints*, 2014.

Appendix A. Proof of Proposition 3.

Proposition 14 *A unique solution $\hat{\beta}$ to the transformed model in Eq. (2)) is the unique solution to the original model in Eq. (1).*

Proof We proceed by reversing each step of the transformation and demonstrating that the solutions do not change. First, note that the logarithm is invertible over the positive reals, which allows us to undo the log transformation without any effect, since absolute value ensures the domain is non-negative. Thus only consider the modified problem using data $\hat{y}_i = |y_i|, \hat{x}_{i,j} = |x_{i,j}|$. Since absolute value distributes under multiplication, $\hat{y}_i = |y_i| = |\prod_{j\in S} x_{i,j}^{\beta_j}| = \prod_{j\in S} |x_{i,j}|^{\beta_j} = \prod_{j\in S} \hat{x}_{i,j}^{\beta_j}$ and the resulting data points $(\hat{x}_i, \hat{y}_i)$ still satisfy the monomial model. Thus, if there is a unique solution on the transformed data $(\log|x_i|, \log|y_i|)$, it must also be the unique solution on all of the data. ■

Appendix B. Supporting Lemmas and Proof for Lemma 9

B.1. Hermite polynomials

We introduce some basic theory about the Hermite polynomial basis which will be useful in the analysis. We take the definition and required basic facts from O'Donnell (2014).

Definition 15 (Hermite orthogonal polynomial basis) *The Hermite basis is an orthogonal basis over $L_2(\mathcal{N}(0,1))$. In particular, we can write*

$$f(a) = \sum_{\ell=0}^{\infty} c_\ell H_\ell(a)$$

where $H_\ell(a)$ is the ℓ^{th} Hermite basis function, and $c_\ell = \mathbb{E}_a[f(a)H_\ell(a)]$. We define

$$H_0(a) = 1, \quad H_1(a) = a,$$

and compute the rest by applying Gram-Schmidt over the function space. We have the following definition for the ℓ^{th} Hermite basis function:

$$H_\ell(a) := \frac{1}{\sqrt{\ell!}} \frac{(-1)^\ell}{\varphi(a)} \frac{d^\ell}{da^\ell} \varphi(a).$$

We also have the recurrence relation

$$H_{\ell+1}(a) = \frac{1}{\sqrt{\ell+1}} \left(aH_\ell(a) - \frac{d}{da} H_\ell(a) \right)$$

and derivative formula

$$\frac{d}{da} H_\ell(a) = \sqrt{\ell} H_{\ell-1}(a).$$

The following important lemma provides a rule for calculating $\mathbb{E}_{a,a'}[f(a)f(a')]$ when a, a' are correlated Gaussian random variables.

Lemma 16 *Let a, a' be standard Gaussian random variables with correlation ρ. Then, we have*

$$\mathbb{E}_{a,a'}[H_\ell(a)H_{\ell'}(a')] = \begin{cases} \rho^\ell & \text{if } \ell = \ell' \\ 0 & \text{otherwise.} \end{cases}$$

and

$$\begin{aligned}\mathbb{E}_{a,a'}[f(a)f(a')] &= \sum_{\ell,\ell'=1}^{\infty} c_\ell c_{\ell'} \mathbb{E}_{a,a'}[H_\ell(a)H_{\ell'}(a')] \\ &= \sum_{\ell=0}^{\infty} c_\ell^2 \rho^\ell.\end{aligned}$$

B.2. Calculating the First and Second Log-Moments

In order to use the ideas from Section B.1, we first need to show that the function $\log|\cdot| \in L_2(\mathcal{N}(0,1))$, e.g., that $\mathbb{E}_{w\sim\mathcal{N}(0,1)}[\log^2|w|] = \alpha < \infty$. Along the way, it will also be useful to calculate and record $\mathbb{E}_{w\sim\mathcal{N}(0,1)}[\log|w|] = \tau < \infty$. In order to directly calculate these quantities, we use an idea from statistical physics called the replica trick (Edwards and Anderson (1975)). The idea is to note that $\frac{d}{d\nu}\mathbb{E}[a^\nu] = \mathbb{E}[\frac{d}{d\nu}a^\nu] = \mathbb{E}[(\log a)a^\nu]$. In general, if one takes m derivatives, the result will be $\frac{d^m}{d\nu^m}\mathbb{E}[a^\nu] = \mathbb{E}[a^\nu \log^m(a)]$. Then, taking the limit as $\nu \to 0$ yields

Lemma 17 (Replica trick) *Let a be a non-negative random variable,*

$$\mathbb{E}_a[\log^m a] = \lim_{\nu\to 0} \frac{d^m}{d\nu^m}\mathbb{E}_a[a^\nu] \tag{5}$$

We refer to the LHS expression in Eq. (5) as the m^{th} *log-moment* of a. Thus, as long as we can get an analytic expression for $\mathbb{E}_a[a^\nu]$ which is valid for $\nu \in \mathbb{R}^+$, we can take the continuous limit and derive expressions for the first and second log-moments of a, where $a = |w|, w \sim \mathcal{N}(0,1)$. We also note that in the upcoming discussion, γ refers to the *Euler-Mascheroni* constant.

We will apply the replica trick to calculate the first two log-moments. We first need to collect some lemmas from the literature.

Lemma 18 (Moments of Absolute Gaussian Distribution) *We have for $\nu \in \mathbb{R}^+$ with $w \sim \mathcal{N}(0,1)$*

$$\mathbb{E}_w[|w|^\nu] = \frac{1}{\sqrt{\pi}} 2^{\nu/2} \Gamma\left(\frac{\nu+1}{2}\right)$$

where Γ is the gamma function.

Proof For derivation, see Winkelbauer (2014). ■

We will need some properties of the gamma function, several of which depend on the polygamma function ψ. We take these facts from Stein and Shakarchi (2003), Abramowitz and Stegun (1964), and Sondow (1998).

Definition 19 (Polygamma function) *The* polygamma function of order 0 *is defined by*

$$\psi(x) := \frac{d}{dx}\log(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

The polygamma function of order $i \geq 1$ *is defined by*

$$\psi^{(i)}(x) := \frac{d^i}{dx^i}\psi(x).$$

Lemma 20 (Properties of the Gamma and Polygamma functions) *The derivative of* $\Gamma(x)$ *is given by*

$$\frac{d}{dx}\Gamma(x) = \Gamma(x)\psi(x).$$

The Taylor series expansions for $\psi^{(i)}(1+x)$ *are*

$$\psi(x+1) = -\gamma + \sum_{j=1}^{\infty}(-1)^{j+1}\zeta(j+1)x^j$$
$$\psi^{(i)}(x+1) = \sum_{j=0}^{\infty}(-1)^{i+j+1}\frac{(i+j)!}{j!}\zeta(i+j+1)x^j \text{ for } i \geq 1$$

(convergence is for $|x| < 1$*). Above,* γ *is the Euler-Mascheroni constant and* $\zeta(s) = \sum_{n=1}^{\infty}\frac{1}{n^s}$ *is the zeta function.*

The first identity above follows from Abramowitz and Stegun (1964) and the antisymmetric formula for γ given in Sondow (1998).

Now, we can use these facts to calculate the first two log-moments of $|w|$.

Lemma 21 (First log-moment (τ))

$$\tau := \mathbb{E}_w[\log|w|] = -\frac{1}{2}\left(\log(2) + \gamma\right) \approx -0.635$$

We also record that $\tau^2 \approx 0.403$.

Proof We apply the replica trick to get

$$\begin{aligned}\mathbb{E}_w[\log|w|] &= \lim_{\nu\to 0}\frac{d}{d\nu}\frac{1}{\sqrt{\pi}}2^{\nu/2}\Gamma(\frac{\nu+1}{2})\\ &= \frac{1}{2}\log(2) + \frac{1}{2\sqrt{\pi}}\lim_{\nu\to 0}\Gamma'\left(\frac{1+\nu}{2}\right)\sqrt{2}^{\nu}\\ &= \frac{1}{2}\log(2) + \frac{1}{2\sqrt{\pi}}\left(1\cdot\sqrt{\pi}\cdot\lim_{\nu\to 0}\psi\left(\frac{1+\nu}{2}\right)\right)\\ &= \frac{1}{2}\left(\log(2) + \psi\left(\frac{1}{2}\right)\right)\end{aligned}$$

where we used the derivative of Γ and the fact that the limit existed individually for each term in the second product. Applying the Taylor expansion of $\psi(x+1)$ and plugging in $x=-1/2$, we get using properties of infinite geometric series that

$$\begin{aligned}
\psi(1/2) &= -\gamma + \sum_{j=1}^{\infty}(-1)^{j+1}\zeta(j+1)(-1/2)^j \\
&= -\left(\gamma + \sum_{j=1}^{\infty}\frac{1}{2^j}\zeta(j+1)\right) \\
&= -\left(\gamma + \sum_{j,n=1}^{\infty}\frac{1}{2^j}\frac{1}{n^{j+1}}\right) \\
&= -\left(\gamma + \sum_{n=1}^{\infty}\frac{1}{n}\sum_{j=1}^{\infty}\frac{1}{(2n)^j}\right) \\
&= -\left(\gamma + \sum_{n=1}^{\infty}\frac{1}{n}\left(\frac{1/2n}{1-1/2n}\right)\right) \\
&= -\left(\gamma + \sum_{n=1}^{\infty}\frac{1}{n(2n-1)}\right)
\end{aligned}$$

Then, consider the Taylor series for $\log(x)$ centered at $x=1$, which has radius of convergence $|x-1|\le 1$. We have, plugging in $x=2$,

$$\begin{aligned}
\log(x) &= \sum_{n=1}^{\infty}(-1)^{n+1}\frac{(x-1)^n}{n} \\
\log(2) &= \sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n} \\
&= (1-\frac{1}{2}) + (\frac{1}{3}-\frac{1}{4}) + (\frac{1}{5}-\frac{1}{6}) + \cdots \\
&= \frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{n(2n-1)}.
\end{aligned}$$

Thus, we conclude that $\psi(1/2) = -(\gamma + 2\log(2))$, and overall that

$$\tau = \frac{1}{2}\big(\log(2) - \gamma - 2\log(2)\big) = -\frac{1}{2}\big(\log(2)+\gamma\big).$$

■

Lemma 22 (Second log-moment (α))

$$\alpha := \mathbb{E}_w[\log^2|w|] = \frac{1}{4}\left(\gamma^2 + \frac{\pi^2}{2} + \log^2(2) + \gamma\log(4)\right) \approx 1.637.$$

We also record that $\alpha - \tau^2 \approx 1.234$.

Proof We calculate the second derivative with respect to ν and evaluate it at $\nu = 0$, using the product rule and the derivatives of Γ and ψ:

$$
\begin{aligned}
\mathbb{E}_w[\log^2 |w|] &= \frac{d^2}{d\nu^2}\mathbb{E}_w[|w|^\nu]|_{\nu=0} \\
&= \frac{1}{2\sqrt{\pi}}\Big[\log(2) \left(\log(\sqrt{2}) e^{\nu \log(\sqrt{2})} \Gamma\left(\frac{1+\nu}{2}\right) + \frac{1}{2} e^{\nu \log(\sqrt{2})} \Gamma\left(\frac{1+\nu}{2}\right) \psi\left(\frac{1+\nu}{2}\right)\right) \\
&+ \psi\left(\frac{1+\nu}{2}\right) \frac{d}{d\nu}\left(\sqrt{2}^\nu \Gamma\left(\frac{1+\nu}{2}\right)\right) + \frac{1}{2}\sqrt{2}^\nu \Gamma\left(\frac{1+\nu}{2}\right) \psi^{(1)}\left(\frac{1+\nu}{2}\right)\Big]\Big|_{\nu=0} \\
&= \frac{1}{4}\left[\left(\log^2(2) - 2\log(2)(\gamma + \log(4)) + (\gamma + \log(4))^2 \right) + \left[\psi^{(1)}\left(\frac{1+\nu}{2}\right)\right]\Big|_{\nu=0} \right]
\end{aligned}
$$

where we used the results from Lemma B.2 to simplify, keeping in mind that we will shortly show that $\psi^{(1)}(1/2)$ exists. We use the Taylor series for $\psi^{(1)}(x+1)$ which converges for $|x| < 1$ and plug in $x = -1/2$:

$$
\begin{aligned}
\psi^{(1)}(x+1) &= \sum_{j=0}^{\infty} (-1)^{j+2} \frac{(j+1)!}{j!} \zeta(j+2) x^j \\
\psi^{(1)}(1/2) &= \sum_{j=0}^{\infty} \frac{j+1}{2^j} \sum_{n=1}^{\infty} \frac{1}{n^{j+2}} \\
&= 2 \sum_{n=1}^{\infty} \frac{1}{n} \sum_{j=1}^{\infty} j \left(\frac{1}{2n}\right)^j \\
&= 4 \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \\
&= 4 \left(\sum_{n=1}^{\infty} \frac{1}{n^2} - \sum_{n=1}^{\infty} \frac{1}{(2n)^2} \right) = 4\left(\frac{\pi^2}{6} - \frac{\pi^2}{24}\right) = \frac{\pi^2}{2}
\end{aligned}
$$

recalling that $\sum_{j=1}^{\infty} jc^j = \frac{c}{1-c} + \frac{c^2}{1-c} + \cdots = \frac{c}{(1-c)^2}$ and the fact that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$. Plugging this value in to our previous formula, we conclude

$$
\begin{aligned}
\alpha &= \frac{1}{4}\left[\log^2(2) - 2\log(2)(\gamma + \log(4)) + (\gamma + \log(4))^2 + \frac{\pi^2}{2} \right] \\
&= \frac{1}{4}\left[\log^2(2) + \frac{\pi^2}{2} - 2\gamma\log(2) - \log^2(4) + \gamma^2 + 2\gamma\log(4) + \log^2(4) \right] \\
&= \frac{1}{4}\left[\gamma^2 + \log^2(2) + \frac{\pi^2}{2} + \gamma\log(4) \right].
\end{aligned}
$$

■

The next lemma will be useful in the next section of the appendix, and uses similar ideas.

Lemma 23

$$\mathbb{E}_w[w^2 \log(|w|)] = 1 + \tau.$$

Proof We have by integration by parts and the fact that $\varphi'(w) = -w\varphi(w)$

$$\begin{aligned}
\int w^2 \log(|w|)\varphi(w)dw\bigg|_{\mathbb{R}} &= \int w \log(|w|)w\varphi(w)dw\bigg|_{\mathbb{R}} \\
&= -\int w \log(|w|)\frac{d}{dw}\varphi(w)dw\bigg|_{\mathbb{R}} \\
&= -\left(w \log(|w|)\varphi(w) - \int (1 + \log(|w|))\varphi(w)dw\right)\bigg|_{\mathbb{R}} \\
&= 0 + 1 + \int \log(|w|)\varphi(w)dw\bigg|_{\mathbb{R}} \\
&= 1 + \tau
\end{aligned}$$

since $\lim_{w\to\pm\infty} w \log(|w|)\varphi(w) = 0$. ■

B.3. Coefficients of the Hermite Expansion of $\log(|\cdot|)$

Lemma 24 (Coefficients of the Hermite Expansion for $\log(|\cdot|)$) *The Hermite expansion of* $\log\big(|\cdot|\big)$

$$\log(a) = \sum_{\ell=0}^{\infty} c_\ell H_\ell(a)$$

has $c_0 = \tau$, and for $\ell \geq 1$,

$$c_{2\ell-1} = 0, \quad c_{2\ell} = \frac{(-1)^{\ell-1}2^{\ell-1}(\ell-1)!}{\sqrt{(2\ell)!}}.$$

Moreover,

$$\lim_{\ell\to\infty} c_{2\ell}^2 \cdot \ell^{3/2} = \frac{\sqrt{\pi}}{4},$$

and for $\ell \geq 2$,

$$c_{2\ell}^2 \geq \frac{1}{5} \cdot \frac{1}{\ell^{3/2}}.$$

Proof Our goal is to calculate $\mathbb{E}_w[H_\ell(w)\log(|w|)]$. Recall that $\varphi(w)$ is the standard Gaussian density. We proceed by making use of several properties of Hermite polynomials from Section

B.1 and applying integration by parts. First define the indefinite integral and apply the property $H'_{i+1}(w) = \sqrt{i+1}H_i(w)$:

$$\begin{aligned}
A_i &= \int \log(|w|)H_i(w)\varphi(w)dw \\
&= \frac{1}{\sqrt{i+1}} \int H'_{i+1}(w) \log(|w|)\varphi(w)dw \\
&= \frac{1}{\sqrt{i+1}} \left(H_{i+1}(w) \log(|w|)\varphi(w) - \int H_{i+1}(w) \left(\frac{1}{w} - w \log(|w|) \right) \varphi(w)dw \right)
\end{aligned}$$

where we used the fact $\frac{d}{dw} \log(|w|)\varphi(w) = \left(\frac{1}{w} - w \log(|w|) \right) \varphi(w)$. Let $V_i(w) = H_i(w) \log(|w|)\varphi(w)$. Then, applying the relation $wH_i(w) = H'_i(w) + \sqrt{i+1}H_{i+1}(w)$, we get

$$\begin{aligned}
A_i = \frac{1}{\sqrt{i+1}} &\left(V_{i+1}(w) + \int \left(H'_{i+1}(w) + \sqrt{i+2}H_{i+2}(w) \right) \log(|w|)\varphi(w)dw \right. \\
&\qquad \left. - \int \frac{1}{\sqrt{i+1}} \frac{wH_i(w) - H'_i(w)}{w} \varphi(w)dw \right) \\
= \frac{1}{\sqrt{i+1}} &\left(V_{i+1}(w) + \sqrt{i+2}A_{i+2} + \sqrt{i+1}A_i - \frac{1}{\sqrt{i+1}} \int \left(H_i(w) - \frac{1}{w}H'_i(w) \right) \varphi(w)dw \right).
\end{aligned}$$

Assuming that $i > 0$, orthogonality implies $\int_{\mathbb{R}} H_i(w)\varphi(w)dw = 0$, and we cancel it out now (since eventually we will evaluate everything over $\mathbb{R}$). We simplify the equation to

$$V_{i+1}(w) + \sqrt{i+2}A_{i+2} + \sqrt{\frac{i}{i+1}} \int \frac{1}{w} H_{i-1}(w)\varphi(w)dw = 0.$$

Then we calculate $\frac{d}{dw}H_{i-1}(w)\varphi(w) = \varphi(w) \left(\sqrt{i-1}H_{i-2}(w) - wH_{i-1}(w) \right)$ and apply integration by parts to the last integral to get

$$\begin{aligned}
&\int \frac{1}{w} H_{i-1}(w)\varphi(w)dw \\
&= V_i(w) - \int \left(\sqrt{i-1} \log(|w|)H_{i-2}(w) - \log(|w|) \left(wH_{i-1}(w) \right) \right) \varphi(w)dw \\
&= V_i(w) - \sqrt{i-1}A_{i-2} + \int \log(|w|) \left(H'_{i-1}(w) + \sqrt{i}H_i(w) \right) \varphi(w)dw \\
&= V_i(w) - \sqrt{i-1}A_{i-2} + \sqrt{i-1}A_{i-2} + \sqrt{i}A_i \\
&= V_i(w) + \sqrt{i}A_i.
\end{aligned}$$

Plugging this equality back in and then evaluating the integrals on $\mathbb{R}$ yields

$$\begin{aligned}
\left[V_{i+1}(w) + \sqrt{\frac{i}{i+1}} V_i(w) \right] \Bigg|_{\mathbb{R}} + \sqrt{i+2}A_{i+2} \Bigg|_{\mathbb{R}} + \frac{i}{\sqrt{i+1}} A_i \Bigg|_{\mathbb{R}} &= 0, \\
\sqrt{i+2}c_{i+2} &= -\frac{i}{\sqrt{i+1}} c_i, \\
\frac{-i}{\sqrt{(i+1)(i+2)}} c_i &= c_{i+2} \qquad (6)
\end{aligned}$$

since $\lim_{w\to\pm\infty} V_i(w) = 0$ for any i, as $\varphi(w)$ decays much faster than $\log(|w|)\cdot \text{poly}(w)$ grows. Note that this recurrence is only valid for $i > 0$, since we used that in the analysis. Now, recall that by definition, $c_0 = \tau$ since $H_0(w) = 1$. Furthermore, since $H_1(w) = w$, $c_1 = \mathbb{E}_w[w\log(|w|)] = 0$ since $w\log(|w|)$ is an odd function and the Gaussian distribution is symmetric. Then, we can calculate $H_2(w) = \frac{1}{\sqrt{2}}(x^2-1)$ and thus that

$$\begin{aligned} c_2 &= \frac{1}{\sqrt{2}}\left(\mathbb{E}_w[w^2\log(|w|)] - \mathbb{E}_w[\log(|w|)]\right) \\ &= \frac{1}{\sqrt{2}}\left(1+\tau-\tau\right) = \frac{1}{\sqrt{2}} \end{aligned}$$

using Lemma 23. The rest of the coefficients are defined recursively by Eq. (6). In particular, we can find a closed form. First, note that since $c_1 = 0$, $c_{2n-1} = 0$ for all strictly positive integers n. Iterating Eq. (6) gives

$$c_{2n} = \frac{(-1)^{n-1}2^{n-1}(n-1)!}{\sqrt{(2n)!}}.$$

Now, we can apply the well-known Stirling's approximation $\left(n! \asymp \sqrt{2\pi n}\,(n/e)^n\right)$ to get the asymptotic behavior of this quantity. We have

$$\begin{aligned} c_{2n} &\asymp (-1)^{n-1}2^{n-1}\sqrt{\frac{\sqrt{\pi}(n-1)}{\sqrt{n}}\frac{e}{n-1}\left(\frac{n-1}{n}\right)^n 2^{-n}} \\ &= (-1)^{n-1}\frac{\pi^{1/4}}{2}\left((n-1)\sqrt{n}\right)^{-1/2} e\left(1-\frac{1}{n}\right)^n \\ &\asymp (-1)^{n-1}\frac{\pi^{1/4}}{2}n^{-3/4} \end{aligned}$$

after noting that $\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n = e^{-1}$. Therefore, the behavior of c_{2n}^2 is given by

$$c_{2n}^2 \asymp \frac{\sqrt{\pi}}{4}\cdot\frac{1}{n^{3/2}}.$$

We note that this asymptotic behavior is quite tight, even up to constants, for sufficiently large n. We can also prove a fairly tight lower bound using Robbins (1955), which gives the bound

$$\sqrt{2\pi n}n^n e^{-n} e^{1/(12n+1)} \le n! \le \sqrt{2\pi n}n^n e^{-n} e^{1/12n}$$

Plugging in the upper bound for $(n+1)!$ and the lower bound for $(2n)!$, we get that

$$\begin{aligned} 2^{n-1}\frac{(n-1)!}{\sqrt{(2n)!}} &\ge \frac{e\pi^{1/4}}{2}\sqrt{\frac{1}{n\sqrt{n}}\left(\frac{n-1}{n}\right)^n e^{\frac{36n+11}{48n(12n-11)}}} \\ &\ge \frac{e\pi^{1/4}}{8}\cdot 1\cdot n^{-3/4} \end{aligned}$$

for $n \geq 2$. Thus, for $n \geq 2$,

$$c_{2n}^2 \geq \frac{e^2\sqrt{\pi}}{64} n^{-3/2} > \frac{1}{5} n^{-3/2}.$$

$\blacksquare$

Lemma 25 (Integrals of the Hermite Coefficients) *Suppose* $|b| < 1$ *and* $a > 0$. *Then*

$$\int_a^\infty \ell^{-3/2} d\ell = \frac{2}{\sqrt{a}}$$

and

$$\int_a^\infty \ell^{-3/2} |b|^{2\ell} d\ell = \frac{2|b|^{2a}}{\sqrt{a}} + 2\sqrt{2\pi \log |b|^{-1}} \left(-1 + \mathrm{Erf}\left(\sqrt{a \log |b|^{-2}} \right) \right)$$

where

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

We also have the following upper bound:

$$\int_a^\infty \ell^{-3/2} |b|^{2\ell} d\ell \leq \frac{2|b|^{2a}}{\sqrt{a}} - 4\sqrt{2 \log |b|^{-1}} \frac{e^{2a \log |b|}}{\sqrt{\log |b|^{-2a}} + \sqrt{2 + \log |b|^{-2a}}}.$$

Proof The first equality is by direct integration. Now we tackle the second equality. We apply integration by parts to get

$$\begin{aligned}
\int \ell^{-3/2} |b|^{2\ell} d\ell &= |b|^{2\ell}(-2\ell^{-1/2}) + \int 2\ell^{-1/2} \cdot (2\log(|b|)) |b|^{2\ell} d\ell \\
&= |b|^{2\ell}(-2\ell^{-1/2}) + 4\log(|b|) \int |b|^{2u^2} 2du \\
&= |b|^{2\ell}(-2\ell^{-1/2}) + 8\log(|b|) \int e^{-u^2/\left(\frac{1}{2}\frac{1}{\log(|b|^{-1})}\right)} du \\
&= |b|^{2\ell}(-2\ell^{-1/2}) + \frac{-8\log(|b|^{-1})}{\sqrt{2\log(|b|^{-1})}} \frac{\sqrt{\pi}}{2} \frac{2}{\sqrt{\pi}} \int e^{-v^2} dv \\
&= -2\left(|b|^{2\ell}(\ell^{-1/2}) + \sqrt{2\pi \log(|b|^{-1})} \mathrm{Erf}\left(\sqrt{2\ell \log(|b|^{-1})} \right) \right).
\end{aligned}$$

Then since $\mathrm{Erf}(\infty) = 1$, we simply evaluate the integral and note that the Erf term (depending on a) is positive:

$$\int_a^\infty \ell^{-3/2} |b|^{2\ell} d\ell = -2\sqrt{2\pi \log(|b|^{-1})} + 2\left(|b|^{2a} a^{-1/2} + \sqrt{2\pi \log(|b|^{-1})} \mathrm{Erf}\left(\sqrt{2a \log(|b|^{-1})} \right) \right).$$

Section 7.1.13 of Abramowitz and Stegun (1964) gives

$$\frac{1}{x+\sqrt{x^2+2}} < e^{x^2}\int_x^\infty e^{-t^2}dt \leq \frac{1}{x+\sqrt{x^2+\frac{4}{\pi}}},$$

$$1-\frac{2}{\sqrt{\pi}}\frac{e^{-x^2}}{x+\sqrt{x^2+\frac{4}{\pi}}} \leq \operatorname{Erf}(x) < 1-\frac{2}{\sqrt{\pi}}\frac{e^{-x^2}}{x+\sqrt{x^2+2}}.$$

We can apply the upper bound on Erf to in order to get the final upper bound on the integral. ■

Appendix C. Supporting Lemmas and Proofs for Theorem 10

C.1. Matrix Inequalities and Hadamard Powers of Matrices

In this section, we record some useful definitions and theorems about matrices and their Hadamard powers.

Theorem 26 (Gershgorin Circle Theorem) *For matrix $A \in \mathbb{R}^{p\times p}$, every eigenvalue $\lambda(A)$ satisfies*

$$\lambda(A) \geq A_{ii} - \sum_{i\neq j}|A_{ij}|.$$

In particular,

$$\lambda_{\min}(A) \geq \min_i |A_{ii}| - (p-1)\max_{i\neq j}|A_{ij}|.$$

Proof See Golub and Van Loan (1996). ■

Definition 27 (Hadamard Product and Power) *The Hadamard product of matrices A, B is given by*

$$[A\circ B]_{i,j} = A_{i,j}B_{i,j}.$$

The m^{th} Hadamard power of A is given by

$$A^{(m)} = \underbrace{A\circ A\circ\cdots\circ A}_{m \text{ times}}.$$

Theorem 28 (Schur Product Theorem (weak version)) *Suppose A, B are both symmetric PSD square matrices. Then $A\circ B$ is also PSD.*

Proof Write eigendecompositions of $A = \sum_i \mu_i a_i a_i^T$ and $B = \sum_i \nu_i b_i b_i^T$. Then

$$\begin{aligned} A\circ B &= \sum_{i,j}\mu_i\nu_j\left(a_i a_i^T\right)\circ\left(b_j b_j^T\right) \\ &= \sum_{i,j}\mu_i\nu_j\left(a_i\circ b_j\right)\left(a_i\circ b_j\right)^T \end{aligned} \tag{7}$$

Then we have that $\mu_i, \nu_i \geq 0$ and $\left(a_i\circ b_j\right)\left(a_i\circ b_j\right)^T$ is PSD. Thus $A\circ B$ is PSD. ■

Theorem 29 (Eigenvalues of Hadamard Powers) *Suppose $A, B \in \mathbb{R}^{p\times p}$ both PSD. Let b denote B's diagonal. Then*

$$\prod_{i=j}^{p} \lambda_i(A \circ B) \geq \prod_{i=j}^{p} \lambda_i(A) b_i \tag{8}$$

for all $j \in [p]$, where λ_i is the i^{th} smallest eigenvalue.

Proof See Theorem 3 from Bapat and Sunder (1985). ■

C.2. Proof of Theorem 10

C.2.1. Lower Bounding the Population Minimal Eigenvalue

As a warm-up, we first prove prove a lower bound on $\lambda_{\min}(\Sigma)$.

Theorem 30 (Minimum Eigenvalue of Population Correlation Matrix) *The following lower bounds on $\lambda_{\min}(\Sigma)$ hold:*

1. $\lambda_{\min}(\Sigma) \geq \frac{\pi^2}{8} \lambda_{\min}(\Phi)$.
2. $\lambda_{\min}(\Sigma) \geq \displaystyle\sum_{\ell=1}^{\frac{1}{2}\left\lfloor \frac{\log(p-1)}{\log(\frac{1}{1-\epsilon})} \right\rfloor} \frac{\lambda_{\min}\left(\Phi^{(2\ell)}\right)}{5\ell^{3/2}} + \frac{2}{5}\sqrt{\frac{2\log((1-\epsilon)^{-1})}{\log\left(\frac{p-1}{1-\epsilon}\right) + \max\{2, \log((1-\epsilon)^{-1})\}}}$.

Note that the first lower bound is positive whenever Φ is full-rank, and the second bound is always strictly positive, even if Φ is low-rank.

Remark 31 (Intuition for Theorem 10) *In the case that Φ is not low-rank, we automatically have a constant multiplicative factor improvement on the minimum eigenvalue when applying the log transformation. However, the true magic happens in the second bound – even if $\lambda_{\min}(\Phi) = 0$, we can still achieve a positive lower bound on $\lambda_{\min}\left(\mathbb{E}_z[zz^T]\right)$. The intuitive reason this phenomenon occurs is because the Hadamard power destroys the potential co-linear structure in Φ – this is precisely how the nonlinearity of the logarithm comes into play.*

Proof For ease of notation, through out this proof, we define $|\rho_{\max}| = 1 - \epsilon$. We recall that $\Sigma = \mathbb{E}[zz^T]$ where $z = \log|x|, x \sim \mathcal{N}(0, \Phi)$. By Lemma 16, we have

$$\Sigma_{i,j} = \sum_{\ell=0}^{\infty} c_\ell^2 \Phi_{i,j}^\ell$$

where $c_\ell = \mathbb{E}_w[H_\ell(w)\log(|w|)]$. This means,

$$\Sigma = \sum_{\ell=0}^{\infty} c_\ell^2 \Phi^{(\ell)}.$$

Continuing with the proof, we have

$$\begin{aligned}\lambda_{\min}\left(\mathbb{E}_z[zz^T]\right) &= \min_{\|v\|_2=1} \sum_{i,j=1}^{p} v_i v_j \sum_{\ell=0}^{\infty} c_\ell^2 \Phi_{i,j}^{\ell} \\ &= \min_{\|v\|_2=1} \sum_{\ell=0}^{\infty} c_\ell^2 \sum_{i,j=1}^{p} v_i v_j \Phi_{ij}^{\ell} \\ &\geq \sum_{\ell=0}^{\infty} c_\ell^2 \left(\min_{\|v\|_2=1} \sum_{i,j=1}^{p} v_i v_j \Phi_{i,j}^{\ell} \right) \qquad (9) \\ &= \sum_{\ell=0}^{\infty} c_\ell^2 \lambda_{\min}\left(\Phi^{(\ell)}\right) \\ &= c_0^2 \overbrace{\lambda_{\min}\left(\mathbf{1}_{p\times p}\right)}^{0} + \sum_{\ell=1}^{\infty} c_\ell^2 \lambda_{\min}\left(\Phi^{(\ell)}\right) \\ &= \sum_{\ell=1}^{\infty} c_\ell^2 \lambda_{\min}\left(\Phi^{(\ell)}\right)\end{aligned}$$

where $\Phi^{(\ell)}$ denotes the ℓ^{th} element-wise (Hadamard) power of Φ (see Definition 27). Then, using Theorem 29, we have that $\lambda_{\min}(A \circ B) \geq \lambda_{\min}(A) \cdot B_{p,p}$, where $\circ$ denotes Hadamard product. Therefore, since the diagonal entries are all 1 and $1^\ell = 1$, we have for all $\ell \geq 1$ that

$$\lambda_{\min}\left(\Phi^{(\ell)}\right) \geq \lambda_{\min}\left(\Phi\right) \cdot 1$$

and we immediately get the bound

$$\lambda_{\min}\left(\mathbb{E}_z[zz^T]\right) \geq \sum_{\ell=1}^{\infty} c_\ell^2 \lambda_{\min}\left(\Phi\right) = \lambda_{\min}(\Phi) \sum_{\ell=1}^{\infty} c_\ell^2 = \left(\alpha - \tau^2\right) \lambda_{\min}(\Phi) = \frac{\pi^2}{8} \lambda_{\min}(\Phi).$$

However, this bound can be greatly improved by judiciously applying the well-known Gershgorin circle theorem (Theorem 26). In order to apply this bound, we need to ensure that the Gershgorin bound will be strictly positive. Therefore, we truncate the summation carefully. Define

$$\ell_{\text{threshold}} = 1 + \left\lceil \frac{\log(p-1)}{\log(1/|\rho_{\max}|)} \right\rceil.$$

Note that for $\ell \geq \ell_{\text{threshold}}$, we have

$$\begin{aligned}(p-1)|\rho_{\max}|^{\ell} &\leq (p-1)|\rho_{\max}|^{1+\left\lceil \frac{\log(p-1)}{\log(1/|\rho_{\max}|)} \right\rceil} \\ &\leq \frac{|\rho_{\max}|(p-1)}{p-1} < 1\end{aligned}$$

Applying Gershgorin to the truncated tail of the sum, we bound

$$\begin{aligned}\lambda_{\min}\left(\mathbb{E}_z[zz^T]\right) &\geq \sum_{\ell=1}^{\infty} c_\ell^2 \lambda_{\min}\left(\Phi^{(\ell)}\right) \\ &= \sum_{\ell=1}^{\ell_{\text{threshold}}-1} c_\ell^2 \lambda_{\min}\left(\Phi^{(\ell)}\right) + \sum_{\ell=\ell_{\text{threshold}}}^{\infty} c_\ell^2 \left(1-(p-1)|\rho_{\max}|^\ell\right).\end{aligned}$$

We know from Theorem 28 that taking the Hadamard power of a PSD matrix yields a PSD matrix, thus the first summation term is non-negative.

We can further control this bound by plugging in estimates for c_ℓ^2 from Lemma 24: Recall that we have $c_{2\ell}^2 \geq \frac{1}{5}\frac{1}{\ell^{3/2}}$ and $c_{2\ell-1}^2 = 0$. Then, supposing $\ell_{\text{threshold}}$ is even for simplicity, we can re-write our bound as

$$\begin{aligned}&\lambda_{\min}\left(\mathbb{E}_z[zz^T]\right) \\ &\geq \sum_{\ell=1}^{(\ell_{\text{threshold}}-2)/2} c_{2\ell}^2 \lambda_{\min}\left(\Phi^{(2\ell)}\right) + \frac{1}{5}\sum_{\ell=\ell_{\text{threshold}}/2}^{\infty} \ell^{-3/2}\left(1-(p-1)|\rho_{\max}|^{2\ell}\right) \\ &= \sum_{\ell=1}^{(\ell_{\text{threshold}}-2)/2} \frac{\lambda_{\min}\left(\Phi^{(2\ell)}\right)}{5\ell^{3/2}} + \frac{1}{5}\left(\sum_{\ell=\ell_{\text{threshold}}/2}^{\infty} \ell^{-3/2} - (p-1)\sum_{\ell=\ell_{\text{threshold}}/2}^{\infty} \ell^{-3/2}|\rho_{\max}|^{2\ell}\right).\end{aligned}$$

We now focus on lower bounding the second term further, letting

$$L = \sum_{\ell=1}^{(\ell_{\text{threshold}}-2)/2} \frac{\lambda_{\min}\left(\Phi^{(2\ell)}\right)}{5\ell^{3/2}}.$$

Recall that for a non-negative function f, we can upper and lower bound its summation as follows:

$$\int_a^\infty f(\ell)d\ell \leq \sum_{\ell=a}^{\infty} f(\ell) \leq f(\ell) + \sum_{\ell=a+1}^{\infty} f(\ell) \leq f(\ell) + \int_a^\infty f(\ell)d\ell.$$

Then, applying Lemma 25, and plugging in the integral bounds, we get

$$\begin{aligned}&\lambda_{\min}\left(\mathbb{E}_z[zz^T]\right) \\ &\geq L + \frac{1}{5}\left(\frac{2-2(p-1)|\rho_{\max}|^{\ell_{\text{threshold}}}}{\sqrt{\ell_{\text{threshold}}/2}} + \frac{4(p-1)\sqrt{2\log(|\rho_{\max}|^{-1})}\,e^{-\ell_{\text{threshold}}\log(|\rho_{\max}|^{-1})}}{\sqrt{\ell_{\text{threshold}}\log(|\rho_{\max}|^{-1})} + \sqrt{2+\ell_{\text{threshold}}\log(|\rho_{\max}|^{-1})}}\right) \\ &= L + \frac{2}{5}\left(\frac{1}{\sqrt{\ell_{\text{threshold}}/2}}(1-|\rho_{\max}|) + \frac{2(p-1)|\rho_{\max}|^{\ell_{\text{threshold}}}\sqrt{2\log(|\rho_{\max}|^{-1})}}{\sqrt{\log(|\rho_{\max}|^{-\ell_{\text{threshold}}})} + \sqrt{2+\log(|\rho_{\max}|^{-\ell_{\text{threshold}}})}}\right) \\ &> L + \frac{2}{5}\left((1-|\rho_{\max}|)\sqrt{\frac{2}{1+\left\lceil\frac{\log(p-1)}{\log(1/|\rho_{\max}|)}\right\rceil}} + |\rho_{\max}|\sqrt{\frac{2\log(|\rho_{\max}|^{-1})}{\log\left(\frac{p-1}{|\rho_{\max}|}\right)+2}}\right) \\ &> L + \frac{2}{5}\left((1-|\rho_{\max}|)\sqrt{\frac{2\log(|\rho_{\max}|^{-1})}{\log\left(\frac{p-1}{|\rho_{\max}|}\right)+\log(|\rho_{\max}|^{-1})}} + |\rho_{\max}|\sqrt{\frac{2\log(|\rho_{\max}|^{-1})}{\log\left(\frac{p-1}{|\rho_{\max}|}\right)+2}}\right)\end{aligned}$$

where we upper bounded $\lceil x \rceil \leq x + 1$. Then, we can simplify the expression to

$$\lambda_{\min}\left(\mathbb{E}_z[zz^T]\right) > L + \frac{2}{5}\sqrt{\frac{2\log(|\rho_{\max}|^{-1})}{\log\left(\frac{p-1}{|\rho_{\max}|}\right) + \max(2, \log(|\rho_{\max}|^{-1}))}}$$

where if $|\rho_{\max}| \geq e^{-2}$, we have that $\log(|\rho_{\max}|^{-1}) \leq 2$, which is the desired result. ■

C.2.2. Bounds that use Sparsity

In this section, we demonstrate bounds on the minimum eigenvalue which are independent of dimension p: instead, the sparsity k plays a role.

In order to fully take advantage of the sparsity assumption, we prove a restricted analogue to the Gershgorin circle theorem (Golub and Van Loan (1996)) we used previously.

Lemma 32 (Restricted Gershgorin Circle Theorem) *Let $A \in \mathbb{R}^{p \times p}$ be a symmetric matrix. Let $\alpha \geq 1$ and $T \subset [p]$. Then,*

$$\tilde{\lambda}(\alpha, T, A^{1/2}) \geq \min_i A_{ii} - |T| \cdot (1+\alpha)^2 \cdot \max_{i \neq j} |A_{ij}|.$$

Proof Given in Appendix E. ■

We can use these results directly to replace dimension p with sparsity k in the statements in Theorem 10. The proof is by direct application of Lemma 32.

Corollary 33 *The following lower bound holds:*

$$\tilde{\lambda}\left(k, \Sigma^{1/2}\right) \geq \sum_{\ell=1}^{\frac{1}{2}\left\lfloor \frac{\log(16k))}{\log(\frac{1}{1-\epsilon})} \right\rfloor} \frac{\tilde{\lambda}\left(k, [\Phi^{(2\ell)}]^{1/2}\right)}{5\ell^{3/2}} + \frac{2}{5}\sqrt{\frac{2\log((1-\epsilon)^{-1})}{\log\left(\frac{16k}{1-\epsilon}\right) + \max\{2, \log((1-\epsilon)^{-1})\}}} \tag{10}$$

This improvement is quite notable in that it completely removes dependence on dimension p. Potentially, the bound could be a lot better as typically $k \ll p$ in high-dimensional settings. This improvement is also valuable because it now shifts dependence on $\lambda_{\min}(\Phi)$ to dependence on $\tilde{\lambda}(k, \Phi^{1/2})$, which is potentially much larger and positive even in the case where $\lambda_{\min}(\Phi) = 0$.

Appendix D. Proof of Lemma 13 and Theorem 4.

D.1. Bounding the Empirical Restricted Eigenvalue

We denote the sample and population covariance matrices of the log-transformed covariates by Σ and $\hat{\Sigma}$:

$$\Sigma := \mathbb{E}_z[zz^T]$$
$$\hat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} z^{(i)} z^{(i)T}.$$

Theorem 10 gives us a bound on $\tilde{\lambda}(k, \Sigma^{1/2})$. In this section we apply the results of Kuchibhotla and Chakrabortty (2018) to convert this into a bound on $\tilde{\lambda}(k, \hat{\Sigma}^{1/2})$ by analyzing $|\tilde{\lambda}(k, \Sigma^{1/2}) - \tilde{\lambda}(k, \hat{\Sigma}^{1/2})|$. The following lemma shows that it is sufficient to analyze $\|\Sigma - \hat{\Sigma}\|_\infty$.

Lemma 34 *We have,*

$$|\tilde{\lambda}(k, \Sigma^{1/2}) - \tilde{\lambda}(k, \hat{\Sigma}^{1/2})| \leq 16k\|\Sigma - \hat{\Sigma}\|_\infty.$$

Where, $\|\cdot\|_\infty$ denotes the entry-wise $\infty-$norm.

Proof We note that,

$$|\tilde{\lambda}(k, \Sigma^{1/2}) - \tilde{\lambda}(k, \hat{\Sigma}^{1/2})| \leq \max_{v:\|v\|_2=1, \|v_S^c\|_1 \leq 3\|v_S\|_1} v^T(\Sigma - \hat{\Sigma})v.$$

Furthermore for any v which satisfies $\|v\|_2 = 1, \|v_{S^c}\|_1 \leq 3\|v_S\|_1$, we have,

$$\begin{aligned} v^T(\Sigma - \hat{\Sigma})v &\overset{(1)}{\leq} \|v\|_1 \|(\Sigma - \hat{\Sigma})v\|_\infty \\ &= \|v\|_1 \max_i |\langle \Sigma_{i,\cdot} - \hat{\Sigma}_{i,\cdot}, v\rangle| \\ &\overset{(2)}{\leq} \|v\|_1^2 \|\Sigma - \hat{\Sigma}\|_\infty. \end{aligned}$$

In the above display, the inequalities marked (1) and (2) both follow from Holder's Inequality. Furthermore,

$$\begin{aligned} \|v\|_1 &= \|v_S\|_1 + \|v_{S^c}\|_1 \\ &\overset{(3)}{\leq} (1+3)\|v_S\|_1 \\ &\leq 4\sqrt{k}\|v_S\|_2 \\ &\overset{(4)}{\leq} 4\sqrt{k}. \end{aligned}$$

In the above display, the inequality marked (3) follows from $\|v_{S^c}\|_1 \leq 3\|v_S\|_1$, the inequality marked (4) follows from $\|v_S\|_2 \leq \|v\|_2 \leq 1$. Consequently, we have,

$$|\tilde{\lambda}(k, \Sigma^{1/2}) - \tilde{\lambda}(k, \hat{\Sigma}^{1/2})| \leq 16k\|\Sigma - \hat{\Sigma}\|_\infty.$$

■

To analyze $\|\Sigma - \hat{\Sigma}\|_\infty$ we appeal to the concentration results from Kuchibhotla and Chakrabortty (2018). To do so we need to verify two conditions on our covariates:

1. The log-transformed covariates z_i are entry-wise (marginally) subexponential. This is done in Lemma 35.
2. An upper bound on the quantity Γ defined as:

$$\Gamma^2 := \max_{j,k\in[p]} \frac{1}{n}\sum_{i=1}^{n} \operatorname{var}\left(z_j^{(i)} z_k^{(i)}\right).$$

   This is done in Lemma 36.

Lemma 35 *Let $w \sim \mathcal{N}(0,1)$. Then, $z = \log(|w|)$ is 1-subexponential.*

Proof It is sufficient to show that, $\forall t > 0$,

$$\mathbb{P}\left[|\log(|w|)| > t\right] \leq 2\exp(-t).$$

To show this, we bound the upper tail and the lower tail separately. First let us consider the upper tail,

$$\begin{aligned}\mathbb{P}[\log(|w|) > t] &= \mathbb{P}[|w| > e^t] \\ &= 2\mathbb{P}[w > e^t] \\ &\overset{(1)}{\leq} \sqrt{\frac{2}{\pi}}\frac{e^{-e^{2t}/2}}{e^t} \\ &\leq \sqrt{\frac{2}{\pi}}e^{-t}.\end{aligned}$$

In the inequality marked (1) we used the standard estimate for Gaussian tails: $\mathbb{P}[w > \delta] \leq \sqrt{\frac{2}{\pi}}\frac{\exp(-\delta^2/2)}{\delta}$. To bound the lower tail, we use standard estimates on Gaussian anti-concentration,

$$\begin{aligned}\mathbb{P}[\log(|w|) < -t] &= \mathbb{P}[|w| < e^{-t}] \\ &= \frac{1}{\sqrt{2\pi}}\int_{-e^{-t}}^{e^{-t}} \exp(-a^2/2)da \\ &\leq \sqrt{\frac{2}{\pi}}e^{-t}.\end{aligned}$$

Combining the estimates of the lower and upper tail, we get,

$$\mathbb{P}\left[|\log(|w|)| > t\right] \leq 2\sqrt{\frac{2}{\pi}}\exp(-t) < 2\exp(-t)$$

as desired. ■

Lemma 36 *We have the following upper bound on Γ:*

$$\Gamma^2 \leq 48.$$

Proof Since z_i are identically distributed,

$$\Gamma^2 = \max_{j,k\in[p]} \operatorname{var}\left(z_j z_k\right).$$

We have,

$$\begin{aligned}\operatorname{var}(z_j z_k) &\leq \mathbb{E}[(z_j z_k)^2] \\ &\overset{(1)}{\leq} \sqrt{\mathbb{E}[z_i^4]\mathbb{E}[z_j^4]} \\ &\overset{(2)}{=} \mathbb{E}[z_i^4].\end{aligned}$$

In the above display we used the Cauchy-Schwarz Inequality to obtain the inequality marked (1) and the fact that z_i and z_j have the same marginal distribution in the equality marked (2). To bound $\mathbb{E}[z_i^4]$ we use the concentration result proved in Lemma 35.

$$\begin{aligned}\mathbb{E}[z_i^4] &= \int_0^\infty \mathbb{P}(z_i^4 > t)\,dt \\ &= \int_0^\infty \mathbb{P}(|z_i| > t^{1/4})\,dt \\ &\overset{(3)}{\leq} \int_0^\infty 2\exp(-t^{1/4})\,dt \\ &= 48.\end{aligned}$$

In the above display, we used Lemma 35 for the inequality marked (3). ■

We can now apply Theorem 4.1 of Kuchibhotla and Chakrabortty (2018) to control $\|\Sigma - \hat{\Sigma}\|_\infty$ and hence control $|\tilde{\lambda}(k, \Sigma^{1/2}) - \tilde{\lambda}(k, \hat{\Sigma}^{1/2})|$.

Lemma 37 *Let $\delta \in (0,1)$ be an arbitrary confidence parameter. With probability $1-\delta$,*

$$|\tilde{\lambda}(k, \Sigma^{1/2}) - \tilde{\lambda}(k, \hat{\Sigma}^{1/2})| \leq Ck\left(\sqrt{\frac{(\log(3/\delta) + 2\log(p))}{n}} + \frac{\log^2(n)(\log(3/\delta) + 2\log(p))^2}{n}\right).$$

In the above display C is a universal constant.

Proof From Lemma 35, we know that the log-transformed covariates are marginally subexponential. Applying Theorem 4.1 of Kuchibhotla and Chakrabortty (2018) for marginally subexponential random variables, we have with probability atleast $1 - 3e^{-t}$,

$$\|\Sigma - \hat{\Sigma}\|_\infty \leq C\left(\sqrt{\frac{\Gamma(t + 2\log(p))}{n}} + \frac{\log^2(n)(t + 2\log(p))^2}{n}\right),$$

where C is a universal constant. Substituting the bound on Γ from Lemma 36 and then applying Lemma 34 we get,

$$|\tilde{\lambda}(k, \Sigma^{1/2}) - \tilde{\lambda}(k, \hat{\Sigma}^{1/2})| \leq Ck\left(\sqrt{\frac{(t + 2\log(p))}{n}} + \frac{\log^2(n)(t + 2\log(p))^2}{n}\right).$$

Substituting $t = \log(3/\delta)$ gives us the required bound. ■

We are now ready to present the proof of Theorem 4 which is restated and proved below.

Theorem 38 *Let $\delta \in (0,1)$ be an arbitrary confidence parameter. Suppose the covariance matrix Φ satisfies $\Phi_{i,i} = 1, \forall i \in [p]$ and $\max_{i \neq j} |\Phi_{i,j}| < 1 - \epsilon$. Then, we have that log-transformed design matrix satisfies the restricted eigenvalue bound:*

$$\tilde{\lambda}(k, \hat{\Sigma}^{1/2}) \geq \frac{1}{5}\sqrt{\frac{\epsilon}{\log(16k) + 2}},$$

with probability $1-\delta$*, provided,*

$$n \geq \frac{Ck^2}{\epsilon} \log^2\left(\frac{2pk}{\delta}\right) \log^2\left(\frac{k}{\epsilon}\log\left(\frac{2pk}{\delta}\right)\right).$$

In the above display, C *is a universal constant.*

Proof For the ease of notation, we define $|\rho_{\max}| = 1-\epsilon$. From Theorem C.2.2, we know that,

$$\begin{aligned}
\tilde{\lambda}(k, \Sigma^{1/2}) &\geq \frac{2}{5}\left(\frac{2\log\left(\frac{1}{1-\epsilon}\right)}{\log\left(\frac{16k}{1-\epsilon}\right) + \max\{2, \log\left(\frac{1}{1-\epsilon}\right)\}}\right)^{1/2} \\
&\overset{(1)}{\geq} \frac{2}{5}\sqrt{\frac{2\log\left(\frac{1}{1-\epsilon}\right)}{\log(16k) + 2 + 2\log\left(\frac{1}{1-\epsilon}\right)}} \\
&\overset{(2)}{\geq} \frac{\sqrt{2}}{5}\min\left(1, \sqrt{\frac{\log\left(\frac{1}{1-\epsilon}\right)}{2+\log(16k)}}\right)
\end{aligned}$$

In the display marked above, we used the fact that $\max(a,b) \leq a+b$ in the inequality marked (1). In the inequality marked (2) we used the fact for any $x, c \geq 0$, we have $\frac{x}{x+c} \geq \frac{1}{2}\min\left(\frac{x}{c}, 1\right)$. By Lemma 37, we know that with probability $1-\delta$,

$$\begin{aligned}
\tilde{\lambda}(k, \hat{\Sigma}^{1/2}) \geq \frac{\sqrt{2}}{5}\min\left(1, \sqrt{\frac{\log\left(\frac{1}{1-\epsilon}\right)}{2+\log(16k)}}\right) \\
- Ck\left(\sqrt{\frac{(\log(3/\delta) + 2\log(p))}{n}} + \frac{\log^2(n)(\log(3/\delta)+2\log(p))^2}{n}\right).
\end{aligned}$$

Hence there exists a constant C such that if,

$$\frac{n}{\log^2(n)} \geq Ck^2\left(\log(3/\delta) + 2\log(p)\right)^2 \max\left(1, \frac{\log(16k)+2}{\log\left(\frac{1}{1-\epsilon}\right)}\right),$$

we have that $\hat{\Sigma}$ satisfies the restricted eigenvalue bound:

$$\tilde{\lambda}(k, \hat{\Sigma}^{1/2}) \geq \frac{1}{5}\min\left(1, \sqrt{\frac{\log\left(\frac{1}{1-\epsilon}\right)}{2+\log(16k)}}\right)$$

Finally, we clean up this bound. First we note that $\log\left(\frac{1}{1-\epsilon}\right) \geq \epsilon$. Hence, if n is large enough so that,

$$\frac{n}{\log^2(n)} \geq \frac{Ck^2\log(2k)}{\epsilon}\log^2\left(\frac{2p}{\delta}\right),$$

we have, with probability $1-\delta$,

$$\tilde{\lambda}(k, \hat{\Sigma}^{1/2}) \geq \frac{1}{5}\sqrt{\frac{\epsilon}{2+\log(16k)}}$$

Finally, we note to satisfy the requirement on the sample size, it is sufficient that,

$$n \geq \frac{Ck^2\log(2k)}{\epsilon}\log^2\left(\frac{2p}{\delta}\right)\log^2\left(\frac{k\log(2k)}{\epsilon}\log\left(\frac{2p}{\delta}\right)\right).$$

■

Appendix E. Proof of the Gershgorin's Circle Theorem for Restricted Eigenvalue

In this section, we prove the Gershgorin's theorem for the restricted eigenvalue. Let A be a $p\times p$ symmetric matrix and S be an arbitrary subset of $[p]$. Let $\tilde{\lambda}(\alpha, S, A^{1/2})$ denote the Restricted eigenvalue defined as:

$$\tilde{\lambda}(\alpha, S, A^{1/2}) = \min v^T A v \text{ subject to: } \|v\|_2 = 1, \|v_{S^c}\|_1 \leq \alpha\|v_S\|_1.$$

The goal is to prove the following theorem.

Theorem 39 *For $\alpha \geq 1$, we have,*

$$\tilde{\lambda}(\alpha, S, A^{1/2}) \geq \min_{i\in[p]} A_{i,i} - (1+\alpha)^2\cdot|S|\cdot\max_{i\neq j}|A_{ij}|.$$

Let $v^\star$ be the optimizer of the Restricted Eigenvalue problem. To simplify notation, we will short hand the optimal objective $\tilde{\lambda}(\alpha, S, A^{1/2})$ as $\lambda^\star$. Without loss of generality we can assume $|v_i^\star| > 0\ \forall i \in [p]$. This is because of the following reason: Let T denote the support of the optimal $v^\star$. It is straightforward to see that $\lambda^\star$ and $v^\star(T)$ are the optimal objective value and the optimizer of the following problem:

$$\min v^T A(T,T) v \text{ subject to: } \|v\|_2 = 1, \|v_{T\cap S^c}\|_1 \leq \alpha\|v_{T\cap S}\|_1.$$

If $T \neq [p]$, then we can make the arguments that follow for the optimization problem defined in the display above.

The proof of the usual Gershgorin Theorem begins with the optimality condition for the unconstrained eigenvalue problem. Taking cue from the original proof, we first derive an optimality condition for the restricted eigenvalue problem. We then utilize this to prove a lower bound on $\lambda^\star$.

Proof We first write the local optimality condition at $v^\star$. For $\lambda \in \mathbb{R}$ and $q \geq 0$, we form the Lagrangian:

$$L(v, \lambda, q) = v^T A v - \lambda\|v\|_2^2 + 2q\left(\|v_{S^c}\|_1 - \alpha\|v_S\|_1\right).$$

Since $|v_i^\star| > 0 \; \forall i$, by the method of Lagrange multipliers, the local optimality condition at $v^\star$ is:

$$\exists \lambda \in \mathbb{R}, q \geq 0 \text{ such that } \nabla_v L(v^\star, \lambda, q) = 0.$$

This means,

$$Av^\star - \lambda v^\star - qu = 0. \tag{11}$$

Where, the vector $u \in \mathbb{R}^p$ is defined as:

$$u_i = \begin{cases} \alpha \text{sign}(v_i^\star) & i \in S \\ -\text{sign}(v_i^\star) & i \notin S. \end{cases}$$

Taking dot-product with $v^\star$ on both sides of equation 11, we get,

$$\begin{aligned} \lambda^\star &= \lambda + q(\alpha \|v_S\|_1 - \|v_{S^c}\|_1) \\ &\geq \lambda. \end{aligned}$$

Hence to lower bound $\lambda^\star$, it is sufficient to lower bound λ. Let i be the coordinate that maximizes $|v_i^\star|$. Then, we have,

$$\sum_{j \neq i} A_{ij} v_j^\star + A_{ii} v_i^\star - \lambda v_i^\star = q u_i. \tag{12}$$

However, since q is unknown, to eliminate it we consider another coordinate k. This coordinate k is chosen so that: If $i \in S$, $k \notin S$ and if $i \notin S$, then $k \in S$. We have,

$$\sum_{j \neq k} A_{kj} v_j^\star + A_{kk} v_k^\star - \lambda v_k^\star = q u_k. \tag{13}$$

Hence, we can eliminate q between equations 12 and 13,

$$\frac{v_i^\star}{u_i} A_{ii} - \frac{v_k^\star}{u_k} A_{kk} - \lambda \left(\frac{v_i^\star}{u_i} - \frac{v_k^\star}{u_k} \right) = \frac{1}{u_k} \left(\sum_{j \neq k} A_{kj} v_j^\star \right) - \frac{1}{u_i} \left(\sum_{j \neq i} A_{ij} v_j^\star \right).$$

Taking absolute values,

$$\left| \frac{v_i^\star}{u_i} A_{ii} - \frac{v_k^\star}{u_k} A_{kk} - \lambda \left(\frac{v_i^\star}{u_i} - \frac{v_k^\star}{u_k} \right) \right| = \left| \frac{1}{u_k} \left(\sum_{j \neq k} A_{kj} v_j^\star \right) - \frac{1}{u_i} \left(\sum_{j \neq i} A_{ij} v_j^\star \right) \right|.$$

Dividing throughout by $|\frac{v_i^\star}{u_i}|$:

$$\begin{aligned} \left| A_{ii} - \frac{v_k^\star u_i}{u_k v_i^\star} A_{kk} - \lambda \left(1 - \frac{v_k^\star u_i}{v_i^\star u_k} \right) \right| &= \left| \frac{u_i}{u_k} \left(\sum_{j \neq k} A_{kj} \frac{v_j^\star}{v_i^\star} \right) - \left(\sum_{j \neq i} A_{ij} \frac{v_j^\star}{v_i^\star} \right) \right| \\ &\leq |S| \cdot (1+\alpha) \cdot \left(\max_{l \neq m} |A_{lm}| \right) \cdot \left(\frac{|u_i|}{|u_k|} + 1 \right) \\ &\leq |S| \cdot (1+\alpha)^2 \cdot \left(\max_{l \neq m} |A_{lm}| \right). \end{aligned}$$

Next we note because of the choice of k (if $i \in S, k \notin S$, if $i \notin S$, $k \in S$) and the definition of u,

$$\rho := -\frac{v_k^\star u_i}{u_k v_i^\star} \geq 0.$$

Dividing through out by $1 + \rho$ gives,

$$\left| \frac{A_{ii} + \rho A_{kk}}{1+\rho} - \lambda \right| \leq \frac{|S| \cdot (1+\alpha)^2 \cdot \left(\max_{l \neq m} |A_{lm}|\right)}{1+\rho} \leq |S| \cdot (1+\alpha)^2 \cdot \left(\max_{l \neq m} |A_{lm}|\right).$$

Next noting that,

$$\frac{A_{ii} + \rho A_{kk}}{1+\rho} \geq \min_i A_{ii},$$

we have the following lower bound,

$$\lambda^\star \geq \lambda \geq (\min_i A_{ii}) - |S| \cdot (1+\alpha)^2 \cdot \left(\max_{l \neq m} |A_{lm}|\right).$$

■