Benign Overfitting in Linear Regression

Peter L. Bartlett
CS and Statistics
UC Berkeley
peter@berkeley.edu

Philip M. Long
Google
plong@google.com

Gábor Lugosi
Economics and Business
Pompeu Fabra University;
ICREA,
Pg. Lluís Companys 23,
08010 Barcelona, Spain;
Barcelona Graduate School of Economics
gabor.lugosi@upf.edu

Alexander Tsigler
Statistics
UC Berkeley
alexander_tsigler@berkeley.edu

June 26, 2019

Abstract

The phenomenon of benign overfitting is one of the key mysteries uncovered by deep learning methodology: deep neural networks seem to predict well, even with a perfect fit to noisy training data. Motivated by this phenomenon, we consider when a perfect fit to training data in linear regression is compatible with accurate prediction. We give a characterization of gaussian linear regression problems for which the minimum norm interpolating prediction rule has near-optimal prediction accuracy. The characterization is in terms of two notions of the effective rank of the data covariance. It shows that overparameterization is essential for benign overfitting in this setting: the number of directions in parameter space that are unimportant for prediction must significantly exceed the sample size.

1 Introduction

Deep learning methodology has revealed a surprising statistical phenomenon: overfitting can perform well. The classical perspective in statistical learning theory is that there should be a tradeoff between the fit to the training data and the complexity of the prediction rule. Whether complexity is measured in terms of the number of parameters, the number of non-zero parameters in a high-dimensional setting, the number of neighbors averaged in a nearest-neighbor estimator, the scale of an estimate in a reproducing kernel Hilbert space, or the bandwidth of a kernel smoother, this tradeoff has been ubiquitous in statistical learning theory. Deep learning seems to operate outside the regime where results of this kind are informative, since deep neural networks can perform well even with a perfect fit to the training data.

As one example of this phenomenon, consider the experiment illustrated in Figure 1(c) in [25]: standard deep network architectures and stochastic gradient algorithms, run until they perfectly fit a standard image classification training set, give respectable prediction performance, *even when significant levels of label noise are introduced*. That paper reported zero loss for training data classifications. However, the deep networks in the experiments reported in [25] also achieved essentially zero cross-entropy loss on the training data.[1] In statistics and machine learning textbooks, an estimate that fits every training example perfectly is often presented as an illustration of overfitting ("... interpolating fits... [are] unlikely to predict future data well at all." [14, p37]). Thus, to arrive at a scientific understanding of the success of deep learning methods, it is a central challenge to understand the performance of prediction rules that interpolate the training data.

In this paper, we consider perhaps the simplest setting where we might hope to witness this phenomenon: linear regression with gaussian data. That is, we assume that the covariates and responses have a gaussian distribution, the loss is quadratic, the prediction rules are linear, and the dimension of the parameter space is large enough that a perfect fit is guaranteed. We consider data in an infinite dimensional space (a separable Hilbert space), but our results apply to a finite-dimensional subspace as a special case. There is an ideal value of the parameters, θ^*, the least squares parameters. We ask when it is possible to fit the data exactly and still compete with the prediction accuracy of θ^*. Since we require more parameters than the sample size in order to fit exactly, the solution might be underdetermined, so there might be many interpolating solutions. We consider the most natural: choose the parameter vector $\hat{\theta}$ with the smallest norm among all vectors that give perfect predictions on the training sample. (This corresponds to using the pseudoinverse to solve the normal equations; see Section 2.) We ask when it is possible to overfit in this way—and embed all of the noise of the labels into the parameter estimate $\hat{\theta}$—without harming prediction accuracy.

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, personal communication, January 2017.

Our main result is a finite sample characterization of when overfitting is benign in this setting. The gaussian linear regression problem depends on the least squares parameters θ^* and the covariance Σ of the covariates x. The properties of Σ turn out to be crucial, since the magnitude of the variance in different directions determines both how the label noise gets distributed across the parameter space and how errors in parameter estimation in different directions in parameter space affect prediction accuracy. There is a classical decomposition of the excess prediction error into a bias term and a variance term. The bias part is rather standard: provided that the scale of the problem (that is, the sum of the eigenvalues of Σ) is small compared to the sample size n, the contribution to $\hat{\theta}$ that we can view as coming from θ^* is not too distorted. The variance part is more interesting, since it reflects the impact of the noise in the labels on prediction accuracy. We show that this part is small if and only if the effective rank of Σ in the subspace corresponding to low variance directions is large compared to n. This necessary and sufficient condition of a large effective rank can be viewed as a property of significant overparameterization: fitting the training data exactly but with near-optimal prediction accuracy occurs if and only if there are many low variance (and hence unimportant) directions in parameter space where the label noise can be hidden.

The details are more complicated. The characterization depends in a specific way on *two* notions of effective rank, r and R; the smaller one, r, determines a split of Σ into large and small eigenvalues, and the excess prediction error depends on the effective rank, as measured by the larger notion R, of the subspace corresponding to the smallest eigenvalues. For the excess prediction error to be small, the smallest eigenvalues of Σ must decay slowly.

The phenomenon of interpolating prediction rules has been an object of study by several authors over the last two years, since it emerged as an intriguing mystery at the Simons Institute program on Foundations of Machine Learning in Spring 2017. Belkin, Ma and Mandal [7] described an experimental study demonstrating that this phenomenon of accurate prediction for functions that interpolate noisy data also occurs for prediction rules chosen from reproducing kernel Hilbert spaces, and explained the mismatch between this phenomenon and classical generalization bounds. Belkin, Hsu and Mitra [3] gave an example of an interpolating nearest neighbor decision rule with an asymptotic consistency property as the input dimension gets large. Belkin, Rakhlin, and Tsybakov [4] showed that kernel smoothing methods based on singular kernels both interpolate and, with suitable bandwidth choice, give optimal rates for nonparametric estimation (building on earlier consistency results [9] for these unusual kernels). Liang and Rakhlin [18] considered minimum norm interpolating kernel regression with kernels defined as nonlinear functions of the Euclidean inner product, and showed that, with certain properties of the training sample (expressed in terms of the empirical kernel matrix), these methods can have good prediction accuracy. Belkin, Hsu, Ma and Mandal [5] studied experimentally the excess risk as a function of the dimension of a sequence of parameter spaces for linear and non-linear classes.

Subsequent to our work, [19] considered the properties of the interpolating linear prediction rule with minimal expected squared error. After this work was presented at the NAS Colloquium on the Science of Deep Learning [2], we became aware of the concurrent work of Belkin, Hsu and Xu [6] and of Hastie, Montanari, Rosset and Tibshirani [13]. Belkin *et al* [6] calculated the excess risk for certain linear models (a regression problem with identity covariance, sparse least squares parameters, both with and without noise, and a problem with random Fourier features with no noise), and Hastie *et al* consider linear regression in an asymptotic regime, where sample size n and input dimension p go to infinity together with asymptotic ratio $p/n \to \gamma$. Rather than exploiting independence in the components of a gaussian distribution, they assume that the covariate vectors are linear transformations of vectors with p i.i.d. entries. They also assume that, as p gets large, the empirical spectral distribution of Σ (the discrete measure on its set of eigenvalues) converges to a fixed measure. They apply random matrix theory to explore the range of behaviors of the asymptotics of the excess prediction error as γ, the noise variance, and the eigenvalue distribution vary. They also study the asymptotics of a model involving random nonlinear features. In contrast, we assume that the data is gaussian, and we give upper and lower bounds on the excess prediction error for data of arbitrary dimension, for arbitrary finite sample size, and for arbitrary covariance matrices.

The next section introduces notation and definitions used throughout the paper, including definitions of the problem of linear regression in a gaussian setting and of various notions of effective rank of the covariance operator. Section 3 gives the characterization of benign overfitting, and illustrates why the effective rank condition corresponds to significant overparameterization, and presents several examples of patterns of eigenvalues that allow benign overfitting. Section 4 gives the proofs of these results.

2 Definitions and Notation

We consider linear regression problems, where a linear function of covariates x from a (potentially infinite dimensional) Hilbert space $\mathbb{H}$ is used to predict a real-valued response variable y. We use vector notation, so that $x^\top \theta$ denotes the inner product between x and θ and $xz^\top$ denotes the tensor product of $x, z \in \mathbb{H}$.

Definition 1 (Linear regression)**.** *A linear regression problem in a separable Hilbert space $\mathbb{H}$ is defined by a random covariate vector $x \in \mathbb{H}$ and outcome $y \in \mathbb{R}$. We assume (x, y) are jointly gaussian and mean zero, and define*

1. *the* covariance operator $\Sigma = \mathbb{E}[xx^\top]$,
2. *the* least squares parameter vector $\theta^* \in \mathbb{H}$, *satisfying* $\mathbb{E}(y - x^\top \theta^*)^2 = \operatorname{argmin}_\theta \mathbb{E}(y - x^\top \theta)^2$, *and*
3. *the* noise variance, $\sigma^2 = \mathbb{E}(y - x^\top \theta^*)^2$.

Given a training sample $(x_1, y_1), \ldots, (x_n, y_n)$ of n i.i.d. pairs with the same distribution as (x, y), an estimator *returns a parameter estimate $\theta \in \mathbb{H}$. The excess risk of the estimator is defined as*

$$R(\theta) := \mathbb{E}_{x,y}\left[\left(y - x^\top \theta\right)^2 - \left(y - x^\top \theta^*\right)^2\right],$$

where $\mathbb{E}_{x,y}$ denotes the conditional expectation given all random quantities other than x, y (in this case, given the estimate θ). Define the vectors $\boldsymbol{y} \in \mathbb{R}^n$ with entries y_i and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ with entries $\varepsilon_i = y_i - x_i^\top \theta^*$. We use infinite matrix notation: X denotes the linear map from $\mathbb{H}$ to $\mathbb{R}^n$ corresponding to $x_1, \ldots, x_n \in \mathbb{H}$, so that $X\theta \in \mathbb{R}^n$ has ith component $x_i^\top \theta$. We use similar notation for the linear map $X^\top$ from $\mathbb{R}^n$ to $\mathbb{H}$.

The assumption that y is gaussian simplifies the statement of the main theorem, but the proof requires only that the conditional distribution of $y - \mathbb{E}[y|x]$ given x is subgaussian, independent of x, and that the conditional expectation $\mathbb{E}[y|x]$ is linear in x, all of which follow from the joint gaussianity assumption.

We shall be concerned with situations where an estimator θ can fit the data perfectly, that is,

$$X\theta = \boldsymbol{y}.$$

Typically this implies that there are many such vectors. We consider the interpolating estimator with minimal norm in $\mathbb{H}$. We use $\|\cdot\|$ to denote both the Euclidean norm of a vector in $\mathbb{R}^n$ and the Hilbert space norm.

Definition 2 (Minimum norm estimator)**.** *Given data $X \in \mathbb{H}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$, the* minimum norm estimator $\hat{\theta}$ *solves the optimization problem*

$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ \textit{such that} \quad & \|X\theta - \boldsymbol{y}\|^2 = \min_{\beta} \|X\beta - \boldsymbol{y}\|^2 . \end{aligned}$$

By the projection theorem, parameter vectors that solve the least squares problem $\min_\beta \|X\beta - \boldsymbol{y}\|^2$ solve the normal equations, so we can equivalently write $\hat{\theta}$ as the minimum norm solution to the normal equations,

$$\begin{aligned} \hat{\theta} &= \arg\min_\theta \left\{ \|\theta\|^2 : X^\top X\theta = X^\top \boldsymbol{y} \right\} \\ &= \left(X^\top X\right)^\dagger X^\top \boldsymbol{y} \\ &= X^\top \left(XX^\top\right)^\dagger \boldsymbol{y}, \end{aligned}$$

where $\left(X^\top X\right)^\dagger$ denotes the pseudoinverse of the bounded linear operator $X^\top X$ (for infinite dimensional $\mathbb{H}$, the existence of the pseudoinverse is guaranteed because $X^\top X$ is bounded and has a closed range; see [8]). When $\mathbb{H}$ is finite dimensional with $p < n$ and X has rank p, there is a unique solution to the normal equations. We shall assume for the remainder of the paper that $\operatorname{rank}(\Sigma) \geq n$, in

which case the normal equations can have many solutions. Since x is gaussian with a positive definite covariance, almost surely $XX^\top$ has full rank, and so we can find a solution $\theta \in \mathbb{H}$ that achieves $X\theta = y$. The minimum norm solution is given by

$$\hat{\theta} = X^\top \left(XX^\top\right)^{-1} \boldsymbol{y}. \tag{1}$$

Our main result gives tight bounds on the excess risk of the minimum norm estimator in terms of certain notions of effective rank of the covariance that are defined in terms of its eigenvalues.

We use $\mu_1(\Sigma) \geq \mu_2(\Sigma) \geq \cdots$ to denote the eigenvalues of Σ in descending order, and we denote the operator norm of Σ by $\|\Sigma\|$. We use I to denote the identity operator on $\mathbb{H}$ and I_n to denote the $n \times n$ identity matrix.

Definition 3 (Effective Ranks)**.** *For the covariance operator Σ, define $\lambda_i = \mu_i(\Sigma)$ for $i = 1, 2, \ldots$. If $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$ for $k \geq 0$, define*

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad\qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

3 Main Result

The following theorem establishes nearly matching upper and lower bounds for the risk of the minimum-norm interpolating estimator.

Theorem 4. *There are universal constants $b, c > 1$ for which the following holds. Consider a linear regression problem with least squares parameter vector $\theta^* \in \mathbb{H}$, covariance operator Σ, noise variance σ^2 and sample size $n \geq c$. Suppose that $\operatorname{rank}(\Sigma) \geq n$, so that the minimum norm estimator satisfies $X\hat{\theta} = \boldsymbol{y}$ almost surely. Define*

$$k^* = \min\left\{k \geq 0 : r_k(\Sigma) \geq bn\right\},$$

where the minimum of the empty set is defined as ∞. Suppose $\delta < 1$ with $\log(1/\delta) < n/c$. Then with probability at least $1 - \delta$,

$$R(\hat{\theta}) \leq c\left(\|\theta^*\|^2 \|\Sigma\| \max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}\right) + c\log(1/\delta)\sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right),$$

and

$$\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c} \min\left\{1, \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right\}.$$

3.1 Effective Ranks and Overparameterization

In order to understand the implications of Theorem 4, we now study relationships between the two notions of effective rank, r_k and R_k, and establish sufficient and necessary conditions for the sequence $\{\lambda_i\}$ of eigenvalues to lead to small excess risk.

The following lemma shows that the two notions of effective rank are closely related. See Appendix A.6 for its proof, and for other properties of r_k and R_k.

Lemma 5. *$r_k(\Sigma) \geq 1$, $r_k^2(\Sigma) = r_k(\Sigma^2)R_k(\Sigma)$, and*

$$r_k(\Sigma^2) \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma).$$

Notice that $r_0(I_p) = R_0(I_p) = p$. More generally, if all the non-zero eigenvalues of Σ are identical, then $r_0(\Sigma) = R_0(\Sigma) = \mathrm{rank}(\Sigma)$. For Σ with finite rank, we can express both $r_0(\Sigma)$ and $R_0(\Sigma)$ as a product of the rank and a notion of symmetry. In particular, for $\mathrm{rank}(\Sigma) = p$ we can write

$$\begin{aligned} & r_0(\Sigma) = \mathrm{rank}(\Sigma) s(\Sigma), && R_0(\Sigma) = \mathrm{rank}(\Sigma) S(\Sigma), \\ \text{with} \qquad & s(\Sigma) = \frac{(1/p)\sum_{i=1}^p \lambda_i}{\lambda_1}, && S(\Sigma) = \frac{\left((1/p)\sum_{i=1}^p \lambda_i\right)^2}{(1/p)\sum_{i=1}^p \lambda_i^2}. \end{aligned}$$

Both notions of symmetry s and S lie between $1/p$ (when $\lambda_2 \to 0$) and 1 (when the λ_i are all equal).

Theorem 4 shows that, for the minimum norm estimator to have near-optimal prediction accuracy, $r_0(\Sigma)$ should be small compared to the sample size n (from the first term) and $r_{k^*}(\Sigma)$ and $R_{k^*}(\Sigma)$ should be large compared to n. Together, these conditions imply that overparameterization is essential for benign overfitting in this setting: the number of non-zero eigenvalues should be large compared to n, they should have a small sum compared to n, and there should be many eigenvalues no larger than λ_{k^*}. If the number of these small eigenvalues is not much larger than n, then they should be roughly equal, but they can be more assymmetric if there are many more of them.

The following theorem shows that the kind of overparameterization that is essential for benign overfitting requires Σ to have a heavy tail. (The proof is in Appendix A.7.) In particular, if we fix Σ in an infinite-dimensional Hilbert space and ask when does the excess risk of the minimum norm estimator approach zero as $n \to \infty$, it imposes tight restrictions on the eigenvalues of Σ. But there are many other possibilities for these asymptotics if Σ can change with n.

We say that a sequence of covariance operators Σ_n is *benign* if

$$\lim_{n\to\infty} \left(\|\Sigma_n\| \sqrt{\frac{r_0(\Sigma_n)}{n}} + \frac{k_n^*}{n} + \frac{n}{R_{k_n^*}(\Sigma_n)} \right) = 0,$$

where $k_n^* = \min\{k \geq 0 : r_k(\Sigma_n) \geq bn\}$ for the universal constant b in Theorem 4.

Theorem 6. *Define $\lambda_{k,n} := \mu_k(\Sigma_n)$ for all k, n.*

1. *If* $\lambda_{k,n} = k^{-\alpha} \ln^{-\beta}(k+1)$, *then* Σ_n *is benign iff* $\alpha = 1$ *and* $\beta > 1$.

2. *If* $\lambda_{k,n} = k^{-(1+\alpha_n)}$, *then* Σ_n *is benign iff* $\omega(1/n) = \alpha_n = o(1)$. *Furthermore,*
$$R(\hat{\theta}) = \Theta\left(\min\left\{\frac{1}{\alpha_n n} + \alpha_n, 1\right\}\right).$$

3. *If*
$$\lambda_{k,n} = \begin{cases} k^{-\alpha} & \text{if } k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$
then Σ_n *is benign iff either* $0 < \alpha < 1$, $p_n = \omega(n)$ *and* $p_n = o\left(n^{1/(1-\alpha)}\right)$ *or* $\alpha = 1$, $p_n = e^{\omega(\sqrt{n})}$ *and* $p_n = e^{o(n)}$.

4. *If*
$$\lambda_{k,n} = \begin{cases} e^{-k} + \epsilon_n & \text{if } k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$
then Σ_n *is benign iff* $p_n = \omega(n)$ *and* $ne^{-o(n)} = \epsilon_n p_n = o(n)$. *Furthermore, for* $p_n = \Omega(n)$ *and* $\epsilon_n p_n = ne^{-o(n)}$,
$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n + 1}{n} + \frac{\ln(n/(\epsilon_n p_n))}{n} + \max\left\{\frac{1}{n}, \frac{n}{p_n}\right\}\right).$$

Many popular learning algorithms may be viewed as the combination of a feature transformation with a linear method applied to the transformed features. In practice, hyperparameter choices such as kernel parameters and deep network architectures can be made with the knowledge of n, and of course these affect the spectrum of the covariance of the transformed features.

It is informative to compare the situations described by Parts 1 and 4 of Theorem 6. Part 1 shows that for infinite-dimensional data with a fixed covariance, benign overfitting occurs iff the eigenvalues of the covariance operator decay just slowly enough for their sum to remain finite. However, Part 4 shows that the situation is very different if the data has finite dimension and a small amount of isotropic noise is added to the covariates. In that case, even if the eigenvalues of the original covariance operator (before the addition of isotropic noise) decay very rapidly, benign overfitting occurs iff both the dimension is large compared to the sample size, and the isotropic component of the covariance is sufficiently small—but not exponentially small—compared to the sample size.

How relevant is Theorem 4 to the phenomenon of benign overfitting in deep neural networks? Certain very wide neural networks trained with suitable random initialization and gradient descent can be accurately approximated by linear functions in a certain randomly chosen Hilbert space, and in this case gradient descent finds an interpolating solution quickly; see [17, 11, 10, 1]. (Note that these papers do not consider prediction accuracy, except when there is no noise; for example, [17, Assumption A1] implies that the network can compute a suitable real-valued response exactly, and the data-dependent bound of [1,

Theorem 5.1] becomes vacuous when independent noise is added to the y_is.) The eigenvalues of the covariance operator in this case appear to have a heavy tail (see [24], where this kernel was introduced, and [15]), as required for benign overfitting. Of course, the assumptions of Theorem 4 do not apply. Even so, the assumption that the network width is large compared to the sample size is somewhat strange, so extending Theorem 4 to more realistic deep network architectures seems to require significant progress beyond the linear case.

4 Proof

Throughout the proofs, we use the symbols b, c, $c_1, c_2, \ldots$ to refer to universal constants whose values are suitably large (and always at least 1) but do not depend on any parameters of the problems we consider.

Bias-variance decomposition

The first step is a standard decomposition of the excess risk into two pieces, one that corresponds to the distortion that is introduced by viewing θ^* through the lens of the finite sample and one that corresponds to the distortion introduced by the noise $\boldsymbol{\varepsilon} = \boldsymbol{y} - X\theta$. The impact of both sources of error in $\hat{\theta}$ on the excess risk is modulated by the covariance Σ, which gives different weight to different directions in parameter space.

Lemma 7. *The excess risk of the minimum norm estimator satisfies*

$$R(\hat{\theta}) = \mathbb{E}_x \left(x^\top \left(\theta^* - \hat{\theta} \right) \right)^2 \leq 2{\theta^*}^\top B \theta^* + 2\boldsymbol{\varepsilon}^\top C \boldsymbol{\varepsilon},$$

and

$$\mathbb{E}_{x,\boldsymbol{\varepsilon}} R(\hat{\theta}) = {\theta^*}^\top B \theta^* + \sigma^2 \operatorname{tr}(C),$$

where

$$\begin{aligned} B &= \left(I - X^\top \left(X X^\top \right)^{-1} X \right) \Sigma \left(I - X^\top \left(X X^\top \right)^{-1} X \right), \\ C &= \left(X X^\top \right)^{-1} X \Sigma X^\top \left(X X^\top \right)^{-1}. \end{aligned}$$

Proof. Since $\varepsilon = y - x^\top \theta^*$ is independent of x and has mean zero,

$$\begin{aligned} R(\hat{\theta}) &= \mathbb{E}_{x,y} \left(y - x^\top \hat{\theta} \right)^2 - \mathbb{E} \left(y - x^\top \theta^* \right)^2 \\ &= \mathbb{E}_{x,y} \left(y - x^\top \theta^* + x^\top \left(\theta^* - \hat{\theta} \right) \right)^2 - \mathbb{E} \left(y - x^\top \theta^* \right)^2 \\ &= \mathbb{E}_x \left(x^\top \left(\theta^* - \hat{\theta} \right) \right)^2. \end{aligned}$$

Using (1), the definition of Σ, and the fact that $\boldsymbol{y} = X\theta^* + \boldsymbol{\varepsilon}$,

$$\begin{aligned}
R(\hat{\theta}) &= \mathbb{E}_x \left(x^\top \left(I - X^\top \left(XX^\top \right)^{-1} X \right) \theta^* - x^\top X^\top \left(XX^\top \right)^{-1} \boldsymbol{\varepsilon} \right)^2 \\
&\leq 2\mathbb{E}_x \left(x^\top \left(I - X^\top \left(XX^\top \right)^{-1} X \right) \theta^* \right)^2 + 2\mathbb{E}_x \left(x^\top X^\top \left(XX^\top \right)^{-1} \boldsymbol{\varepsilon} \right)^2 \\
&= 2\theta^{*\top} \left(I - X^\top \left(XX^\top \right)^{-1} X \right) \Sigma \left(I - X^\top \left(XX^\top \right)^{-1} X \right) \theta^* \\
&\quad + 2\boldsymbol{\varepsilon}^\top \left(XX^\top \right)^{-1} X\Sigma X^\top \left(XX^\top \right)^{-1} \boldsymbol{\varepsilon} \\
&= 2\theta^{*\top} B\theta^* + 2\boldsymbol{\varepsilon}^\top C\boldsymbol{\varepsilon}.
\end{aligned}$$

Also, since $\boldsymbol{\varepsilon}$ has zero mean and is independent of x and X, we have

$$\begin{aligned}
\mathbb{E}_{x,\boldsymbol{\varepsilon}} R(\hat{\theta}) &= \mathbb{E}_{x,\boldsymbol{\varepsilon}} \left(x^\top \left(I - X^\top \left(XX^\top \right)^{-1} X \right) \theta^* - x^\top X^\top \left(XX^\top \right)^{-1} \boldsymbol{\varepsilon} \right)^2 \\
&= \mathbb{E}_{x,\boldsymbol{\varepsilon}} \left[\left(x^\top \left(I - X^\top \left(XX^\top \right)^{-1} X \right) \theta^* \right)^2 + \left(x^\top X^\top \left(XX^\top \right)^{-1} \boldsymbol{\varepsilon} \right)^2 \right] \\
&= \theta^{*\top} \left(I - X^\top \left(XX^\top \right)^{-1} X \right) \Sigma \left(I - X^\top \left(XX^\top \right)^{-1} X \right) \theta^* \\
&\quad + \operatorname{tr} \left(\left(XX^\top \right)^{-1} X\Sigma X^\top \left(XX^\top \right)^{-1} \mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \right) \\
&= \theta^{*\top} B\theta^* + \sigma^2 \operatorname{tr}(C).
\end{aligned}$$

□

We can control the term $\theta^{*\top} B\theta^*$ using a standard argument. The proof of the following lemma is in Appendix A.1.

Lemma 8. *There is a universal constant c such that for any $1 < t < n$, with probability at least $1 - e^{-t}$,*

$$\theta^{*\top} B\theta^* \leq c\|\theta^*\|^2 \|\Sigma\| \max\left\{ \sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{t}{n}} \right\}.$$

The following lemma shows that we can obtain a high-probability upper bound on the term $\boldsymbol{\varepsilon}^\top C\boldsymbol{\varepsilon}$ in terms of the trace of C. It is Lemma 36 in [20].

Lemma 9. *Consider random variables $\varepsilon_1, \ldots, \varepsilon_n$, conditionally independent given X and conditionally σ^2-subgaussian, that is, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\exp(\lambda\varepsilon_i)|X] \leq \exp(\sigma^2\lambda^2/2).$$

Suppose that, given X, $M \in \mathbb{R}^{n\times n}$ is a.s. positive semidefinite. Then a.s. on X, with conditional probability at least $1 - e^{-t}$,

$$\boldsymbol{\varepsilon}^\top M\boldsymbol{\varepsilon} \leq \sigma^2 \operatorname{tr}(M) + 2\sigma^2\|M\|t + 2\sigma^2\sqrt{\|M\|^2t^2 + \operatorname{tr}(M^2)\,t}.$$

Since $\|C\| \leq \operatorname{tr}(C)$ and $\operatorname{tr}(C^2) \leq \operatorname{tr}(C)^2$, with probability at least $1 - e^{-t}$,

$$\boldsymbol{\varepsilon}^\top C\boldsymbol{\varepsilon} \leq \sigma^2 \operatorname{tr}(C)(2t+1) + 2\sigma^2\sqrt{\operatorname{tr}(C)^2(t^2+t)} \leq (4t+2)\sigma^2 \operatorname{tr}(C).$$

Combining this with Lemma 7, both upper and lower bounds follow from suitable bounds on $\operatorname{tr}(C)$.

Standard normals

The gaussian distribution of the covariates allows the trace of C to be expressed as a function of many standard normal vectors.

Lemma 10. *Consider a covariance operator Σ with $\lambda_i = \mu_i(\Sigma)$ and $\lambda_n > 0$. Write its spectral decomposition $\Sigma = \sum_{j=1}^{\infty} \lambda_j v_j v_j^\top$, where the orthonormal $v_j \in \mathbb{H}$ are the eigenvectors corresponding to the λ_j. Define $z_i = X v_i / \sqrt{\lambda_i}$. Then we can write*

$$\operatorname{tr}(C) = \sum_{i=1}^{\infty} \left[\lambda_i^2 z_i^\top \left(\sum_{j=1}^{\infty} \lambda_j z_j z_j^\top \right)^{-2} z_i \right],$$

and these $z_i \in \mathbb{R}^n$ are independent with distribution $\mathcal{N}(0, I_n)$. Furthermore, for any $i \geq 1$,

$$\lambda_i^2 z_i^\top \left(\sum_{j=1}^{\infty} \lambda_j z_j z_j^\top \right)^{-2} z_i = \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2},$$

where $A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top$.

Proof. Each $x^\top v_j$ has distribution $\mathcal{N}(0, \lambda_j)$, and they are uncorrelated, hence independent. Without loss of generality, we can assume that the rows of X have coordinates corresponding to the v_i, that is, we can think of Σ as diagonal. We can write

$$\begin{aligned}
\operatorname{tr}(C) &= \operatorname{tr}\left(\left(XX^\top\right)^{-1} X \Sigma X^\top \left(XX^\top\right)^{-1} \right) \\
&= \operatorname{tr}\left(\Sigma X^\top \left(XX^\top\right)^{-2} X \right) \\
&= \sum_{i=1}^{\infty} \left[\lambda_i^2 z_i^\top \left(\sum_{i=1}^{\infty} \lambda_i z_i z_i^\top \right)^{-2} z_i \right].
\end{aligned}$$

For the second part, we use Lemma 19, which is a consequence of the Sherman-Woodbury-Morrison formula; see Appendix A.2.

$$\begin{aligned}
\lambda_i^2 z_i^\top \left(\sum_{i=1}^{\infty} \lambda_i z_i z_i^\top \right)^{-2} z_i &= \lambda_i^2 z_i^\top \left(\lambda_i z_i z_i^\top + A_{-i} \right)^{-2} z_i \\
&= \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2},
\end{aligned}$$

by Lemma 19, for the case $k = 1$ and $Z = \sqrt{\lambda_i} z_i$. □

The weighted sum of outer products of these gaussian vectors plays a central role in the rest of the proof. Define

$$A = \sum_{i=1}^{\infty} \lambda_i z_i z_i^\top, \qquad A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top, \qquad A_k = \sum_{i>k} \lambda_i z_i z_i^\top,$$

where the $z_i \in \mathbb{R}^n$ are independent with distribution $\mathcal{N}(0, I_n)$. Note that the vector z_i is independent of the matrix A_{-i}, so the last part of Lemma 10 allows us to write $\operatorname{tr}(C)$ in terms of ratios of quantities involving only independent vectors.

Concentration of A

The next step is to show that eigenvalues of A, A_{-i} and A_k are concentrated. The proof of the following inequality is in Appendix A.3. Recall that $\mu_1(A)$ and $\mu_n(A)$ denote the largest and the smallest eigenvalues of the matrix A.

Lemma 11. *With probability at least* $1 - 2e^{-t}$,

$$\sum_{i=1}^{\infty} \lambda_i - \lozenge \leq \mu_n(A) \leq \mu_1(A) \leq \sum_{i=1}^{\infty} \lambda_i + \lozenge,$$

where

$$\lozenge = \frac{32}{9} \left(\lambda_1 (t + n \ln 9) + \sqrt{(t + n \ln 9) \sum_{i=1}^{\infty} \lambda_i^2} \right).$$

Hence, there is a universal constant c such that with probability at least $1 - 2e^{-n/c}$,

$$\frac{1}{c} \sum_{i=1}^{\infty} \lambda_i - c\lambda_1 n \leq \mu_n(A) \leq \mu_1(A) \leq c \left(\sum_{i=1}^{\infty} \lambda_i + \lambda_1 n \right).$$

The following lemma uses this result to give bounds on the eigenvalues of A_k, which in turn give bounds on some eigenvalues of A_{-i} and A. For these upper and lower bounds to match up to a constant factor, the sum of the eigenvalues of A_k should dominate the term involving its leading eigenvalue, which is a condition on the effective rank $r_k(\Sigma)$.

Lemma 12. *There are universal constants $b, c \geq 1$ such that for any $k \geq 0$, with probability at least $1 - 2e^{-n/c}$,*

1. *for all $i \geq 1$,*

$$\mu_{k+1}(A_{-i}) \leq \mu_{k+1}(A) \leq \mu_1(A_k) \leq c \left(\sum_{j>k} \lambda_j + \lambda_{k+1} n \right),$$

2. for all $1 \le i \le k$,

$$\mu_n(A) \ge \mu_n(A_{-i}) \ge \mu_n(A_k) \ge \frac{1}{c}\sum_{j>k}\lambda_j - c\lambda_{k+1}n,$$

3. if $r_k(\Sigma) \ge bn$, *then*

$$\frac{1}{c}\lambda_{k+1}r_k(\Sigma) \le \mu_n(A_k) \le \mu_1(A_k) \le c\lambda_{k+1}r_k(\Sigma).$$

Proof. By Lemma 11, we know that with probability at least $1 - 2e^{-n/c}$,

$$\frac{1}{c_1}\sum_{j>k}\lambda_j - c_1\lambda_{k+1}n \le \mu_n(A_k) \le \mu_1(A_k) \le c_1\left(\sum_{j>k}\lambda_j + \lambda_{k+1}n\right).$$

First, the matrix $A - A_k$ has rank at most k (as a sum of k matrices of rank 1). Thus, there is a linear space $\mathscr{L}$ of dimension $n-k$ such that for all $v \in \mathscr{L}$, $v^\top A v = v^\top A_k v \le \mu_1(A_k)\|v\|^2$, and so $\mu_{k+1}(A) \le \mu_1(A_k)$.

Second, by the Courant-Fischer-Weyl Theorem, for all i and j, $\mu_j(A_{-i}) \le \mu_j(A)$ (see Lemma 25). On the other hand, for $i \le k$, $A_k \preceq A_{-i}$, so all the eigenvalues of A_{-i} are lower bounded by $\mu_n(A_k)$.

Finally, if $r_k(\Sigma) \ge c_2 n$,

$$\sum_{j>k}\lambda_j + \lambda_{k+1}n = \lambda_{k+1}r_k(\Sigma) + \lambda_{k+1}n \le \left(1 + \frac{1}{c_2}\right)\lambda_{k+1}r_k(\Sigma),$$

$$\frac{1}{c_1}\sum_{j>k}\lambda_j - c_1\lambda_{k+1}n = \frac{1}{c_1}\lambda_{k+1}r_k(\Sigma) - c_1\lambda_{k+1}n \ge \left(\frac{1}{c_1} - \frac{c_1}{c_2}\right)\lambda_{k+1}r_k(\Sigma).$$

Choosing b and c sufficiently large gives the third claim of the lemma. □

Upper bound on the trace term

Lemma 13. *There are universal constants* $b, c \ge 1$ *such that if* $0 \le k \le n/c$, $r_k(\Sigma) \ge bn$, $l \le k$ *then with probability at least* $1 - 6e^{-n/c}$,

$$\operatorname{tr}(C) \le c\left(\frac{l}{n} + n\frac{\sum_{i>l}\lambda_i^2}{\left(\sum_{i>k}\lambda_i\right)^2}\right).$$

The proof uses the following lemma. Its proof is in Appendix A.3.

Lemma 14 (Concentration of a weighted sum of χ^2)**.** *Suppose* $\{\xi_i\}_{i=1}^\infty$ *are i.i.d. random variables distributed as* $\chi^2(1)$. *Then for any* $t \ge 0$ *and any non-negative non-increasing sequence* $\{\lambda_i\}_{i=1}^\infty$ *such that* $\sum_{i=1}^\infty \lambda_i < \infty$, *with probability at least* $1 - 2e^{-t}$,

$$-2\sqrt{t\sum\lambda_i^2} \le \sum_{i=1}^\infty \lambda_i(\xi_i - 1) \le 2\lambda_1 t + 2\sqrt{t\sum\lambda_i^2}.$$

Proof. (of Lemma 13) Fix b to its value in Lemma 12. By Lemma 10,

$$\begin{aligned}\mathrm{tr}(C) &= \sum_{i=1}^{\infty} \lambda_i^2 z_i^\top A^{-2} z_i \\ &= \sum_{i=1}^{l} \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} + \sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i. \end{aligned} \tag{2}$$

First, consider the sum up to l. If $r_k(\Sigma) \geq bn$, Lemma 12 shows that with probability at least $1 - 2e^{-n/c}$, for all $i \leq k$, $\mu_n(A_{-i}) \geq \lambda_{k+1} r_k(\Sigma)/c_1$, and, for all i, $\mu_{k+1}(A_{-i}) \leq c_1 \lambda_{k+1} r_k(\Sigma)$. The lower bounds on the $\mu_n(A_{-i})$'s imply that, for all $z \in \mathbb{R}^n$ and $1 \leq i \leq l$,

$$z^\top A_{-i}^{-2} z \leq \frac{c_1^2 \|z\|^2}{\left(\lambda_{k+1} r_k(\Sigma)\right)^2},$$

and the upper bounds on the $\mu_{k+1}(A_{-i})$'s give

$$z^\top A_{-i}^{-1} z \geq \left(\Pi_{\mathscr{L}_i} z\right)^\top A_{-i}^{-1} \Pi_{\mathscr{L}_i} z \geq \frac{\|\Pi_{\mathscr{L}_i} z\|^2}{c_1 \lambda_{k+1} r_k(\Sigma)},$$

where $\Pi_{\mathscr{L}_i}$ is the orthogonal projection on $\mathscr{L}_i$, the span of the $n-k$ eigenvectors of A_{-i} corresponding to its smallest $n-k$ eigenvalues, using the upper bound on the $(k+1)$-th largest eigenvalue of A_{-i}. So for $i \leq l$,

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \leq \frac{z_i^\top A_{-i}^{-2} z_i}{(z_i^\top A_{-i}^{-1} z_i)^2} \leq c_1^3 \frac{\|z_i\|^2}{\|\Pi_{\mathscr{L}_i} z_i\|^4}. \tag{3}$$

Next, we apply Lemma 14 l times, together with a union bound, to show that with probability at least $1 - 2e^{-t}$, for all $1 \leq i \leq l$,

$$\|z_i\|^2 \leq n + 2(t + \ln k) + 2\sqrt{n(t+\ln k)} \leq c_2 n, \tag{4}$$

$$\|\Pi_{\mathscr{L}_i} z_i\|^2 \geq n - k - 2\sqrt{(t+\ln k)(n-k)} \geq n/c_3, \tag{5}$$

provided that the c in the lemma is sufficiently large. Combining (3), (4), and (5), with probability at least $1 - 4e^{-n/c}$,

$$\sum_{i=1}^{l} \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \leq c_4 \frac{l}{n}.$$

Second, consider the second sum in (2). Lemma 12 shows that, on the same high probability event that we considered in bounding the first half of the sum, $\mu_n(A) \geq \lambda_{k+1} r_k(\Sigma)/c_1$. Hence,

$$\sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i \leq \frac{c_1^2 \sum_{i>l} \lambda_i^2 \|z_i\|^2}{\left(\lambda_{k+1} r_k(\Sigma)\right)^2}.$$

Notice that $\sum_{i>l} \lambda_i^2 \|z_i\|^2$ is a weighted sum of $\chi^2(1)$ random variables, with the weights given by the λ_i^2 in blocks of size n. Lemma 14 implies that, with probability at least $1 - 2e^{-t}$,

$$\begin{aligned}
\sum_{i>l} \lambda_i^2 \|z_i\|^2 &\le n \sum_{i>l} \lambda_i^2 + 2\lambda_{l+1}^2 t + 2\sqrt{tn \sum_{i>l} \lambda_i^4} \\
&\le n \sum_{i>l} \lambda_i^2 + 2\lambda_{l+1}^2 t + 2\sqrt{\lambda_{l+1}^2 tn \sum_{i>l} \lambda_i^2} \\
&\le n \sum_{i>l} \lambda_i^2 + c_5 \lambda_{l+1}^2 t + n \sum_{i>l} \lambda_i^2 \\
&\le c_6 n \sum_{i>l} \lambda_i^2.
\end{aligned}$$

where the third inequality follows from the arithmetic mean-geometric mean inequality. Combining the above gives

$$\sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i \le c_7 n \frac{\sum_{i>l} \lambda_i^2}{\left(\lambda_{k+1} r_k(\Sigma)\right)^2}.$$

□

Lower bound on the trace term

We first give a bound on a single term in $\mathrm{tr}(C)$ that holds regardless of $r_k(\Sigma)$.

Lemma 15. *There is a universal constant $c \ge 1$ such that for any $i \ge 1$, $0 \le k \le n/c$, with probability at least $1 - 4e^{-n/c}$,*

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \ge \frac{1}{cn} \left(1 + \frac{\sum_{j>k} \lambda_j + n\lambda_{k+1}}{n\lambda_i}\right)^{-2}.$$

Proof. Fix $i \ge 1$ and $k \ge 0$. By Lemma 12, with probability at least $1 - 2e^{-n/c}$,

$$\mu_{k+1}(A_{-i}) \le c_1 \left(\sum_{i>k} \lambda_i + \lambda_{k+1} n\right),$$

and hence

$$z_i^\top A_{-i}^{-1} z_i \ge \frac{\|\Pi_{\mathscr{L}_i} z_i\|^2}{c_1 \left(\sum_{i>k} \lambda_i + \lambda_{k+1} n\right)},$$

where $\Pi_{\mathscr{L}_i}$ is the orthogonal projector on the span of the last $n - k$ eigenvectors of A_{-i}. So $\|\Pi_{\mathscr{L}_i} z_i\|^2 \sim \chi^2(n-k)$, independent of A_{-i}. By Lemma 14, with probability at least $1 - 2e^{-t}$,

$$\|\Pi_{\mathscr{L}_i} z_i\|^2 \ge n - k - 2\sqrt{t(n-k)} \ge n/c_2.$$

Thus, with probability at least $1 - 2e^{-n/c}$,

$$z_i^\top A_{-i}^{-1} z_i \geq \frac{n}{c_3 \left(\sum_{i>k} \lambda_i + \lambda_{k+1} n \right)},$$

hence

$$1 + \lambda_i z_i^\top A_{-i}^{-1} z_i \leq \left(\frac{c_3 \left(\sum_{i>k} \lambda_i + \lambda_{k+1} n \right)}{\lambda_i n} + 1 \right) \lambda_i z_i^\top A_{-i}^{-1} z_i.$$

Dividing $z_i^\top A_{-i}^{-2} z_i$ by the square of both sides and multiplying each by λ_i^2, we have

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \geq \left(\frac{c_3 \left(\sum_{i>k} \lambda_i + \lambda_{k+1} n \right)}{\lambda_i n} + 1 \right)^{-2} \frac{z_i^\top A_{-i}^{-2} z_i}{(z_i^\top A_{-i}^{-1} z_i)^2}.$$

Applying the Cauchy-Schwarz inequality and then Lemma 14, we have that with probability at least $1 - 2e^{-t}$,

$$\frac{z_i^\top A_{-i}^{-2} z_i}{(z_i^\top A_{-i}^{-1} z_i)^2} \geq \frac{z_i^\top A_{-i}^{-2} z_i}{\left\| A_{-i}^{-1} z_i \right\|^2 \|z_i\|^2} = \frac{1}{\|z_i\|^2} \geq \frac{1}{n + 2t + 2\sqrt{nt}} \geq \frac{1}{c_4 n}.$$

Choosing c suitably large gives the lemma. □

We can extend these high probability lower bounds to a lower bound on $\operatorname{tr}(C)$ using the following lemma. The proof is in Appendix A.4.

Lemma 16. *Suppose $n \leq \infty$ and $\{\eta_i\}_{i=1}^n$ is a sequence of non-negative random variables, $\{t_i\}_{i=1}^n$ is a sequence of non-negative real numbers (at least one of which is strictly positive) such that for some $\delta \in (0, 1)$ and any natural $i \leq n$,*

$$\mathbb{P}(\eta_i > t_i) \geq 1 - \delta.$$

Then

$$\mathbb{P}\left(\sum_{i=1}^n \eta_i \geq \frac{1}{2} \sum_{i=1}^n t_i \right) \geq 1 - 2\delta.$$

These two lemmas imply the following lower bound.

Lemma 17. *There is a universal constant $c \geq 1$ such that for any $0 \leq k \leq n/c$, and $a \geq 1$, with probability at least $1 - 8e^{-n/c}$,*

1. *If $r_k(\Sigma) < bn$, then $\operatorname{tr}(C) \geq k/(cb^2 n)$.*
2. *If $r_k(\Sigma) \geq bn$, then*

$$\operatorname{tr}(C) \geq \frac{1}{cb^2} \min_{l \leq k} \left(\frac{l}{n} + n \frac{\sum_{i>l} \lambda_i^2}{\left(\lambda_{k+1} r_k(\Sigma) \right)^2} \right).$$

In particular, if all choices of $k \le n/c$ give $r_k(\Sigma) < bn$, then $r_{n/c}(\Sigma) < bn$ implies that with probability at least $1 - 8e^{-n/c}$, $\mathrm{tr}(C) \ge 1/(cb^2)$.

Proof. From Lemmas 10, 15 and 16, with probability at least $1 - 8e^{-n/c}$,

$$\begin{aligned}
\mathrm{tr}(C) &\ge \frac{1}{c_1 n} \sum_i \left(1 + \frac{c_1 \left(\sum_{j>k} \lambda_j + n\lambda_{k+1}\right)}{n\lambda_i}\right)^{-2} \\
&\ge \frac{1}{c_2 n} \sum_i \left(1 + \frac{\sum_{j>k} \lambda_j + n\lambda_{k+1}}{n\lambda_i}\right)^{-2} \\
&\ge \frac{1}{c_2 n} \sum_i \min\left\{1, \frac{n^2\lambda_i^2}{\left(\sum_{j>k} \lambda_j\right)^2}, \frac{\lambda_i^2}{\lambda_{k+1}^2}\right\} \\
&\ge \frac{1}{c_2 b^2 n} \sum_i \min\left\{1, \left(\frac{bn}{r_k(\Sigma)}\right)^2 \frac{\lambda_i^2}{\lambda_{k+1}^2}, \frac{\lambda_i^2}{\lambda_{k+1}^2}\right\}.
\end{aligned}$$

Now, if $r_k(\Sigma) < bn$, then the second term in the minimum is always bigger than the third term, and in that case,

$$\mathrm{tr}(C) \ge \frac{1}{c_2 b^2 n} \sum_i \min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2}\right\} \ge \frac{k+1}{c_2 b^2 n}.$$

On the other hand, if $r_k(\lambda) \ge bn$,

$$\begin{aligned}
\mathrm{tr}(C) &\ge \frac{1}{c_2 b^2} \sum_i \min\left\{\frac{1}{n}, \frac{b^2 n \lambda_i^2}{\left(\lambda_{k+1} r_k(\Sigma)\right)^2}\right\} \\
&= \frac{1}{c_2 b^2} \min_{l \le k} \left(\frac{l}{n} + \frac{b^2 n \sum_{i>l} \lambda_i^2}{\left(\lambda_{k+1} r_k(\Sigma)\right)^2}\right),
\end{aligned}$$

where the equality follows from the fact that the λ_is are non-increasing. □

A simple choice of k

If some $k \le n/c$ has $r_k(\Sigma) \ge bn$, then the upper and lower bounds of Lemmas 13 and 17 are constant multiples of

$$\min_{l \le k} \left(\frac{l}{n} + n \frac{\sum_{i>l} \lambda_i^2}{\left(\lambda_{k+1} r_k(\Sigma)\right)^2}\right).$$

It might seem surprising that any suitable choice of k suffices to give upper and lower bounds: what prevents one choice of k from giving an upper bound that falls below the lower bound that arises from another choice of k? However, the freedom to choose k is somewhat illusory: Lemma 12 shows that, for any

qualifying value of k, the smallest eigenvalue of A is within a constant factor of $\lambda_{k+1} r_k(\Sigma)$. Thus, any two choices of k satisfying $k \leq n/c$ and $r_k(\Sigma) \geq bn$ must have values of $\lambda_{k+1} r_k(\Sigma)$ within constant factors. The smallest such k simplifies the bound on $\mathrm{tr}(C)$, as the following lemma shows.

Lemma 18. *For any $b \geq 1$ and $k^* := \min\{k : r_k(\Sigma) \geq bn\}$, we have*

$$\min_{l \leq k^*} \left(\frac{l}{bn} + \frac{bn \sum_{i>l} \lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}(\Sigma)\right)^2} \right) = \frac{k^*}{bn} + \frac{bn \sum_{i>k^*} \lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}(\Sigma)\right)^2} = \frac{k^*}{bn} + \frac{bn}{R_{k^*}(\Sigma)}.$$

Proof. We can write the function of l being minimized as

$$\begin{aligned}
\frac{l}{bn} + \frac{bn \sum_{i>l} \lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}(\Sigma)\right)^2} &= \sum_{i=1}^{l} \frac{1}{bn} + \sum_{i>l} \frac{bn\lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}(\Sigma)\right)^2} \\
&\geq \sum_{i=1}^{k^*} \min\left\{ \frac{1}{bn}, \frac{bn\lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}(\Sigma)\right)^2} \right\} \\
&\quad + \sum_{i>k^*} \frac{bn\lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}(\Sigma)\right)^2} \\
&= \sum_{i=1}^{l^*} \frac{1}{bn} + \sum_{i>l^*} \frac{bn\lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}(\Sigma)\right)^2},
\end{aligned}$$

where l^* is the largest value of $i \leq k^*$ for which

$$\frac{1}{bn} \leq \frac{bn\lambda_i^2}{\left(\lambda_{k^*+1} r_{k^*}(\Sigma)\right)^2},$$

since the λ_i^2 are non-increasing. This condition is equivalent to

$$\lambda_i \geq \frac{\lambda_{k^*+1} r_{k^*}(\Sigma)}{bn}.$$

The definition of k^* implies $r_{k^*-1}(\Sigma) < bn$. So we can write

$$\begin{aligned}
r_{k^*}(\Sigma) &= \frac{\sum_{i>k^*} \lambda_i}{\lambda_{k^*+1}} \\
&= \frac{\sum_{i>k^*-1} \lambda_i - \lambda_{k^*}}{\lambda_{k^*+1}} \\
&= \frac{\lambda_{k^*}}{\lambda_{k^*+1}} (r_{k^*-1}(\Sigma) - 1) \\
&< \frac{\lambda_{k^*}}{\lambda_{k^*+1}} (bn - 1),
\end{aligned}$$

and so

$$\lambda_{k^*} > \frac{\lambda_{k^*+1} r_{k^*}(\Sigma)}{bn},$$

which implies that the minimizing l is k^*. Also,

$$\frac{\sum_{i>k^*}\lambda_i^2}{\left(\lambda_{k^*+1}r_{k^*}(\Sigma)\right)^2} = \frac{\sum_{i>k^*}\lambda_i^2}{\left(\sum_{i>k^*}\lambda_i\right)^2} = \frac{1}{R_{k^*}(\Sigma)}.$$

□

Setting b in Lemmas 17 and 18 to the universal constant b that suffices for the condition $r_k(\Sigma) \geq bn$ in Lemma 13 and combining these lemmas with Lemma 7 completes the proof of Theorem 4.

5 Conclusions and Further Work

Our results characterize when the phenomenon of benign overfitting occurs in high dimensional linear regression with gaussian data. We give finite sample excess risk bounds that reveal the covariance structure that ensures that the minimum norm interpolating prediction rule has near-optimal prediction accuracy. The characterization depends on two notions of the effective rank of the data covariance operator. It shows that overparameterization, that is, the existence of many low-variance and hence unimportant directions in parameter space, is necessary and sufficient for benign overfitting.

There are several natural future directions. We have extended our results beyond the case of gaussian data, but we would also like to understand how our results extend to other loss functions besides squared error and what we can say about interpolating estimators beyond the minimum norm estimator. One of the most interesting future directions is understanding how these ideas could apply to nonlinearly parameterized function classes such as neural networks, the methodology that uncovered the phenomenon of benign overfitting.

Acknowledgements

We gratefully acknowledge the support of the NSF through grant IIS-1619362 and of Google through a Google Research Award. Part of this work was done as part the Fall 2018 program on Foundations of Data Science at the Simons Institute for the Theory of Computing. Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P and FEDER, EU; "High-dimensional problems in structured probabilistic models - Ayudas Fundación BBVA a Equipos de Investigación Cientifica 2017"; and Google Focused Award "Algorithms and Learning for AI".

References

[1] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized

two-layer neural networks. Technical Report 1901.08584 [cs.LG], arXiv, 2019.

[2] Peter L. Bartlett. Accurate prediction from interpolation: A new challenge for statistical learning theory. Presentation at the National Academy of Sciences workshop, *The Science of Deep Learning,* March 14, 2019. https://youtu.be/1y2sB38T6FU, 2019.

[3] M. Belkin, D. Hsu, and P. Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Proceedings of NIPS 2018*, 2018.

[4] M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? Technical Report 1806.09471, arXiv, 2018.

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. Technical Report 1812.11118, arXiv, 2018.

[6] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. Technical Report 1903.07571 [cs.LG], arXiv, 2019.

[7] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. Technical Report 1802.01396v3 [stat.ML], arXiv, 2018.

[8] C. A. Desoer and B. H. Whalen. A note on pseudoinverses. *Journal of the Society of Industrial and Applied Mathematics*, 11(1):442–446, 1963.

[9] L. Devroye, L. Györfi, and A. Krzyżak. The Hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227, 1998.

[10] Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. Technical Report 1811.03804 [cs.LG], arXiv, 2018.

[11] Simon S. Du, Barnabás Poczós, Xiyu Zhai, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. Technical Report 1810.02054 [cs.LG], arXiv, 2018.

[12] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 1996.

[13] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. Technical Report 1903.08560 [math.ST], arXiv, 2019.

[14] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *Elements of Statistical Learning.* Springer, 2001.

[15] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2018.

[16] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, February 2017.

[17] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. Technical Report 1808.01204 [cs.LG], arXiv, 2018.

[18] Tengyuan Liang and Alexander Rakhlin. Just interpolate: kernel "ridgeless" regression can generalize. Technical Report 1808.00387, arXiv, 2018. To appear in Annals of Statistics.

[19] Vidya Muthukumar, Kailas Vodrahalli, and Anant Sahai. Harmless interpolation of noisy data in regression. Technical Report 1903.09139 [cs.LG], arXiv, 2019.

[20] Stephen Page and Steffen Grünewälder. Ivanov-regularised least-squares estimators over large RKHSs and their interpolation spaces. Technical Report arXiv:1706.03678 [math.ST], ArXiv, June 2017.

[21] Mark Rudelson and Roman Vershynin. Sampling from large matrices: an approach through geometric functional analysis. *Journal of the ACM*, 54(4):21, 2007.

[22] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

[23] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, page 210–268. Cambridge University Press, 2012.

[24] Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. Technical Report 1611.03131 [cs.LG], arXiv, 2016.

[25] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

A Additional Proofs

A.1 Proof of Lemma 8

Note that

$$\left(I - X^\top \left(XX^\top\right)^{-1} X\right) X^\top = X^\top - X^\top \left(XX^\top\right)^{-1} (XX^\top) = 0. \tag{6}$$

Moreover, for any v in the orthogonal complement to the span of the columns of $X^\top$,

$$\left(I - X^\top \left(XX^\top\right)^{-1} X\right) v = v.$$

Thus,

$$\|I - X^\top \left(XX^\top\right)^{-1} X\| \leq 1. \tag{7}$$

Now we can apply (6) to write

$$\begin{aligned}
\theta^{*\top} B \theta^* &= \theta^{*\top} \left(I - X^\top \left(XX^\top\right)^{-1} X\right) \Sigma \left(I - X^\top \left(XX^\top\right)^{-1} X\right) \theta^* \\
&= \theta^{*\top} \left(I - X^\top \left(XX^\top\right)^{-1} X\right) \left(\Sigma - \frac{1}{n} X^\top X\right) \\
&\qquad \left(I - X^\top \left(XX^\top\right)^{-1} X\right) \theta^*.
\end{aligned}$$

Combining with (7) shows that

$$\theta^{*\top} B \theta^* \leq \left\| \Sigma - \frac{1}{n} X^\top X \right\| \|\theta^*\|^2.$$

Thus, due to Corollary 2 of Theorem 6 in [16], there is an absolute constant c such that for any $t > 1$ with probability at least $1 - e^{-t}$,

$$\theta^{*\top} B \theta^* \leq c \|\theta^*\|^2 \|\Sigma\| \max\left\{ \sqrt{\frac{r(\Sigma)}{n}}, \frac{r(\Sigma)}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n} \right\},$$

where

$$r(\Sigma) := \frac{(\mathbb{E}\|x\|)^2}{\|\Sigma\|} \leq \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|} = \frac{1}{\lambda_1} \sum_{i=1}^{\infty} \lambda_i = r_0(\Sigma).$$

A.2 An Algebraic Property

.

Lemma 19. *Suppose $k < n$, $A \in \mathbb{R}^{n \times n}$ is an invertible matrix, and $Z \in \mathbb{R}^{n \times k}$ is such that $ZZ^\top + A$ is invertible. Then*

$$Z^\top (ZZ^\top + A)^{-2} Z = (I + Z^\top A^{-1} Z)^{-1} Z^\top A^{-2} Z (I + Z^\top A^{-1} Z)^{-1}.$$

Proof. We use the Sherman–Morrison–Woodbury formula to write

$$(ZZ^\top + A)^{-1} = A^{-1} - A^{-1}Z(I + Z^\top A^{-1}Z)^{-1}Z^\top A^{-1}. \tag{8}$$

Denote $M_1 := Z^\top A^{-1}Z$ and $M_2 := Z^\top A^{-2}Z$. Applying (8), we get

$$\begin{aligned}
&Z^\top (ZZ^\top + A)^{-2} Z \\
&= Z^\top \Big(A^{-1} - A^{-1}Z(I + Z^\top A^{-1}Z)^{-1}Z^\top A^{-1}\Big)^2 Z \\
&= Z^\top \Big(A^{-1} - A^{-1}Z(I + M_1)^{-1}Z^\top A^{-1}\Big)^2 Z \\
&= Z^\top \Big(A^{-2} - A^{-2}Z(I + M_1)^{-1}Z^\top A^{-1} \\
&\qquad - A^{-1}Z(I + M_1)^{-1}Z^\top A^{-2} \\
&\qquad + A^{-1}Z(I + M_1)^{-1}Z^\top A^{-2}Z(I + M_1)^{-1}Z^\top A^{-1}\Big) Z \\
&= M_2 - M_2(I + M_1)^{-1}M_1 \\
&\quad - M_1(I + M_1)^{-1}M_2 + M_1(I + M_1)^{-1}M_2(I + M_1)^{-1}M_1 \\
&= M_2 - M_2(I + M_1)^{-1}M_1 - M_1(I + M_1)^{-1}M_2(I - (I + M_1)^{-1}M_1) \\
&= M_2(I + M_1)^{-1} - M_1(I + M_1)^{-1}M_2(I + M_1)^{-1} \\
&= (I + M_1)^{-1}M_2(I + M_1)^{-1},
\end{aligned}$$

where we used the identity $I - (I + M_1)^{-1}M_1 = (I + M_1)^{-1}$ twice in the second last equality and the identity $I - M_1(I + M_1)^{-1} = (I + M_1)^{-1}$ in the last equality. □

A.3 Proof of concentration inequalities

We prove the upper and lower bounds of Lemma 14 separately.

Lemma 20 (Upper bound for weighted sum of χ^2)**.** *Suppose $\{\xi_i\}_{i=1}^\infty$ are i.i.d. random variables distributed as $\chi^2(1)$. Then for any $x \geq 0$ and any non-negative non-increasing sequence $\{\lambda_i\}_{i=1}^\infty$ such that $\sum_{i=1}^\infty \lambda_i < \infty$, with probability at least $1 - e^{-x}$,*

$$\sum_{i=1}^\infty \lambda_i(\xi_i - 1) \leq 2\lambda_1 x + 2\sqrt{x\sum \lambda_i^2}.$$

Proof. For any $\tau > 0$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^\infty \lambda_i(\xi_i - 1) > t\right) &\leq e^{-\tau(t + \sum \lambda_i)} \prod_i \mathbb{E}e^{\tau\lambda_i\xi_i} \\
&= \exp\left(-\tau t - \frac{1}{2}\sum_i \left(\ln(1 - 2\lambda_i\tau) + 2\lambda_i\tau\right)\right).
\end{aligned}$$

Also, for $x \in (0,1)$,

$$\begin{aligned} x + \ln(1-x) &= -\int_0^x \int_0^y \frac{1}{(1-z)^2}\, dz\, dy \\ &\geq -\int_0^x \int_0^y \frac{1}{(1-z)^3}\, dz\, dy \\ &= -\frac{x^2}{2(1-x)}. \end{aligned}$$

Hence if $2\lambda_1\tau < 1$ we have

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{\infty} \lambda_i(\xi_i - 1) > t\right) &\leq \exp\left(-\tau t - \frac{1}{2}\sum_i \left(\ln(1-2\lambda_i\tau) + 2\lambda_i\tau\right)\right) \\ &\leq \exp\left(-\tau t + \sum_i \frac{\lambda_i^2\tau^2}{(1-2\lambda_i\tau)}\right) \\ &\leq \exp\left(-\tau t + \frac{\tau^2}{(1-2\lambda_1\tau)}\sum_i \lambda_i^2\right). \end{aligned}$$

Set

$$x = \tau t - \frac{\tau^2}{(1-2\lambda_1\tau)}\sum_i \lambda_i^2;$$

we'll check later that $x > 0$. Then we have

$$t = \frac{\tau}{(1-2\lambda_1\tau)}\sum_i \lambda_i^2 + \frac{x}{\tau} = (\tau^{-1} - 2\lambda_1)^{-1}\sum \lambda_i^2 + x\tau^{-1},$$

and minimizing over $1/\tau$ gives the optimal choice

$$\frac{1}{\tau} = 2\lambda_1 + \sqrt{\frac{\sum \lambda_i^2}{x}}, \tag{9}$$

and substituting gives

$$t = 2\lambda_1 x + 2\sqrt{x\sum \lambda_i^2}. \tag{10}$$

To see that these choices are valid, fix $x > 0$, define τ by (9) and t by (10), and observe that this implies $2\lambda_1\tau < 1$ and so

$$\mathbb{P}\left(\sum_{i=1}^{\infty} \lambda_i(\xi_i - 1) > 2\lambda_1 x + 2\sqrt{x\sum \lambda_i^2}\right) \leq \exp(-x).$$

□

Lemma 21 (Lower bound for weighted sum of χ^2)**.** *Suppose $\{\xi_i\}_{i=1}^{\infty}$ are i.i.d. random variables distributed as $\chi^2(1)$. Then for any $x \geq 0$ and any non-negative non-increasing sequence $\{\lambda_i\}_{i=1}^{\infty}$ such that $\sum_{i=1}^{\infty}\lambda_i < \infty$, with probability at least $1 - e^{-x}$,*

$$\sum_{i=1}^{\infty} \lambda_i(\xi_i - 1) \geq -2\sqrt{x\sum \lambda_i^2}.$$

Proof. For any $\tau > 0$,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{\infty}\lambda_i(\xi_i-1) < -t\right) &\le e^{\tau(-t+\sum\lambda_i)}\prod_i \mathbb{E}e^{-\tau\lambda_i\xi_i} \\ &= \exp\left(-\tau t - \frac{1}{2}\sum_i\left(\ln(1+2\lambda_i\tau) - 2\tau\lambda_i\right)\right).\end{aligned}$$

For $x > 0$,

$$\begin{aligned}\ln(1+x) - x &= -\int_0^x\int_0^y \frac{1}{(1+z)^2}\,dz\,dy \\ &\ge -\int_0^x\int_0^y 1\,dt\,dy \\ &= -\frac{x^2}{2}.\end{aligned}$$

Hence we have

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{\infty}\lambda_i(\xi_i-1) < -t\right) &\le \exp\left(-\tau t - \frac{1}{2}\sum_i\left(\ln(1+2\lambda_i\tau) - 2\tau\lambda_i\right)\right) \\ &\le \exp\left(-\tau t + \tau^2\sum_i\lambda_i^2\right).\end{aligned}$$

Set $x = \tau t - \tau^2\sum_i\lambda_i^2$, so that $t = \tau\sum_i\lambda_i^2 + x/\tau$. Optimizing gives $\tau = \sqrt{x/\sum\lambda_i^2}$, and hence $t = 2\sqrt{x\sum\lambda_i^2}$. As in the previous proof, we can check that this gives a valid choice by fixing a non-negative x and confirming that substituting these choices of τ and t gives $\tau t - \tau^2\sum_i\lambda_i^2 = 2x - x = x$. □

Lemma 22 (ϵ-net argument)**.** *Suppose $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, and $\mathcal{N}_\epsilon$ is an ϵ-net on the unit sphere $\mathcal{S}^{n-1}$ in the Euclidean norm, where $\epsilon < \frac{1}{2}$. Then*

$$\|A\| \le (1-\epsilon)^{-2}\max_{x\in\mathcal{N}_\epsilon}|x^\top Ax|.$$

Proof. Denote the eigenvalues of A as $\lambda_1,\ldots,\lambda_n$ and assume $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n|$. Denote the first eigenvector of A as $v \in \mathcal{S}^{n-1}$, and take $\Delta v \in \mathbb{R}^n$ such that $v + \Delta v \in \mathcal{N}_\epsilon$ and $\|\Delta v\| \le \epsilon$. Denote the coordinates of Δv in the eigenbasis of A

as $\Delta v_1, \ldots, \Delta v_n$. Now we can write

$$\begin{aligned}
\left|(v+\Delta v)^\top A(v+\Delta v)\right| &= \left|\lambda_1 + 2\lambda_1 \Delta v_1 + \sum_{i=1}^n \lambda_i \Delta v_i^2\right| \\
&= |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 + \sum_{i=2}^n \frac{\lambda_i}{\lambda_1} \Delta v_i^2\right| \\
&\geq |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 - \sum_{i=2}^n \Delta v_i^2\right| \\
&= |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 - \|\Delta v\|^2 + \Delta v_1^2\right| \\
&= |\lambda_1| \cdot \left|1 + 2\left(\Delta v_1 + \Delta v_1^2\right) - \|\Delta v\|^2\right| \\
&\geq |\lambda_1| \cdot \left|1 + 2\left(-\|\Delta v\| + (-\|\Delta v\|)^2\right) - \|\Delta v\|^2\right| \\
&= |\lambda_1| \cdot \left|1 - 2\|\Delta v\| + \|\Delta v\|^2\right| \\
&\geq |\lambda_1| \cdot |1 - 2\epsilon + \epsilon^2| \\
&= \|A\|(1-\epsilon)^2,
\end{aligned}$$

where the first inequality holds because the λ_is are decreasing in magnitude, and the last two inequalities hold since the functions $x + x^2$ and $2x + x^2$ are both increasing on $(-\frac{1}{2}, \infty)$ and $\Delta v_1 \geq -\|\Delta v\| \geq -\epsilon \geq -\frac{1}{2}$. □

We restate Lemma 11.

Lemma 23. *With probability at least* $1 - 2e^{-t}$,

$$\sum_{i=1}^\infty \lambda_i - \Diamond \leq \mu_n(A) \leq \mu_1(A) \leq \sum_{i=1}^\infty \lambda_i + \Diamond,$$

where

$$\Diamond = \frac{32}{9}\left(\lambda_1(t + n\ln 9) + \sqrt{(t+n\ln 9)\sum_{i=1}^\infty \lambda_i^2}\right).$$

Hence, there is a universal constant c such that if $t < n/c$, then with probability at least $1 - 2e^{-t}$,

$$\frac{1}{c}\sum_{i=1}^\infty \lambda_i - c\lambda_1 n \leq \mu_n(A) \leq \mu_1(A) \leq c\left(\sum_{i=1}^\infty \lambda_i + \lambda_1 n\right).$$

Proof. For a fixed unit vector $v \in \mathbb{R}^n$, Lemmas 20 and 21 imply that with probability at least $1 - 2e^{-t}$,

$$\left|v^\top A v - \sum \lambda_i\right| \leq 2\left(\lambda_1 t + \sqrt{t\sum \lambda_i^2}\right).$$

Let $\mathcal{N}$ be a $\frac{1}{4}$-net on the sphere $\mathcal{S}^{n-1}$ with respect to the Euclidean distance such that $|\mathcal{N}| \leq 9^n$. Applying the union bound over the elements of $\mathcal{N}$, we see that with probability $1 - 2e^{-t}$, every $v \in \mathcal{N}$ satisfies

$$\left| v^\top A v - \sum \lambda_i \right| \leq 2 \left(\lambda_1 (t + n \ln 9) + \sqrt{(t + n \ln 9) \sum_{i=1}^{\infty} \lambda_i^2} \right).$$

Since $\mathcal{N}$ is a $\frac{1}{4}$-net, by Lemma 22, we need to multiply the quantity above by $(1 - 1/4)^{-2}$ to get the bound on the norm of the $A - I_n \sum_{i=1}^{\infty} \lambda_i$. Thus, with probability at least $1 - 2e^{-t}$,

$$\left\| A - I_n \sum_{i=1}^{\infty} \lambda_i \right\| \leq \Diamond.$$

When $t < n/c_1$ we can write $t + n \ln 9 \leq c_2 n$, and we have

$$\begin{aligned}
\Diamond &= \frac{32}{9} \left(\lambda_1 (t + n \ln 9) + \sqrt{(t + n \ln 9) \sum_{i=1}^{\infty} \lambda_i^2} \right) \\
&\leq c_3 \left(\lambda_1 n + \sqrt{n \sum_{i=1}^{\infty} \lambda_i^2} \right) \\
&\leq c_3 \lambda_1 n + \sqrt{(c_3^2 \lambda_1 n) \sum_{i=1}^{\infty} \lambda_i} \\
&\leq c_4 \lambda_1 n + \frac{1}{2} \sum_{i=1}^{\infty} \lambda_i,
\end{aligned}$$

by the AMGM inequality. (Recall that $c_1, c_2, \ldots$ denote universal constants with value at least 1.) □

A.4 Proof of Lemma 16

We know that, for all $i \leq n$,

$$\mathbb{P}(\eta_i > t_i) \geq 1 - \delta.$$

Consider the following event:

$$E = \left\{ \sum_{i=1}^{n} \eta_i < \frac{1}{2} \sum_{i=1}^{n} t_i \right\}$$

and denote its probability as $c\delta$ for some $c \in (0, \delta^{-1})$.

On the one hand, by the definition of the event, we have

$$\frac{1}{\mathbb{P}(E)}\mathbb{E}\left[1_E \sum_{i=1}^{n} \eta_i\right] \leq \frac{1}{2}\sum_{i=1}^{n} t_i.$$

On the other hand, note that for any i ,

$$\begin{aligned}\mathbb{E}[\eta_i 1_E] \geq & \mathbb{E}[t_i 1_{\{\eta_i \geq t_i\} \cap E}] \\ & = t_i \mathbb{P}(\{\eta_i \geq t_i\} \cap E) \\ & \geq t_i(\mathbb{P}\{\eta_i \geq t_i\} + \mathbb{P}(E) - 1) \\ & \geq t_i(c-1)\delta.\end{aligned}$$

So

$$\begin{aligned}\mathbb{E}\left[1_E \sum_{i=1}^{n} \eta_i\right] &\geq (c-1)\delta \sum_{i=1}^{n} t_i, \\ \frac{1}{\mathbb{P}(E)}\mathbb{E}\left[1_E \sum_{i=1}^{n} \eta_i\right] &\geq (1-c^{-1}) \sum_{i=1}^{n} t_i.\end{aligned}$$

Thus, we obtain

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^{n} t_i &\geq (1-c^{-1}) \sum_{i=1}^{n} t_i, \\ c &\leq 2, \\ \mathbb{P}\left(\sum_{i=1}^{n} \eta_i < \frac{1}{2}\sum_{i=1}^{n} t_i\right) &= c\delta \leq 2\delta.\end{aligned}$$

A.5 Eigenvalue monotonicity

Recall (half of) the Courant-Fischer-Weyl theorem.

Lemma 24. *For any symmetric $n \times n$ matrix A, and any $i \in [n]$, $\mu_i(A)$ is the minimum, over all subspaces U of $\mathbb{R}^n$ of dimension $n-i$, of the maximum, over all unit-length $u \in U$, of $u^\top A u$.*

Lemma 25 (Monotonicity of eigenvalues)**.** *If symmetric matrices A and B satisfy $A \preceq B$, then, for any $i \in [n]$, we have $\mu_i(A) \leq \mu_i(B)$.*

Proof. Let U be the subspace of $\mathbb{R}^n$ of dimension $n-i$ that minimizes the maximum over all unit-length $u \in U$, of $u^\top A u$, and let V be the analogous

subspace for B. We have

$$\begin{aligned}\mu_i(A) &= \max_{u \in U : ||u||=1} u^\top A u \quad \text{(by Lemma 24)} \\ &\le \max_{v \in V : ||v||=1} v^\top A v \quad \text{(since } U \text{ is the minimizer)} \\ &\le \max_{v \in V : ||v||=1} v^\top B v \quad \text{(since } A \preceq B\text{)} \\ &= \mu_i(B),\end{aligned}$$

by Lemma 24, completing the proof. □

A.6 Rank facts

The quantity $r_0(\Sigma)$ is an important complexity parameter for covariance estimation problems, where it has been called the 'effective rank' [23, 16]. Earlier, $r_0(\Sigma^2)$ was called the 'stable rank' [22] and the 'numerical rank' [21], although that term has a different meaning in computational linear algebra [12, p261].

We restate Lemma 5.

Lemma 26. *$r_k(\Sigma) \ge 1$, $r_k^2(\Sigma) = r_k(\Sigma^2) R_k(\Sigma)$, and*

$$r_k(\Sigma^2) \le r_k(\Sigma) \le R_k(\Sigma) \le r_k^2(\Sigma).$$

Proof. The first inequality and the equality are immediate from the definitions. For the second inequality,

$$r_k(\Sigma^2) = \frac{\sum_{i>k} \lambda_i^2}{\lambda_{k+1}^2} \le \frac{\lambda_{k+1} \sum_{i>k} \lambda_i}{\lambda_{k+1}^2} = r_k(\Sigma).$$

The last two inequalities follow. □

Lemma 27. *Writing r_k and R_k for $r_k(\Sigma)$ and $R_k(\Sigma)$,*

$$\frac{1}{R_{k+1}} = \frac{\frac{1}{R_k} - \frac{1}{r_k^2}}{1 - \left(2 - \frac{1}{r_k}\right)\frac{1}{r_k}}.$$

Thus, the function $\phi(k) = k/(b^2 n) + n/R_k$ satisfies the monotonicity property $\phi(k+1) > \phi(k)$ whenever $r_k > bn \ge 1$.

Proof. Writing

$$q = \sum_{i>k+1} \lambda_i^2, \qquad\qquad s = \sum_{i>k+1} \lambda_i,$$

so that $R_{k+1} = s^2/q$, we have

$$\begin{aligned}
\frac{1}{R_k} - \frac{1}{R_{k+1}} &= \frac{\lambda_{k+1}^2 + q}{(\lambda_{k+1} + s)^2} - \frac{q}{s^2} \\
&= \frac{\left(\lambda_{k+1}^2 + q\right) s^2 - q\left(\lambda_{k+1} + s\right)^2}{s^2 \left(\lambda_{k+1} + s\right)^2} \\
&= \frac{1}{r_k^2} - \frac{q\lambda_{k+1}\left(\lambda_{k+1} + 2s\right)}{s^2 \left(\lambda_{k+1} + s\right)^2} \\
&= \frac{1}{r_k^2} - \frac{2\left(\lambda_{k+1} + s\right) - \lambda_{k+1}}{R_{k+1} r_k \left(\lambda_{k+1} + s\right)} \\
&= \frac{1}{r_k^2} - \frac{2 - 1/r_k}{R_{k+1} r_k}.
\end{aligned}$$

Hence

$$\frac{1}{R_{k+1}} = \frac{1/R_k - 1/r_k^2}{1 - \left(2 - \frac{1}{r_k}\right)\frac{1}{r_k}}.$$

Since $r_k > 1$, $0 < 1 - (2 - 1/r_k)/r_k < 1$, so

$$\frac{n}{R_{k+1}} > \frac{n}{R_k} - \frac{n}{r_k^2},$$

and if $r_k > bn$,

$$\begin{aligned}
\phi(k+1) - \phi(k) &= \frac{k+1}{b^2 n} + \frac{n}{R_{k+1}} - \left(\frac{k}{b^2 n} + \frac{n}{R_k}\right) \\
&> \frac{1}{b^2 n} - \frac{n}{r_k^2} \\
&> 0.
\end{aligned}$$

□

A.7 Conditions on eigenvalues

In this section, we prove Theorem 6. We build up the proof in stages. First, we characterize those sequences of effective ranks that can arise.

Theorem 28. *Consider some positive summable sequence $\{\lambda_i\}_{i=1}^{\infty}$, and for any non-negative integer i denote*

$$r_i := \lambda_{i+1}^{-1} \sum_{j>i} \lambda_j.$$

Then $r_i > 1$ and

$$\sum_{i=0}^{\infty} r_i^{-1} = \infty.$$

Moreover, for any positive sequence $\{u_i\}$ such that $\sum_{i=0}^{\infty} u_i^{-1} = \infty$ and for every i $u_i > 1$, there exists a positive sequence $\{\lambda_i\}$ (unique up to constant multiplier) such that $r_i \equiv u_i$. The sequence is (a constant rescaling of)

$$\lambda_k = u_{k-1}^{-1} \prod_{i=0}^{k-2} (1 - u_i^{-1}).$$

Proof.

$$\begin{aligned}\sum_{i=k+1}^{\infty} \lambda_i =& \sum_{i=k}^{\infty} \lambda_i - \lambda_k \\ =& (1 - r_{k-1}^{-1}) \sum_{i=k}^{\infty} \lambda_i.\end{aligned}$$

Thus,

$$\sum_{i=k+1}^{\infty} \lambda_i = \prod_{i=0}^{k-1} \left(1 - r_i^{-1}\right) \cdot \sum_{i=1}^{\infty} \lambda_i,$$

which goes to zero if and only if $\sum_{i=0}^{\infty} r_i^{-1} = \infty$.

On the other hand, we may rewrite the first equality in the proof as

$$\lambda_{k+1} r_k = \lambda_k r_{k-1} (1 - r_{k-1}^{-1}),$$

and hence

$$\lambda_k r_{k-1} = \prod_{i=0}^{k-2} \left(1 - r_i^{-1}\right) \lambda_1 r_0.$$

So for any sequence $\{u_i\}$ we can uniquely (up to a constant multiplier) recover the sequence $\{\lambda_i\}$ such that $r_i = u_i$ — the only candidate is

$$\lambda_k = u_{k-1}^{-1} \prod_{i=0}^{k-2} (1 - u_i^{-1}).$$

However, for such $\{\lambda_i\}$ one can compute

$$\sum_{i=1}^{k} \lambda_i = 1 - \prod_{i=0}^{k-1} (1 - u_i^{-1}),$$

so the resulting sequence $\{\lambda_i\}$ sums to 1, and

$$r_k = \lambda_{k+1}^{-1} \sum_{i>k} \lambda_i = \lambda_{k+1}^{-1} \prod_{i=0}^{k-1} (1 - u_i^{-1}) = u_k.$$

□

Theorem 29. *Suppose b is some constant, and $k^*(n) = \min\{k : r_k \geq bn\}$. Suppose also that the sequence $\{r_n\}$ is increasing. Then, as n goes to infinity, $k^*(n)/n$ goes to zero if and only if r_n/n goes to infinity.*

Proof. We prove the "if" part separately from the "only if" part.

1. **If $k^*(n)/n \to 0$ then $r_n/n \to \infty$.**

   Fix some $C > 1$. Since $k^*(n)/n \to 0$, there exists some N_C such that for any $n \geq N_C$, $k^*(n) < n/C$. Thus, for all $n > N_C$,

   $$k^*(\lfloor Cn \rfloor) \leq n,$$
   $$r_n \geq r_{k^*(\lfloor Cn \rfloor)} \geq b\lfloor Cn \rfloor.$$

   Since the constant C is arbitrary, r_n/n goes to infinity.

2. **If $r_n/n \to \infty$ then $k^*(n)/n \to 0$.**

   Fix some constant $C > 1$. Since $r_n/n \to \infty$ there exists some N_C such that for any $n \geq N_C$, $r_n > Cn$. Thus, for any $n > CN_C/b$

   $$r_{\lceil nb/C \rceil} \geq bn,$$
   $$k^*(n) \leq \lceil nb/C \rceil.$$

   Since the constant C is arbitrary, $k^*(n)/n$ goes to zero.

□

Theorem 30. *Suppose the sequence $\{r_i\}$ is increasing and $r_n/n \to \infty$ as $n \to \infty$. Then a sufficient condition for $\frac{n}{R_{k^*(n)}} \to 0$ is*

$$r_k^{-2} = o(r_k^{-1} - r_{k+1}^{-1}) \text{ as } k \to \infty.$$

For example, this condition holds for $r_n = n \log n$.

Proof. We need to show that

$$\frac{n}{R_{k^*(n)}} = \frac{n \sum_{i>k^*(n)} \lambda_i^2}{\left(\sum_{i>k^*(n)} \lambda_i\right)^2} = \frac{n \sum_{i>k^*(n)} \lambda_i^2}{\lambda_{k^*(n)+1}^2 r_{k^*(n)}^2} \to 0.$$

Since $r_{k^*(n)} \geq bn$ and $\lim_{n\to\infty} k^*(n) = \infty$, it is enough to prove that $\frac{\sum_{i>k} \lambda_i^2}{\lambda_{k+1}^2 r_k} \to 0$ as k goes to infinity. Since

$$\lambda_{k+2} r_{k+1} = \lambda_{k+1} r_k (1 - r_k^{-1}),$$

we can write that

$$\lambda_{k+1+l} r_{k+l} = \lambda_{k+1} r_k \prod_{i=k}^{k+l-1} (1 - r_i^{-1}) \leq \lambda_{k+1} r_k \exp\left(-\sum_{i=k}^{k+l-1} r_i^{-1}\right)$$

which yields

$$\frac{\lambda_{k+1+l}}{\lambda_{k+1} r_k} \leq r_{k+l}^{-1} \exp\left(- \sum_{i=k}^{k+l-1} r_i^{-1} \right).$$

Thus, we obtain

$$\frac{\sum_{i>k} \lambda_i^2}{\lambda_{k+1}^2 r_k} \leq r_k \sum_{i \geq k} r_i^{-2} \exp\left(-2 \sum_{j=k}^{i-1} r_j^{-1} \right),$$

and it is sufficient to prove that the latter quantity goes to zero. We write

$$\begin{aligned} r_k \sum_{i \geq k} r_i^{-2} \exp\left(-2 \sum_{j=k}^{i-1} r_j^{-1} \right) &= \frac{\sum_{i \geq k} r_i^{-2} \exp\left(-2 \sum_{j=k}^{i-1} r_j^{-1} \right)}{r_k^{-1}} \\ &= \frac{\sum_{i \geq k} r_i^{-2} \exp\left(-2 \sum_{j=0}^{i-1} r_j^{-1} \right)}{r_k^{-1} \exp\left(-2 \sum_{j=0}^{k-1} r_j^{-1} \right)}. \end{aligned}$$

Since both numerator and denominator are decreasing in k and go to zero as $k \to \infty$, we can apply the Stolz–Cesáro theorem (an analog of L'Hôpital's rule for discrete sequences):

$$\begin{aligned} &\lim_{k \to \infty} \frac{\sum_{i \geq k} r_i^{-2} \exp\left(-2 \sum_{j=0}^{i-1} r_j^{-1} \right)}{r_k^{-1} \exp\left(-2 \sum_{j=0}^{k-1} r_j^{-1} \right)} \\ &= \lim_{k \to \infty} \frac{r_k^{-2} \exp\left(-2 \sum_{j=0}^{k-1} r_j^{-1} \right)}{(r_k^{-1} - e^{-2 r_k^{-1}} r_{k+1}^{-1}) \exp\left(-2 \sum_{j=0}^{k-1} r_j^{-1} \right)} \\ &= \lim_{k \to \infty} \frac{r_k^{-2}}{(r_k^{-1} - e^{-2 r_k^{-1}} r_{k+1}^{-1})} \\ &\quad \text{(since for } k \text{ large enough } e^{-2 r_k^{-1}} \leq 1 - r_k^{-1}) \\ &\leq \lim_{k \to \infty} \frac{r_k^{-2}}{r_k^{-1} - r_{k+1}^{-1} + r_k^{-1} r_{k+1}^{-1}} \\ &= 0 \end{aligned}$$

where the last line is due to our sufficient condition. □

Now we are ready to prove Theorem 6.

Part 1, if direction, first term: We have

$$||\Sigma_n|| \sqrt{r_0(\Sigma_n)} = \sqrt{\lambda_1 \sum_{i=1}^{\infty} \lambda_i} = O\left(\sqrt{\sum_{i=1}^{\infty} \frac{1}{i \log^{\beta}(1+i)}} \right),$$

which is $O(1)$ for $\beta > 1$.

Part 1, if direction, second term: By Theorem 29, it suffices to prove that $\lim_{n\to\infty} \frac{r_n}{n} = \infty$. This holds because

$$r_n = \frac{\sum_{i>n} \frac{1}{i \log^\beta(1+i)}}{\frac{1}{(n+1)\log^\beta(2+n)}} = \Theta(n \log n),$$

since $\beta > 1$.

Part 1, if direction, third term: By Theorem 30, it suffices to prove that $r_k^{-2} = o(r_k^{-1} - r_{k+1}^{-1})$, that is

$$\lim_{k\to\infty} \frac{r_k^{-2}}{r_k^{-1} - r_{k+1}^{-1}} = 0$$

or, equivalently,

$$\lim_{k\to\infty} \frac{r_{k+1}}{r_k(r_{k+1} - r_k)} = 0.$$

As argued above, when $\alpha = 1$ and $\beta > 1$, $r_k = \Theta(k \log k)$, so it suffices to show that $\lim_{k\to\infty}(r_{k+1} - r_k) = \infty$. We have

$$\begin{aligned} r_{k+1} - r_k &= \frac{\sum_{i>k+1} \lambda_i}{\lambda_{k+2}} - \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \\ &= \frac{\left((\lambda_{k+1} - \lambda_{k+2}) \sum_{i>k+1} \lambda_i\right) - \lambda_{k+1}\lambda_{k+2}}{\lambda_{k+1}\lambda_{k+2}} \\ &= \left(\left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \sum_{i>k+1} \lambda_i\right) - 1 \end{aligned}$$

so it suffices to show that

$$\lim_{k\to\infty} \left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \sum_{i>k+1} \lambda_i = \infty.$$

Since λ_i is non-increasing, we have

$$\begin{aligned} &\left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \sum_{i>k+1} \lambda_i \\ &\geq \left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \int_{k+1}^{\infty} \frac{1}{x \log^\beta x}\, dx \\ &= \left(\frac{1}{\lambda_{k+2}} - \frac{1}{\lambda_{k+1}}\right) \frac{1}{(\beta-1)\log^{\beta-1}(k+1)} \\ &= \frac{(k+2)\log^\beta(k+3) - (k+1)\log^\beta(k+2)}{(\beta-1)\log^{\beta-1}(k+1)}. \end{aligned}$$

If we define f on the positive reals by $f(x) = x\log^\beta(x+1)$, then f is convex, and, since $f'(x) = \frac{\beta x \log^{\beta-1}(x+1)}{x+1} + \log^\beta(x+1)$, we have

$$\frac{(k+2)\log^\beta(k+3) - (k+1)\log^\beta(k+2)}{(\beta-1)\log^{\beta-1}(k+1)}$$
$$\geq \frac{\frac{\beta(k+1)\log^{\beta-1}(k+2)}{k+2} + \log^\beta(k+2)}{(\beta-1)\log^{\beta-1}(k+1)},$$

which goes to infinity for large k, completing the proof of the "if" direction of the third term of Part 1.

Part 1, only if direction, $\alpha > 1$: If $\alpha > 1$, then

$$\begin{aligned} r_n &= \frac{\sum_{i>n} \frac{1}{i^a \log^\beta(1+i)}}{\frac{1}{n^a \log^\beta(1+n)}} \\ &\leq n^\alpha \sum_{i>n} \frac{\log^\beta(1+n)}{i^a \log^\beta(1+i)} \\ &\leq n^\alpha \sum_{i>n} \frac{1}{i^a} \\ &= n^\alpha O(n^{1-\alpha}), \end{aligned}$$

which does not grow faster than n. Thus, by Theorem 29, $k^*(n)/n$ does not go to zero.

Part 1, only if direction, $\alpha < 1$, or $\alpha = 1$ and $\beta \leq 1$: In this case, since, as above

$$||\Sigma_n||\sqrt{r_0(\Sigma_n)} \geq \sqrt{\sum_{i=1}^{\infty} \lambda_i},$$

and $\sum_{i=1}^{\infty} \frac{1}{i^\alpha \log^\beta(1+i)}$ diverges in this case, $\frac{||\Sigma_n||\sqrt{r_0(\Sigma_n)}}{n}$ does not go to zero.

Before starting on Part 2, let us define $r_{k,n} = r_k(\Sigma_n)$ and $R_{k,n} = R_k(\Sigma_n)$.

Part 2, if direction, first term: We have

$$||\Sigma_n||\sqrt{r_{0,n}} = \sqrt{\lambda_{1,n}\sum_{i=1}^{\infty}\lambda_{i.n}} = \sqrt{\sum_{i=1}^{\infty}\frac{1}{i^{1+\alpha_n}}} \leq \sqrt{1+\frac{1}{\alpha_n}}$$

so $||\Sigma_n||\sqrt{\frac{r_{0,n}}{n}} \leq \sqrt{\frac{1+\frac{1}{\alpha_n}}{n}}$ which goes to zero with n if $\alpha_n = \omega(1/n)$.

Part 2, if direction, second term: First,

$$\begin{aligned} r_{k,n} &= (k+1)^{1+\alpha_n} \sum_{i>k} i^{-(1+\alpha_n)} \\ &\geq (k+1)^{1+\alpha_n} \int_{k+1}^{\infty} x^{-(1+\alpha_n)} dx \\ &= \frac{k+1}{\alpha_n}. \end{aligned}$$

Thus, $k^*(n) = O(\alpha_n n)$, so that $\frac{k^*(n)}{n} = O(\alpha_n) = o(1)$.
Part 2, if direction, third term: We bound $R_{k,n}$ from below by separately bounding its numerator and denominator:

$$\begin{aligned} \sum_{i>k} i^{-(1+\alpha_n)} &\geq \int_{k+1}^{\infty} x^{-(1+\alpha_n)}\, dx \\ &= \frac{1}{\alpha_n (k+1)^{\alpha_n}} \end{aligned}$$

and

$$\begin{aligned} \sum_{i>k} i^{-2(1+\alpha_n)} &\leq \int_{k}^{\infty} x^{-2(1+\alpha_n)}\, dx \\ &= \frac{1}{k^{1+2\alpha_n}(2\alpha_n+1)} \end{aligned}$$

so that

$$R_{k,n} \geq \frac{k^{1+2\alpha_n}(2\alpha_n+1)}{\alpha_n^2 (k+1)^{2\alpha_n}} \geq \frac{k}{\alpha_n^2} \times \left(1 - \frac{1}{k+1}\right)^{2\alpha_n}. \tag{11}$$

So now we want a lower bound on $k^*(n)$. For that, we need an upper bound on $r_{k,n}$, and

$$\begin{aligned} r_{k,n} &\leq (k+1)^{1+\alpha_n} \int_{k}^{\infty} x^{-(1+\alpha_n)} dx \\ &= \frac{(k+1)}{\alpha_n} \times \left(1 + \frac{1}{k}\right)^{\alpha_n} \\ &\leq \frac{2k}{\alpha_n} e^{\alpha_n/k}. \end{aligned}$$

This implies $\frac{2k^*(n)}{\alpha_n} e^{\alpha_n/k^*(n)} \geq bn$. This, together with the fact that, for $u > 1$, $ue^{1/u}$ is an increasing function of u, implies that, for large enough n, $k^*(n) \geq \alpha_n bn/3$. Since $\alpha_n = \omega(1/n)$, this implies that $k^*(n) = \omega(1)$. Combining this with (11), for large enough n

$$R_{k^*(n),n} \geq \frac{k^*(n)}{\alpha_n^2} e^{-\alpha_n/k^*(n)} \geq \frac{k^*(n)}{2\alpha_n^2} \geq \frac{bn}{6\alpha_n}.$$

Thus $n/R_{k^*(n),n} = O(\alpha_n) = o(1)$.
Part 2, only if direction, $\alpha_n = O(1/n)$: We have

$$||\Sigma_n||\sqrt{r_{0,n}} = \sqrt{\sum_{i=1}^{\infty} \frac{1}{i^{1+\alpha_n}}} \geq \sqrt{\frac{1}{\alpha_n}}$$

so $||\Sigma_n||\sqrt{\frac{r_{0,n}}{n}} \geq \sqrt{\frac{1}{\alpha_n n}}$, which is bounded below by a constant for large n if $\alpha_n = O(1/n)$.
Part 2, only if direction, $\alpha_n = \Omega(1)$: Recall that, if the proof of the "if" direction of the third term, we showed that $k^*(n) \geq \alpha_n bn/3$. This implies that $\frac{k^*(n)}{n} = \Omega(\alpha_n)$.
Part 3: Suppose that Σ_n is benign. Then because $R_k(\Sigma_n) \leq p_n - k$, we must have $p_n = \omega(n)$. Thus, we can restrict our attention to the sequences for which $p_n = \omega(n)$ and find the necessary and sufficient conditions for that class.

Next, for any positive α and any natural $k \in [1, p_n)$, we can write

$$\int_k^{p_n} x^{-\alpha}\, dx \geq \sum_{i=k+1}^{p_n} i^{-\alpha} \geq \int_{k+1}^{p_n} x^{-\alpha}\, dx,$$

$$F(p_n) - F(k) \geq \sum_{i=k+1}^{p_n} i^{-\alpha} \geq F(p_n) - F(k+1),$$

where

$$F(x) = \begin{cases} \frac{1}{1-\alpha} x^{1-\alpha}, & \text{for } \alpha \neq 1, \\ \ln(x), & \text{for } \alpha = 1. \end{cases}$$

As the sequence can only be benign if $k^* = o(n)$, we can only consider values of k, that don't exceed some constant fraction of n, e.g. $n/2$.

Since $p_n = \omega(n)$, noting that, for $x > 0$, the sign of $\frac{1}{1-\alpha} x^{1-\alpha}$ flips when α crosses 1, we can write, uniformly for all $k \in [1, n/2]$,

$$\sum_{i=k+1}^{p_n} i^{-\alpha} = \begin{cases} \Theta_\alpha\left(p_n^{1-\alpha}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(\ln(p_n/k)\right), & \text{for } \alpha = 1, \\ \Theta_\alpha\left(k^{1-\alpha}\right), & \text{for } \alpha > 1. \end{cases}$$

Recall that we consider $\lambda_{i,n} = i^{-\alpha}$ for $i \leq p_n$. Using the formula above, we get uniformly for all $k \in [1, n/2]$

$$r_k(\Sigma_n) = \begin{cases} \Theta_\alpha\left(k^\alpha p_n^{1-\alpha}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(k \ln(p_n/k)\right), & \text{for } \alpha = 1, \\ \Theta_\alpha\left(k\right), & \text{for } \alpha > 1. \end{cases}$$

Recall that $k^* = \min\{k : r_k(\Sigma_n) \geq bn\}$. We compute

$$k^* = \begin{cases} \Theta_\alpha\left(p_n^{1-\frac{1}{\alpha}} n^{\frac{1}{\alpha}}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(\frac{n}{\ln(p_n/n)}\right), & \text{for } \alpha = 1, \\ \Theta_\alpha\left(n\right), & \text{for } \alpha > 1. \end{cases}$$

One can see that for $\alpha > 1$, $k^* = \Omega_\alpha(n)$, so the sequence is not benign for $\alpha > 1$. On the other hand, $k^* = o(n)$ for $\alpha \leq 1$.

Next, analogously to the asymptotics for $r_k(\Sigma)$, we have

$$r_k(\Sigma_n^2) = \begin{cases} \Theta_\alpha\left(k^{2\alpha} p_n^{1-2\alpha}\right), & \text{for } \alpha \in (0, 0.5), \\ \Theta_\alpha\left(k \ln(p_n/k)\right), & \text{for } \alpha = 0.5, \\ \Theta_\alpha\left(k\right), & \text{for } \alpha \in (0.5, 1]. \end{cases}$$

Since $R_k = \frac{r_k(\Sigma)^2}{r_k(\Sigma^2)}$, we can write uniformly for all $k \in [1, n/2]$

$$R_k = \begin{cases} \Theta_\alpha\left(p_n\right), & \text{for } \alpha \in (0, 0.5), \\ \Theta_\alpha\left(\frac{p_n}{\ln(p_n/k)}\right), & \text{for } \alpha = 0.5, \\ \Theta_\alpha\left(k^{2\alpha-1} p_n^{2-2\alpha}\right), & \text{for } \alpha \in (0.5, 1), \\ \Theta_\alpha\left(\ln(p_n/k)^2\right), & \text{for } \alpha = 1. \end{cases}$$

Now we plug in k^* instead of k. Recall that $p_n/k^* = \Theta_\alpha\left((p_n/n)^{1/\alpha}\right)$ for $\alpha \in (0,1)$, and $p_n/k^* = \Theta_\alpha\left(p_n/n \ln(p_n/n)\right)$ for $\alpha = 1$. We get

$$R_{k^*} = \begin{cases} \Theta_\alpha\left(p_n\right), & \text{for } \alpha \in (0, 0.5), \\ \Theta_\alpha\left(n \frac{p_n/n}{\ln(p_n/n)}\right), & \text{for } \alpha = 0.5, \\ \Theta_\alpha\left(n \left(\frac{p_n}{n}\right)^{\frac{1}{\alpha}-1}\right), & \text{for } \alpha \in (0.5, 1), \\ \Theta_\alpha\left(\ln(p_n/n)^2\right), & \text{for } \alpha = 1. \end{cases}$$

Since $p_n = \omega(n)$, for any $\alpha \in (0,1)$, $R_{k^*} = \omega(n)$. For $\alpha = 1$ the necessary and sufficient for $R_{k^*} = \omega(n)$ is $\ln(p_n/n) = \omega(\sqrt{n})$.

So far, we obtained the necessary and sufficient conditions for the variance term to go to zero. Now let's look at the upper bound for the bias: since $\lambda_{1,n} \equiv 1$, we just need $r_0/n \to 0$. We write, for $\alpha \in (0,1]$,

$$r_0 = \sum_{i=1}^{p_n} i^{-\alpha} = \begin{cases} \Theta_\alpha\left(p_n^{1-\alpha}\right), & \text{for } \alpha \in (0,1), \\ \Theta_\alpha\left(\ln p_n\right), & \text{for } \alpha = 1. \end{cases}$$

Thus, for $\alpha < 1$, $r_0(\Sigma_n)/n$ goes to zero if and only if $p_n = o\left(n^{1/(1-\alpha)}\right)$, and for $\alpha = 1$, $r_0(\Sigma_n)/n$ goes to zero if and only if $\ln(p_n) = o(n)$.

Part 4: Suppose that Σ_n is benign. Then because $R_k(\Sigma_n) \leq p_n - k$, we must have $p_n = \omega(n)$. Also,

$$\begin{aligned} \operatorname{tr}(\Sigma_n) &= \Theta\left(1 - e^{-p_n} + p_n \epsilon_n\right) \\ &= \Theta\left(1 + p_n \epsilon_n\right), \end{aligned}$$

and so $p_n\epsilon_n = o(n)$. Since Σ_n benign implies $k^* = o(n)$, and hence $k^* = o(p_n)$, we consider $k = o(p_n)$. In this regime,

$$\begin{aligned}\sum_{i>k}\lambda_i &= \Theta\left(e^{-k} - e^{-p_n} + (p_n - k)\epsilon_n\right)\\ &\leq \Theta\left(e^{-k} + p_n\epsilon_n\right).\end{aligned}$$

Thus, whenever $k \leq p_n$,

$$r_k(\Sigma_n) \leq \Theta\left(\frac{e^{-k} + p_n\epsilon_n}{e^{-k} + \epsilon_n}\right).$$

Notice that

$$\frac{d}{dx}\frac{x + p_n\epsilon_n}{x + \epsilon_n} = \frac{\epsilon_n - p_n\epsilon_n}{(x + \epsilon_n)^2} < 0,$$

so k^* must be large enough to make

$$\frac{e^{-k} + p_n\epsilon_n}{e^{-k} + \epsilon_n} = \Omega(n).$$

Substituting $k = \ln(n/(p_n\epsilon_n)) - \ln c$ gives

$$\begin{aligned}r_k(\Sigma_n) &\leq \Theta\left(\frac{cp_n\epsilon_n/n + p_n\epsilon_n}{cp_n\epsilon_n/n + \epsilon_n}\right)\\ &= \Theta\left(\frac{p_n\epsilon_n}{cp_n\epsilon_n/n}\right)\\ &= \Theta\left(cn\right),\end{aligned}$$

which shows that $k^* \geq \ln(n/(p_n\epsilon_n)) - O(1)$. Thus, if Σ_n is benign, we must have $k^* = o(n)$, that is, $\epsilon_n p_n = ne^{-o(n)}$.

Conversely, assume $p_n = \Omega(n)$ and $\epsilon_n p_n = ne^{-o(n)}$ (that is, $\ln(n/(p_n\epsilon_n)) = o(n)$). Set $k = \ln(n/(p_n\epsilon_n)) - c$, for some c, which we shall see is $\Theta(1)$. Notice that $k = o(n)$, so $p_n - k = \Omega(p_n)$ and $e^{-p_n} = o(e^{-k})$. Thus,

$$\begin{aligned}\sum_{i>k}\lambda_i &= \Theta\left(e^{-k} - e^{-p_n} + (p_n - k)\epsilon_n\right)\\ &= \Theta\left(e^{-k} + p_n\epsilon_n\right),\\ \sum_{i>k}\lambda_i^2 &= \Theta\left(e^{-2k} - e^{-2p_n} + (p_n - k)\epsilon_n^2\right)\\ &= \Theta\left(e^{-2k} + p_n\epsilon_n^2\right).\end{aligned}$$

These imply

$$\begin{aligned}
\operatorname{tr}(\Sigma_n) &= \Theta(\epsilon_n p_n + 1), \\
r_k(\Sigma_n) &= \Theta\left(\frac{e^{-k} + p_n\epsilon_n}{e^{-k} + \epsilon_n}\right) \\
&= \Theta\left(\frac{cp_n\epsilon_n/n + p_n\epsilon_n}{cp_n\epsilon_n/n + \epsilon_n}\right) \\
&= \Theta\left(\frac{p_n\epsilon_n}{cp_n\epsilon_n/n}\right) \\
&= \Theta\left(cn\right),
\end{aligned}$$

which shows that $k^* = \ln(n/(p_n\epsilon_n)) + O(1)$. Also, we have

$$\begin{aligned}
R_k(\Sigma_n) &= \Theta\left(\frac{\left(e^{-k} + p_n\epsilon_n\right)^2}{e^{-2k} + p_n\epsilon_n^2}\right) \\
&= \Theta\left(\frac{\left(p_n\epsilon_n/n + p_n\epsilon_n\right)^2}{p_n^2\epsilon_n^2/n^2 + p_n\epsilon_n^2}\right) \\
&= \Theta\left(\frac{p_n^2\epsilon_n^2}{p_n^2\epsilon_n^2/n^2 + p_n\epsilon_n^2}\right) \\
&= \Theta\left(\min\left\{n^2, p_n\right\}\right).
\end{aligned}$$

Combining gives

$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n + 1}{n} + \frac{\ln(n/(\epsilon_n p_n))}{n} + \max\left\{\frac{1}{n}, \frac{n}{p_n}\right\}\right).$$

Now, it is clear that $p_n = \omega(n)$, $\epsilon_n p_n = o(n)$, and $\epsilon_n p_n = ne^{-o(n)}$ imply that Σ_n is benign.