Journal of Machine Learning Research 24 (2023) 1-76 Submitted 12/22; Revised 3/23; Published 3/23

Benign overfitting in ridge regression

Alexander Tsigler ALEXANDER_TSIGLER@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
367 Evans Hall, Berkeley, CA 94720-3860

Peter L. Bartlett PETER@BERKELEY.EDU
Departments of Statistics and Computer Science
University of California, Berkeley and Google Research
367 Evans Hall, Berkeley, CA 94720-3860

Editor: Daniel Roy

Abstract

In many modern applications of deep learning the neural network has many more parameters than the data points used for its training. Motivated by those practices, a large body of recent theoretical research has been devoted to studying overparameterized models. One of the central phenomena in this regime is the ability of the model to interpolate noisy data, but still have test error lower than the amount of noise in that data. Bartlett et al. (2020) characterized for which covariance structure of the data such a phenomenon can happen in linear regression if one considers the interpolating solution with minimum ℓ_2-norm and the data has independent components: they gave a sharp bound on the variance term and showed that it can be small if and only if the data covariance has high effective rank in a subspace of small co-dimension. We strengthen and complete their results by eliminating the independence assumption and providing sharp bounds for the bias term. Thus, our results apply in a much more general setting than those of Bartlett et al. (2020), e.g., kernel regression, and not only characterize how the noise is damped but also which part of the true signal is learned. Moreover, we extend the result to the setting of ridge regression, which allows us to explain another interesting phenomenon: we give general sufficient conditions under which the optimal regularization is negative.

Keywords: ridge regression, overparameterization, interpolation, generalization, concentration inequalities, high-dimensional probability.

1. Introduction

1.1 Motivation and our contribution

The bias-variance tradeoff is well known in statistics and machine learning. The classical theory suggests that large models overfit the data and that one needs significant regularization to make them generalize. This intuition is, however, in contrast with the empirical study of modern machine learning techniques. It was repeatedly observed that even models with enough capacity to exactly interpolate the data can generalize with little regularization, or no regularization at all (Belkin et al., 2019a; Zhang et al., 2016). In some cases, the best value of the regularizer can be zero (Liang and Rakhlin, 2018) or even negative (Kobak et al., 2020) for such models.

The aim of this paper is to provide a theoretical understanding of these phenomena, and to do that we consider one of the simplest settings in which they can be observed—ridge regression in dimension p with $n < p$ i.i.d. noisy observations. Despite being a classical statistical methodology, ridge regression and its ridgeless limit are still not completely understood in such a regime: when $n < p$ classical theory suggests that the regularization parameter should be large enough to provide additional capacity control (see, e.g., Hsu et al. (2014) and references therein). The basis of our work was set by Bartlett et al. (2020), who studied the variance term for ridgeless regression with $n < p$ under the additional assumption that the data vectors have independent components. The main discovery of their work is that the variance term can be small if and only if there exists $k^* \ll n$ such that if one removes the first k^* largest eigenvalues of the covariance operator, the remaining tail of the sequence of eigenvalues has large effective rank compared to n. In our work we start afresh and use the same separation of eigendirections from the very beginning, which allows us to substitute the independence assumption by a weaker assumption on the condition number of the Gram matrix of the tails of the data vectors. Moreover, we show how the same separation of the eigenvalues gives tight bounds for the bias term too. Finally, by virtue of algebra, our argument extends very easily to the setting of ridge regression, which allows for comparison with the above mentioned classical results and investigation of the case when the regularization is even less than zero. We show that we extend (with different constants) the results of Hsu et al. (2014) to a larger range of regularization parameters, and give general conditions under which negative regularization is optimal and can provide arbitrarily high multiplicative gain in excess risk.

The structure of the paper is the following: in Section 1.2, we provide an overview of the field of overparameterized ridge regression. We postpone a more technical overview to Section 9, where we also explain how our paper relates to other works. We start the presentation of our results with introducing the setting of ridge regression in Section 2. After that, we use Section 3 to introduce the separation of eigendirections and define the relevant important objects: Subsection 3.1 shows two simple sketches aimed at building up intuition, Subsection 3.2 explains the results of Bartlett et al. (2020) in terms of that intuition and Subsection 3.3 explains how our work completes the story. The aim of this discussion is to elucidate the meaning behind the rigorous assumptions and results that we show in Section 4. Then Section 5 provides a more technical discussion of the main assumption. Section 6 provides an outline of the proof and explains where it uses the assumption that the data is sub-Gaussian. In Section 7 we note that as a side product of the proof an alternative form of the main bound arises, which makes it convenient to compare our bounds to the results of other papers. In Section 8, we derive the sufficient conditions for optimality of negative regularization. Finally, we conclude the paper with Section 10.

1.2 Related work

Motivated by the empirical success of overparametrized models, there has recently been a flurry of work aimed at understanding theoretically whether the corresponding effects can be seen in overparametrized linear regression; see, e.g., (Liang et al., 2019; Muthukumar et al., 2019; Belkin et al., 2019b; Bibas et al., 2019; Nakkiran, 2019; Xu and Hsu, 2019; Zhou et al., 2021; Negrea et al., 2020) and other references in this section.

The results that aim at characterizing the generalization performance of linear methods can be split roughly into three categories. The first category is results that give exact expressions of the excess risk in the asymptotic setting with ambient dimension and the number of data points going

to infinity, while their ratio goes to a constant, and the spectral density of the covariance operator converges weakly to some limiting distribution (Dobriban and Wager, 2015; Hastie et al., 2019; Wu and Xu, 2020; Richards et al., 2020).

The second category is results that make strong assumptions on the distribution of data (e.g., that data vectors have i.i.d. components or come from a uniform distribution on a sphere) and derive bounds on excess risk of linear regression with some specific features, or kernel regression with a kernel that has some specific properties (Montanari and Zhong, 2020; Ghorbani et al., 2020b; Mei and Montanari, 2019; Ghorbani et al., 2020a; Liang et al., 2020). Some of these results are also asymptotic, and some are non-asymptotic.

The third category is results that prove non-asymptotic bounds depending on the arbitrary structure of the covariance of the data. This is the category to which this paper belongs. We already mentioned the work of Bartlett et al. (2020). The other works in this category are (Kobak et al., 2020), (Chinot and Lerasle, 2021), (Dereziński et al., 2019) and (Dereziński et al., 2020).

We provide more detailed comparison and discuss some technical aspects in Section 9.

There have been many related works since the arXiv version of this paper (Tsigler and Bartlett, 2020) was posted (Mei et al., 2021a,b; Ghosh et al., 2021; Misiakiewicz and Mei, 2021; Bartlett et al., 2021; Celentano et al., 2021; Muthukumar et al., 2021; Narang et al., 2021; McRae et al., 2021; Shamir, 2022; Koehler et al., 2021; Bunea et al., 2022; Cheng and Montanari, 2022) etc. Hastie et al. (2020) obtained a finite sample version of the asymptotic results of the old version of their paper (Hastie et al., 2019). In Section 7.3 we provide an explicit comparison with our results. More recently, Mei et al. (2021a) obtained generalization bounds for kernel ridge regression under similar assumptions to those we consider here (see their Assumption 1). Koehler et al. (2021) used the idea of separating the firs k eigendirections of the covariance to study excess risk of minimum norm interpolators with arbitrary norms and Gaussian data. Bartlett et al. (2021) obtained results which belong to the intersection of the first and the second categories which we described in Section 1.2 (see their Theorem 4.1). Shamir (2022) constructed an example of a misspecified setting (i.e., the noise is not independent from the data) in which our results don't hold even though the condition number of the matrix A_k is a constant (see their Example 1).

2. Ridge regression setup

The learning problem we consider is ridge regression. Its goal is to learn an unknown real-valued function on $\mathbb{R}^p$ given noisy observations of its values in n points. We operate in the overparameterized regime, i.e., $p > n$.

2.1 Covariate model

We assume that the data set consists of n i.i.d. vectors sampled from some distribution on $\mathbb{R}^p$, whose mean is zero. Throughout the paper x denotes an independent draw from that distribution. Denote $X \in \mathbb{R}^{n\times p}$ to be the matrix whose rows are the (transposed) data vectors.

Our results depend on the spectrum of the covariance matrix $\Sigma = \mathbb{E}[xx^\top]$. We fix an orthonormal basis in which Σ is diagonal:

$$\Sigma = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_p), \tag{1}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ is the non-increasing sequence of eigenvalues of Σ.

We assume sub-Gaussianity: denote $Z := X\Sigma^{-1/2}$ (whitened data matrix). Rows of Z are isotropic centered i.i.d. random vectors. We assume that rows of Z are sub-Gaussian with sub-Gaussian norm σ_x as defined in Appendix A.1.

Sub-Gaussianity is a classical assumption, which provides a convenient framework for controlling deviations of various quantities of interest (see Vershynin (2018) for an introduction). We discuss whether it is actually needed in Section 6.4.

2.2 Response model

Denote $y \in \mathbb{R}^n$ to be the vector whose coordinates are noisy measurements of the values of an unknown function in the corresponding data points. We assume that the true function is linear with coefficients $\theta^* \in \mathbb{R}^p$, i.e.,

$$y = X\theta^* + \varepsilon,$$

where ε is the noise vector. We assume that components of ε are i.i.d. centered random variables with variance v_ε^2.

2.3 Learning procedure

Ridge regression with regularization parameter λ is a classical learning algorithm that estimates θ^* from X, y according to the following formula:

$$\hat{\theta}(y) := X^\top (XX^\top + \lambda I_n)^{-1} y.$$

See Appendix B for a discussion. The matrix $\lambda I_n + XX^\top$ will play an important role in our analysis, so we denote

$$A := \lambda I_n + XX^\top.$$

In the ridgeless case ($\lambda = 0$), A is the Gram matrix of the data. Ridge regularization shifts all its eigenvalues by λ.

2.4 Excess risk and its bias-variance decomposition

The quantity of interest is excess risk that we define in the following way: recall that x is a new data point from the same distribution as rows of X. The error that our predictor incurs on this data point is $x^\top(\hat{\theta}(y) - \theta^*)$. We define excess risk as the average squared error over the population, i.e.,

$$\mathbb{E}_x \left[(x^\top(\hat{\theta}(y) - \theta^*))^2 \right] = \|\hat{\theta}(y) - \theta^*\|_\Sigma^2,$$

where we define $\|x\|_M := \sqrt{x^\top M x}$ for any positive semi-definite (PSD) matrix M and any vector x of the corresponding dimension.

Note that $\hat{\theta}(y)$ is linear in y, which allows us to write

$$\begin{gathered}
\hat{\theta}(y) = \hat{\theta}(X\theta^*) + \hat{\theta}(\varepsilon), \\
\mathbb{E}_\varepsilon \left[\|\hat{\theta}(y) - \theta^*\|_\Sigma^2 \right] = \|\hat{\theta}(X\theta^*) - \theta^*\|_\Sigma^2 + \mathbb{E}_\varepsilon \left[\|\hat{\theta}(\varepsilon)\|_\Sigma^2 \right], \\
\|\hat{\theta}(y) - \theta^*\|_\Sigma^2 \le 2(\|\hat{\theta}(X\theta^*) - \theta^*\|_\Sigma^2 + \|\hat{\theta}(\varepsilon)\|_\Sigma^2).
\end{gathered}$$

The term $\|\hat{\theta}(X\theta^*) - \theta^*\|_\Sigma^2$ is the error in the noiseless regime; it is caused by rows of X not spanning the whole space and by regularization. The term $\|\hat{\theta}(\varepsilon)\|_\Sigma^2$ is the error of learning the zero function from pure noise. One can see that these two terms nicely decouple from each other and can be studied separately. Moreover, note that $\|\hat{\theta}(\varepsilon)\|_\Sigma^2$ is a quadratic form in ε. Its expectation scales linearly with v_ε^2 (variance of the noise):

$$\mathbb{E}_\varepsilon \left[\|\hat{\theta}(\varepsilon)\|_\Sigma^2\right] = v_\varepsilon^2 \operatorname{tr}(A^{-1}X\Sigma X^\top A^{-1}).$$

If the noise is sub-Gaussian with sub-Gaussian norm σ_ε, then by Lemma 22 from the appendix for some absolute constant c and any $t > 1$, with probability at least $1 - ce^{-n/c}$,

$$\begin{aligned}\|\hat{\theta}(\varepsilon)\|_\Sigma^2 =& \varepsilon^\top A^{-1}X\Sigma X^\top A^{-1}\varepsilon \\ \leq& ct\sigma_\varepsilon^2 \operatorname{tr}(A^{-1}X\Sigma X^\top A^{-1}).\end{aligned}$$

Therefore, both expectation and deviations of the term $\|\hat{\theta}(\varepsilon)\|_\Sigma^2$ are controlled by the quantity $\operatorname{tr}(A^{-1}X\Sigma X^\top A^{-1})$. Thus, we define:

$$\begin{aligned} B :=&\quad \|\hat{\theta}(X\theta^*) - \theta^*\|_\Sigma^2 &=&\quad \|(X^\top A^{-1}X - I_p)\theta^*\|_\Sigma^2 &&\text{— bias,} \\ V :=&\quad \mathbb{E}_\varepsilon\left[\|\hat{\theta}(\varepsilon)\|_\Sigma^2/v_\varepsilon^2\right] &=&\quad \operatorname{tr}(A^{-1}X\Sigma X^\top A^{-1}) &&\text{— variance.}\end{aligned} \tag{2}$$

These quantities don't depend on the distribution of the noise. The goal of this paper is to provide sharp non-asymptotic bounds for them.

3. The story of separating the first k eigendirections and our contribution

3.1 Essentially high-dimensional linear regression vs. essentially low-dimensional

Before we present our results, we develop some intuition by considering two easy scenarios: "essentially low-dimensional" and "essentially high-dimensional". For each scenario we will do an informal computation of the excess risk and give a geometric interpretation.

- **Essentially low-dimensional linear regression.** Consider least squares regression in which data lives in $\mathbb{R}^k$ and $k \ll n$: $X \in \mathbb{R}^{n\times k}$ with i.i.d. centered rows from a distribution with covariance $\Sigma \in \mathbb{R}^{k\times k}$ and $y = X\theta^* + \varepsilon$, where ε has i.i.d. centered components with variances v_ε^2. Our estimator of choice in this regime is OLS:

  $$\hat{\theta} = \arg\min_\theta \|X\theta - y\|^2 = \arg\min_\theta \|X(\theta - \theta^*) - \varepsilon\|^2$$

  As θ takes all possible values in $\mathbb{R}^k$, $X(\theta - \theta^*)$ takes all possible values in the span of columns of X, which means that

  $$X(\hat{\theta} - \theta^*) = \Pi_X \varepsilon,$$

  where Π_X is the projection on the span of columns of X. This allows us to write the following informal computation, which leads to the classical k/n rate:

  $$v_\varepsilon^2 k = \mathbb{E}_\varepsilon\|\Pi_X\varepsilon\|^2 = \mathbb{E}_\varepsilon\|X(\hat{\theta} - \theta^*)\|^2 = \mathbb{E}_\varepsilon\left[(\hat{\theta} - \theta^*)^\top \underbrace{X^\top X}_{\approx n\Sigma}(\hat{\theta} - \theta^*)\right],$$

$$v_\varepsilon^2 \cdot k/n \approx (\hat{\theta} - \theta^*)^\top \Sigma (\hat{\theta} - \theta^*) = \mathbb{E}_{x \sim \mathcal{N}(0,\Sigma)} \langle x, \hat{\theta} - \theta^* \rangle^2.$$

Here we used the informal transition $\|n^{-1} X^\top X - \Sigma\| \approx 0$ — the population covariance matrix is well-approximated by the sample covariance matrix uniformly in all directions. If $k \ll n$ this results holds with very few additional assumptions (see (Tikhomirov, 2017) and references therein).

What we have obtained is an example of a classical argument: the training error $\|X(\hat{\theta} - \theta^*)\|^2$ is a good proxy for the population error $\|\Sigma^{1/2}(\hat{\theta} - \theta^*)\|^2$ uniformly over all $\hat{\theta} \in \mathbb{R}^k$, and the model helps eliminate the noise because it gets projected on a subspace of low dimension. The larger the model, the more error comes from the noise.

Such a result leads to a classical bias-variance trade-off: the larger the model is, the better it can approximate the true dependence, but also the more noise it picks up. A classical cartoon is shown in Figure 1: Figures 1a–1c show the result of performing least squares regression with features $\{\cos(m\pi x)\}_{m=0}^p$. As the number of features grows, the ability of the model to approximate the signal grows too, but at the cost of increasing sensitivity to the noise. As the number of features approaches the number of data points (the "interpolation threshold"), this leads to overfitting.

- **Essentially high-dimensional linear regression.** Now consider linear regression in which $p \gg n$, but with isotropic data: assume that the matrix X has i.i.d. standard normal elements and $y = X\theta_* + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0_n, v_\varepsilon^2 I_n)$ — independent from X. We consider the minimum ℓ_2-norm interpolating solution:

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^p : X\theta = y} \|\theta\| = X^\top (XX^\top)^{-1} y = X^\top (XX^\top)^{-1} (X\theta^* + \varepsilon).$$

According to our definitions of bias and variance from Equation (2) with $\lambda = 0$,

$$\begin{aligned} B =& \|\big(I_p - X^\top (XX^\top)^{-1} X\big)\theta^*\|, \\ V =& \mathbb{E}_\varepsilon \|X^\top (XX^\top)^{-1} \varepsilon\|^2 / v_\varepsilon^2 = \operatorname{tr}\Big(\big(\underbrace{XX^\top}_{\approx p I_n}\big)^{-1}\Big). \end{aligned}$$

Here we see the following: the matrix $X^\top (XX^\top)^{-1} X$ is the projection on the span of the data. This is a random n-dimensional subspace in p-dimensional space. Thus, with high probability $\|X^\top (XX^\top)^{-1} X\theta^*\|^2 / \|\theta^*\|^2 \approx n/p$, so the projection only preserves an n/p fraction of the energy of the signal. When it comes to the variance term, we can use the same concentration result for the sample covariance as we did in the low-dimensional case, but for the transposed data matrix, meaning $XX^\top \approx pI_n$. Finishing the computation yields

$$B \approx (1 - n/p)\|\theta^*\|^2, \quad \mathbb{E}_\varepsilon V \approx n/p.$$

We see that the signal is almost not learned at all in this regime (the bias term is close to the full energy of the signal), but the noise is also damped by the factor p/n.

The geometric interpretation is as follows: if $p \gg n$, the span of n data points is almost orthogonal to θ^* with high probability. The data just does not measure θ^* in most directions,

so almost the whole signal is lost. On the other hand, despite the noise fully propagating into in-sample predictions, a new data point x is also almost orthogonal to all the old data points with high probability, so those noisy predictions don't influence the prediction in x. Overall, despite interpolating the data, we effectively learn a zero estimate out of sample. The zero estimator can be a very good estimator, e.g., if the true signal is zero. This hints at the possibility of learning via high-dimensional interpolation: the model can use the directions in which the signal is not learned to smear the noise over them.

The learning cartoon for this regime is given in Figures 1d–1e: as the number of cosine features becomes large compared to the number of data points, the learning procedure predicts zero out of sample, despite interpolating the values in sample. However, if we add multiplicative weights to the cosine features, down-weighting higher frequencies, it causes the minimum norm solution to learn the low frequency signal and interpolate the noise using the high frequency components.

3.2 The ground provided by the previous work

Bartlett et al. (2020) studied the variance term for ridgeless regression under the additional assumption that the data vectors have independent components. To give an overview of their results, introduce the following quantities: for any $k \in \{0, 1, 2, \dots, p-1\}$ define

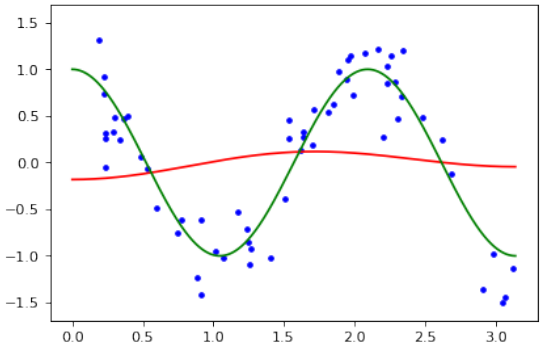
$$r_k := \frac{1}{\lambda_{k+1}} \left(\lambda + \sum_{i>k} \lambda_i \right), \quad \rho_k := r_k / n.$$

In the ridgeless setting, meaning $\lambda = 0$, r_0 is a well-known effective rank of the matrix Σ, r_k is the effective rank of the same matrix, but after restricting it to the span of its last $p-k$ eigenvectors, and ρ_k measures how large that effective rank is compared to the number of data points.

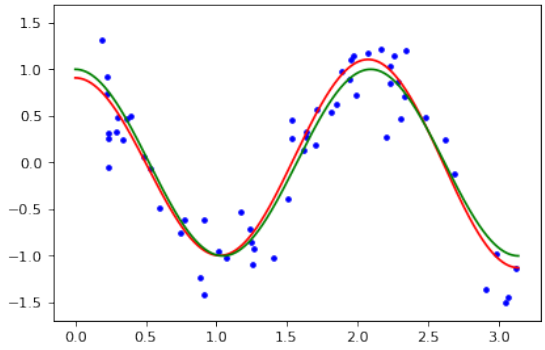
Given this notation, Bartlett et al. (2020) defined k^* as the minimum k for which ρ_k is larger than a universal constant. Their result is then that if such a k^* doesn't exist or if k^*/n is at least a constant, then V is lower bounded by a constant. Otherwise, they show that with high probability V is equal up to a constant factor to the following quantity:

$$\frac{k^*}{n} + \frac{n \sum_{i>k^*} \lambda_i^2}{\left(\sum_{i>k^*} \lambda_i \right)^2}.$$

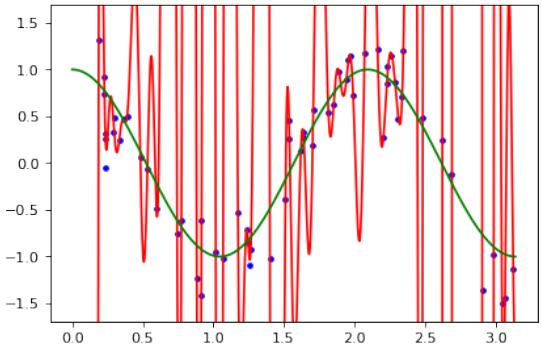
Inspection of the proof shows that the "essentially low-dimensional" rate k/n comes from the first k components of the vector $\hat{\theta}(\varepsilon)$ and the term $\left(n \sum_{i>k^*} \lambda_i^2\right) / \left(\sum_{i>k^*} \lambda_i\right)^2$ comes from the rest of the components of $\hat{\theta}(\varepsilon)$. Note that if one plugs in $\lambda_i = \lambda_j$ for all $i, j > k^*$, then it becomes $\left(n \sum_{i>k^*} \lambda_i^2\right) / \left(\sum_{i>k^*} \lambda_i\right)^2 = n/(p-k^*)$ — exactly the variance term of the "essentially high-dimensional" regime of Section 3.1. The conclusion of Bartlett et al. (2020) is therefore that the only way that an interpolating solution can damp the noise by more than a constant factor is the following: the data is such that after removing k components, it becomes "essentially high-dimensional", meaning that the effective rank of its covariance is large compared to the number of data points. After that the variance in the first k components is the same as for the classical least squares, and the variance in the rest of the components corresponds to the "essentially high-dimensional" case, where you cannot learn but the noise is still damped. Note, however, that the story was not complete because only the variance term was bounded sharply in that work.



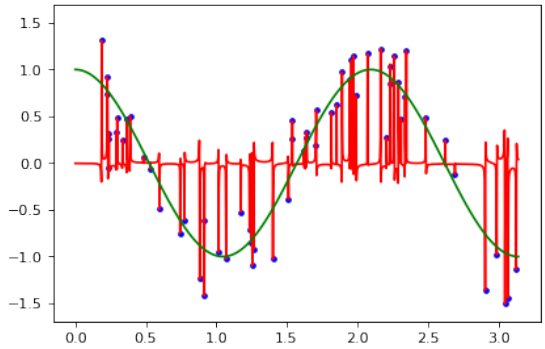
(a) Features $\{\cos(mx)\}_{m=1}^{2}$: underfitting. A linear combination of features cannot approximate the true dependence.



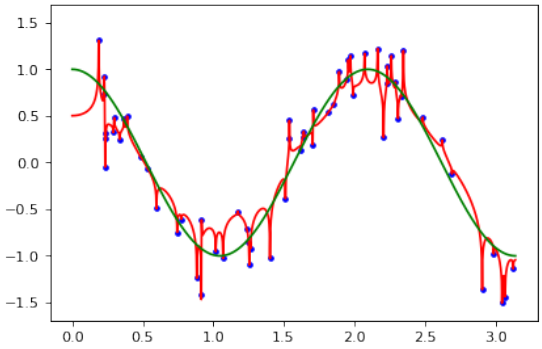
(b) Features $\{\cos(mx)\}_{m=1}^{3}$: the best fit. This is the minimum number of features that span the true dependence.



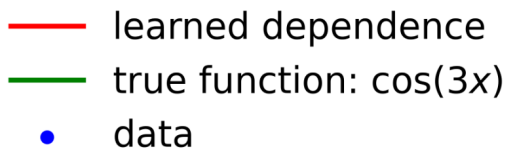
(c) Features $\{\cos(mx)\}_{m=1}^{50}$: overfitting. As the number of features approaches the number of data points, the effect of the noise becomes stronger.



(d) Features $\{\cos(mx)\}_{m=1}^{2000}$: isotropic overparameterization. As the number of cosine features grows above the interpolation threshold, the learned solution goes to zero out of sample.



(e) Features $\{\cos(mx)/m\}_{m=1}^{2000}$: benign overfitting. Adding weights to cosine features results in interpolating the noise with high frequency features and learning the signal with low frequency features.



(f) Legend for all the plots.

Figure 1: Learning $\cos(3x)$ using linear regression with different featurizations. The data points $(x_i, y_i)_{i=1}^{60}$ were generated i.i.d. such that x_i have uniform distribution on $[0, \pi]$ and y_i have normal distribution with mean $\cos(3x_i)$ and standard deviation 0.4. The OLS estimator was used when n is larger than the number of features, and the minimum norm interpolating solution was used otherwise.

3.3 Our contribution

We complete the story of Bartlett et al. (2020) by providing sharp bounds on the bias term, extending the results to the setting of ridge regression with nonzero λ, and replacing the assumption of independence of the components by a much broader sufficient condition. From our point of view, k^* is the main discovery of Bartlett et al. (2020). In our work we also start with separation of the first k eigendirections and show that the same split leads to a bound for the bias term that is in full alignment with the intuitive explanation given above.

Let's introduce some notation. Recall that we fixed the basis to be the eigenbasis of the covariance in (1). For any $k \in \{0, 1, \dots, p\}$ we denote $X_{0:k}$ and $Z_{0:k}$ to be the matrices comprised of the first k columns of X and Z respectively.[1] Analogously, we denote $X_{k:\infty}$ and $Z_{k:\infty}$ to be the matrices comprised of the last $p-k$ columns of X and Z, $\Sigma_{0:k} = \operatorname{diag}(\lambda_1, \dots, \lambda_k)$ and $\Sigma_{k:\infty} = \operatorname{diag}(\lambda_{k+1}, \dots, \lambda_p)$. For any $\theta \in \mathbb{R}^p$ we denote $\theta_{0:k}$ to be the vector comprised of the first k components of θ, and $\theta_{k:\infty}$ — of the remaining components. We choose the $k : \infty$ notation instead of $k : p$ to emphasize that our results don't depend on p, and only the notions of effective dimension implicitly given by the sequence $\{\lambda_i\}_{i=1}^p$ matter. For example, if one increases the dimension to $p' > p$ and pads the sequence $\{\lambda_i\}_{i=1}^p$ with $p' - p$ zeros, our results will still hold.

The central object in our analysis is the following matrix:

$$A_k := X_{k:\infty} X_{k:\infty}^\top + \lambda I_n. \tag{3}$$

The matrix $X_{k:\infty} X_{k:\infty}^\top$ is the Gram matrix of the data after removing the first k components. A_k is obtained from that Gram matrix by shifting all eigenvalues by the ridge regularization parameter λ.

In Bartlett et al. (2020), the crucial step was to show that the singular values of A_k are within a constant factor of each other for $k = k^*$ (see their Lemma 5). When the components of data vectors are independent, such control over the condition number is a consequence of high effective rank. In this paper, the roles of effective rank and condition number of A_k are reversed. We prove sharp bounds assuming that there is some oracle that guarantees that with high probability all eigenvalues of A_k are within a constant factor of each other. Independence of components is not needed. Moreover, such control implies that ρ_k is at least a constant, which, in turn, implies sharpness of the bounds. In other words, we provide a more general condition under which the tail of the data is "essentially high dimensional" — instead of assuming independent components and high effective rank, only oracle control of condition number of A_k is needed. In Section 5 we provide an extensive discussion of this assumption: we show that a version of a small-ball condition for the tails of the data is required and that a stronger version of the same condition is sufficient if the data is sub-Gaussian.

The bound that we obtain for the bias term is given informally by the following expression:

$$B \approx \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k} \lambda_i}{n} \right)^2.$$

One can see how it aligns with the intuition of "essentially low-dimensional" and "essentially high-dimensional" parts: one cannot estimate the signal in the high dimensional part, so almost all of its energy $\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}$ goes into the error. When it comes to the low-dimensional part, the high-dimensional part acts as a ridge regularizer for it, so the bias in the first k components is the same as that of ridge regression with regularization coefficient $\lambda + \sum_{i>k} \lambda_i$ (i.e., the full regularization is

1. When $k = 0$ these matrices are just empty and all the terms that involve $0 : k$ index become zero.

equal to the explicitly imposed part λ plus "implicit regularization", which is equal to the energy of the tail.)

Our extension of the results to the ridge regression scenario allows us to answer the following question: can it happen that the "essentially high dimensional part" is too high dimensional, meaning that it provides too much regularization and negative λ is needed to compensate for that? In Section 8, we show that this indeed can happen and that the following is sufficient for it to be true: the noise and the energy of the signal in the tail (components $k:\infty$) are small compared to the signal in the spiked[2] part (components $0:k$), but the effective rank of the tail abruptly becomes much larger than n.

3.4 Additional notation

For any symmetric matrix $M \in \mathbb{R}^{n\times n}$ and any $i \in \{1, 2, \dots, n\}$ we write $\mu_i(M)$ for the i-th largest eigenvalue of M. For example, $\mu_1(M)$ is its largest eigenvalue, and $\mu_n(M)$ is the smallest. We write $M[i, j]$ for the element of M standing in the i-th row and j-th column.

Throughout the paper the following objects will be needed: for any i denote z_i to be the i-th column of Z. Then define

$$A_{-i} := X_{0:i-1}X_{0:i-1}^\top + X_{i:\infty}X_{i:\infty}^\top + \lambda I_n = \lambda I_n + \sum_{j\neq i} \lambda_j z_j z_j^\top,$$

an analogue of the matrix A, but we throw away the i-th component of the data vectors. Denote also

$$\rho_k(0) = \frac{1}{n\lambda_{k+1}} \sum_{i>k} \lambda_i,$$

the ratio of the effective rank of the tail to the number of data points without taking regularization λ into account.

Finally, denote our data points to be $\{x^i\}_{i=1}^n$, i.e., $X^\top = [x^1, \dots, x^n] \in \mathbb{R}^{p\times n}$.

For the readers convenience, we compile all the notation in Appendix A.

4. Main results

As we explained in the previous section, the central objects in our proof are A_k and ρ_k. In principle, any control of the spectrum of A_k leads to some upper bound on B and V (see our Theorem 5), the question is when that bound is tight. The intuitive answer is the following: the bound is tight when the condition number of A_k is a constant and k is chosen correctly, meaning that either ρ_k is a constant or k is the smallest number such that ρ_k is larger than a constant (i.e., $k = k^*$).[3] Our arguments, however, only support this intuition when the following technical assumption holds for some constant $\gamma < 1$:

NoncritReg(k, γ) Assume that $\lambda > -\gamma \sum_{i>k} \lambda_i$.

2. Here we use the word "spiked" as in the "spiked covariance models", which usually assume that the eigenvalues of $\Sigma_{k:\infty}$ are all equal and of smaller order than eigenvalues of $\Sigma_{0:k}$. One way to interpret our results is that only spiked-covariance-like models can exhibit benign overfitting, and we derive general conditions for a model to be spiked-covariance-like.
3. Note that there may be several values of k that satisfy these conditions. Applying our upper bound for any of those k will yield the same result up to a constant factor.

The reason why this assumption is needed is that as λ approaches $-\sum_{i>k}\lambda_i$, $\mathbb{E}A_k$ approaches zero. It still can be possible to bound the eigenvalues of A_k with high probability in such regime, but their magnitude will be smaller, and some error terms that were dominated before become significant. We do investigate such a regime in Section 8, where we show that negative regularization may give better rates than any value of non-negative regularization, but we only provide an upper bound there. For all the results we discuss in this section, we make Assumption *NoncritReg*(k,γ).

The focus of our work was to obtain the tight upper bound on the excess risk under minimal assumptions. Such minimal assumption turns out to be

CondNum(k,δ,L) Assume that with probability at least $1-\delta$ the matrix A_k is positive-definite (PD) with condition number at most L.

We provide a thorough discussion of this assumption in Section 5, for example we derive sufficient and almost matching necessary conditions for it to hold when the distribution is sub-Gaussian. The reason why we don't just assume those sufficient conditions is that we believe that sub-Gaussianity is not essential for our results to hold, as we discuss in Section 6.4. Moreover, the matrix A_k is the central object in our argument, and making an assumption on its condition number explicitly makes presentation easier.

A careful reader will notice that we have just stated that another condition is needed for the bound to be tight: k should be chosen in the right way. This, however, can be achieved by shifting k to k^* if necessary: indeed, assumptions *NoncritReg*(k,γ) and *CondNum*(k,δ,L) imply a constant lower bound on ρ_k (see Corollary 6). That means that either ρ_k is a constant, or it is more than a constant, i.e., $k > k^*$. In the latter case one can shift from k to k^* meaning that Assumption *CondNum*(k^*,δ',L') also holds with modified constants δ', L' (see Lemma 11 for the exact statement). Now applying the upper bound (Corollary 6) with $k = k^*$ gives tight result, as given by the following.

Theorem 1 *Fix any constants $b > 0$, $\gamma \in [0,1)$, $L > 0$. Denote*

$$k^* = \min\{\kappa : \rho_\kappa > b\}.$$

There exists a constant c which only depends on σ_x, b, γ, L such that the following holds: suppose NoncritReg$(\bar{k},\gamma)$ and CondNum$(\bar{k},\delta,L)$ are satisfied for some $\bar{k} < n/c$ and $\delta < 1 - ce^{-n/c}$. Take $k = \min(\bar{k}, k^*)$. Then with probability at least $1 - ce^{-n/c} - \delta$

$$B/c \le \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2, \tag{4}$$

$$V/c \le \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}. \tag{5}$$

Moreover $\rho_k \ge c^{-1}$, NoncritReg(k,γ) holds, and there exist L', c' that only depend on σ_x, b, γ, L s.t. CondNum$(k, \delta + c'e^{-n/c'}, L')$ holds.[4]

Proof In this proof let's call any quantities that only depend on σ_x, γ, b and L "constants". First of all, if $\bar{k} \le k^*$ then $k = \bar{k}$. Since we are given that *NoncritReg*$(\bar{k},\gamma)$ and *CondNum*$(\bar{k},\delta,L)$ are satisfied, we immediately get that *NoncritReg*(k,γ) and *CondNum*$(k,\delta + c'e^{-n/c'}, L')$ are satisfied

4. That is, the assumptions still hold if we substitute $\bar{k}$ by k, but with different L, δ. Further we will see that satisfaction of these assumptions implies tightness of the bounds for the chosen k.

with $L' = L$ and any $c' > 0$. However, if $\bar{k} > k^*$ then $k = k^*$ and by Lemma 11 *NoncritReg*(k, γ) and *CondNum*$(k, \delta + c'e^{-n/c'}, L')$ are still satisfied for some constants c', L'. Note that the larger the constants, the looser the assumptions, so we can take our final choice of c', L' to be the maximum over two cases.

Now that we know that *NoncritReg*(k, γ) and *CondNum*$(k, \delta + c'e^{-n/c'}, L')$ are satisfied, by Corollary 6, there is a constant c_1 such that $\rho_k > 1/c_1$ and with probability at least $1 - c_1 e^{-n/c_1} - c'e^{-n/c'} - \delta$

$$
\begin{aligned}
B/c_1 \leq & \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k} \lambda_i}{n} \right)^2, \\
V/c_1 \leq & \frac{k}{n} + \frac{n \sum_{i>k} \lambda_i^2}{\left(\lambda + \sum_{i>k} \lambda_i\right)^2}.
\end{aligned}
$$

Taking $c \geq c_1 + c'$ gives the first part. ■

Algebraically, under Assumption *CondNum*(k, δ, L) all eigenvalues of A_k^{-1} are within a constant factor of each other, so one can pull its operator norm from the expressions and obtain an upper bound without losing tightness. This strategy, however, doesn't produce lower bounds, so we derive them in a different way. Because of that, we impose different assumptions, namely

IndepCoord Assume that all elements of matrix X are independent (i.e., data vectors have independent coordinates).[5]

for the variance term, and

ExchCoord Assume that the sequence of coordinates of $\Sigma^{-1/2}x$ is exchangeable (any deterministic permutation of the coordinates of whitened data vectors doesn't change their distribution).

PriorSigns$(\bar{\theta})$ Assume that θ^* is sampled from a prior distribution in the following way: one starts with vector $\bar{\theta}$ and flips signs of all its coordinates with probability 0.5 independently.

for the bias term. The reason we introduce them is purely technical: taking expectation over the prior signs kills the cross-terms in the expression for bias, after which we decompose bias and variance into sums with respect to individual coordinates of the predictor, and bound each term in each sum from below. The latter is possible because of the Assumptions *IndepCoord* and *ExchCoord*. We don't believe those assumptions to be necessary for our results to be tight, but because of this mismatch in assumptions, our lower bounds don't formally show that our upper bound is always tight. What they show is that one needs some specific knowledge about the distribution to obtain better bounds. We provide a more detailed discussion of the relations between those assumptions in Section 6.2. The lower bounds themselves are given by the following

5. Recall that we fix the basis to be the eigenbasis of the covariance from the very beginning. Because of that, Assumption *IndepCoord* is stronger than the assumption that elements of $X\Sigma^{-1/2}$ are independent in some basis, that is often made in Random Matrix Theory literature.

Theorem 2 *Fix any constants $b > a > 0$, $\gamma \in [0, 1)$, $L > 0$. Denote*

$$k^* = \min\{\kappa : \rho_\kappa > b\}.$$

There exists a constant c which only depends on σ_x, a, b, γ, L such that all the following hold:

1. *For any $k \in \{0, 1, \ldots, k^*\}$ under assumptions IndepCoord, NoncritReg(k, γ), if $\rho_k > a$ then with probability at least $1 - 2\delta - ce^{-c/n}$*

$$V \geq \frac{1}{c}\left(\frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}\right).$$

2. *For any $k \in \{1, 2, \ldots, k^*\}$ under assumptions NoncritReg(k, γ), CondNum(k, δ, L), PriorSigns$(\bar{\theta})$ and ExchCoord, if $\rho_k > a$ then with probability at least $1 - 2\delta - ce^{-c/n}$*

$$\mathbb{E}_{\theta^*} B \geq \frac{1}{c}\left(\|\bar{\theta}_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\bar{\theta}_{0:k}\|^2_{\Sigma^{-1}_{0:k}}\left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2\right),$$

where $\mathbb{E}_{\theta^*}$ denotes expectation over a random draw of θ^* from the distribution described in assumption PriorSigns$(\bar{\theta})$.[6]

Proof Lemma 7 gives a lower bound for V, and Lemmas 8 and 9 give the lower bound for B. Those lower bounds have the desired probability, but different algebraic form. To bring them to the same form as the upper bounds one needs the right k to be chosen. We assumed that $\rho_k > a$. Moreover, since $k \leq k^*$ by definition of k^* we either have $\rho_k \leq b$ or $k = k^*$. In both of those cases Theorem 10 guarantees that these lower bounds are the same as what we need up to multiplicative constants that only depend on σ_x, γ, a, b and L. ■

One can notice from this proof that having separate arguments for the lower bounds results in a different algebraic form of the same bound. This different form turns out to be convenient to draw explicit connections between our results and results from earlier works. We do that in Section 7.

5. Effective ranks and control of the spectrum of A_k

The central assumption that we need to compute the excess risk is Assumption *CondNum*(k, δ, L), which provides control over condition number of A_k. In this section we discuss when this assumption is known to be satisfied and what are the necessary conditions for it to happen.

5.1 Effect of λ on the condition number

Recall that $A_k = X_{k:\infty}X_{k:\infty}^\top + \lambda I_n$, so its spectrum is the shift by λ of the spectrum of $X_{k:\infty}X_{k:\infty}^\top$, the random matrix that is equal to the Gram matrix of the projected data. There are therefore three ways of establishing a constant upper bound on the condition number of A_k:

1. Establish an upper bound $\bar{\mu}$ on $\mu_1(X_{k:\infty}X_{k:\infty}^\top)$ and take $\lambda > \bar{\mu}/c$ for some constant $c > 0$. In this case, the singular values of A_k are all equal to λ (and greater than $\bar{\mu}$) up to a constant multiplier.

6. Note that under this distribution $\|\bar{\theta}_{k:\infty}\|_{\Sigma_{k:\infty}} = \|{\theta^*}_{k:\infty}\|_{\Sigma_{k:\infty}}$ and $\|\bar{\theta}_{0:k}\|_{\Sigma^{-1}_{0:k}} = \|{\theta^*}_{0:k}\|_{\Sigma^{-1}_{0:k}}$ almost surely.

2. Establish upper and lower bounds $\bar{\mu}$ and $\underline{\mu}$ on $\mu_1(X_{k:\infty}X_{k:\infty}^\top)$ and $\mu_n(X_{k:\infty}X_{k:\infty}^\top)$ respectively, such that $\bar{\mu}/\underline{\mu}$ is a constant. Then take $\lambda > -\underline{\mu}/c$ for some constant $c > 1$. In this case, the singular values of A_k are all equal to $\bar{\mu}$ (or $\underline{\mu}$) up to a constant multiplier.

3. Establish upper and lower bounds $\bar{\mu}$ and $\underline{\mu}$ on $\mu_1(X_{k:\infty}X_{k:\infty}^\top)$ and $\mu_n(X_{k:\infty}X_{k:\infty}^\top)$ respectively, and take $\lambda = -\underline{\mu} + \lozenge$, where $\lozenge \geq c(\bar{\mu} - \underline{\mu})$ for a constant $c > 0$. In this case, the singular values of A_k are all equal to $\lozenge$ up to a constant multiplier. This case can be substantially different from the previous case when the singular values of $X_{k:\infty}X_{k:\infty}^\top$ are very well concentrated, i.e., the gap $\bar{\mu} - \underline{\mu}$ is of smaller order than $\underline{\mu}$ itself. In this case $\lozenge$ can be a smaller order term.

Our bounds are sharp when assumption *NoncritReg*(γ) is satisfied for some $\gamma < 1$, i.e., in the first and the second case above. The third case is quite rare because it requires very good concentration of the spectrum of $X_{k:\infty}X_{k:\infty}^\top$. Moreover, in this case λ is very close to the critical negative value under which it is impossible to even guarantee that A_k is PD as it becomes negative definite in expectation. We use this regime to investigate how negative regularization can improve excess risk by more than a constant factor in Section 8. However, we don't expect our bounds to always be sharp in this regime.

Therefore, we focus our attention on the first two cases. In Section 5.2 we discuss informally what conditions on the distribution are necessary to bound $\mu_1(X_{k:\infty}X_{k:\infty}^\top)$ and $\mu_n(X_{k:\infty}X_{k:\infty}^\top)$, and show how notions of high effective rank and norm concentration condition arise. In Section 5.3 we combine those bounds for sub-Gaussian data with the choice of λ to provide necessary and almost matching sufficient conditions for the condition number of A_k to be constant under sub-Gaussianity. In Section 5.4 we show that sub-Gaussianity is not actually required for the condition number of A_k to be controlled with high probability: Theorem 4 states that norm concentration condition and a modified version of high effective rank condition are sufficient even if the data only has bounded $4 + \varepsilon$ moments.

5.2 Informal necessary conditions

There are several easy observations that help understand what is needed for the condition number of A_k to be bounded.

1. The first observation is that $X_{k:\infty}X_{k:\infty}^\top \succeq \lambda_{k+1} z_{k+1} z_{k+1}^\top$, where z_{k+1} is the first column of $Z_{k:\infty}$ —a vector with n i.i.d. coordinates with unit variance. By the law of large numbers, $\|z_{k+1}\|^2 \approx n$, meaning that $\|\lambda_{k+1} z_{k+1} z_{k+1}^\top\| \approx \lambda_{k+1} n$. Therefore, $\bar{\mu} \gtrsim \lambda_{k+1} n$.

2. The second observation is that the diagonal elements of $X_{k:\infty}X_{k:\infty}^\top$ are squared norms of the tails of data vectors. Recall that we denoted the data points to be $\{x^i\}_{i=1}^n$. We can write

$$(X_{k:\infty}X_{k:\infty}^\top)[i,i] = \|x_{k:\infty}^i\|^2 \text{ — i.i.d. r.v's.}$$

   Once again, by the law of large numbers, $\operatorname{tr}(X_{k:\infty}X_{k:\infty}^\top) \approx n\sum_{i>k}\lambda_i$, which implies that $\bar{\mu} \gtrsim \sum_{i>k}\lambda_i \gtrsim \underline{\mu}$. Combining it with the first observation shows that $\bar{\mu}$ and $\underline{\mu}$ can only be within a constant multiplier of each other when $\sum_{i>k}\lambda_i \geq c\lambda_{k+1} n$ for some constant c. This is exactly the high effective rank condition $\rho_k > c$ for $\lambda = 0$.

3. The third observation is that the diagonal elements of a PD matrix themselves provide bounds on the singular values:

$$\mu_n(X_{k:\infty}X_{k:\infty}^\top) \le \min_{i\in[n]}(X_{k:\infty}X_{k:\infty}^\top)[i,i] \le \max_{i\in[n]}(X_{k:\infty}X_{k:\infty}^\top)[i,i] \le \mu_1(X_{k:\infty}X_{k:\infty}^\top).$$

Therefore, to control condition number of $X_{k:\infty}X_{k:\infty}^\top$ by a constant L with probability $1-\delta$, it is necessary to guarantee that

$$\max_i \|x_{k:\infty}^i\|^2 \le L \min_j \|x_{k:\infty}^j\|^2,$$

i.e., n independent random draws of the random variable $\|x_{k:\infty}\|^2$ should all lie within a constant factor of some value, meaning that the norm of the tail of a data vector should be within a constant factor of a fixed value with probability $(1-\delta)^{1/n}$.

5.3 Controlling condition number under sub-Gaussianity

Sub-Gaussianity of the data implies an upper bound on $\mu_1(A_k)$, but doesn't help with $\mu_n(A_k)$. To see this one can consider a well-known construction: take a sub-Gaussian distribution and construct another distribution in the following way: to sample from this new distribution take a vector from the old distribution and multiply it by $\sqrt{2}$ with probability $1/2$ and by zero otherwise. The new distribution is still sub-Gaussian with the same covariance, but the Gram matrix of n i.i.d. samples from it is degenerate with probability at least $1-2^{-n}$. Therefore, an additional assumption is needed to lower bound $\mu_n(A_k)$. As we already mentioned in Section 5.2, we need norm concentration. Since sub-Gaussianity allows to bound the norm from above, it reduces to a version of the small-ball condition: $\|x_{k:\infty}\|$ should be lower-bounded with high probability. The formal result is given by the following

Lemma 3 (Controlling $\mu_1(A_k)/\mu_n(A_k)$ under sub-Gaussianity) *For any $\gamma \in [0,1)$ and $\sigma_x > 0$ there exists $c > 0$ that only depends on σ_x and γ such that under Assumption NoncritReg(k,γ) the following holds: for any $L \ge 1$*

- *If $\rho_k \ge L^2$ and with probability at least $(1-\delta)^{1/n}$*

$$\lambda + \|x_{k:\infty}\|^2 \ge \frac{c}{L}\left(\lambda + \mathbb{E}\|x_{k:\infty}\|^2\right),$$

then with probability at least $1-\delta-ce^{-n/c}$

$$\mu_n(A_k) \ge L^{-1}\mu_1(A_k).$$

- *Suppose that it is known that with probability at least $ce^{-n/c}$ $\mu_n(A_k) \ge L^{-1}\mu_1(A_k)$. Then $\rho_k \ge \frac{1}{cL}$ and with probability at least $\left(1-ce^{-n/c}\right)^{1/n}$*

$$\lambda + \|x_{k:\infty}\|^2 \ge \frac{1}{cL}\left(\lambda + \mathbb{E}\|x_{k:\infty}\|^2\right).$$

The proof is given in Appendix D. One can see that both the necessary and the sufficient conditions are that ρ_k is lower bounded by a constant and a version of small-ball condition that says that the regularized squared norm of the data exceeds a constant fraction of its expectation with probability $(1-\delta)^{1/n}$. There is, however, a gap in those constants.

5.4 Heavy-tailed case

The following is a direct corollary of Theorem 2.1 from Guédon et al. (2017)

Theorem 4 *Suppose that the distribution of the tail satisfies the following two assumptions:*

1. **Norm concentration:** *For some $\delta \in (0, 1/n)$, $L > 1$ and $M > 0$*

$$\mathbb{P}(L^{-1} \leq \|x_{k:\infty}\|/M \leq L) \geq 1 - \delta.$$

2. **Heavy-tailed effective rank:** *for some $h > 4$ denote $r_{h,k} > 0$ to be the maximum number such that for any $a \in \mathcal{S}^{p-k-1}$ and $t > 0$*

$$\mathbb{P}\left(\frac{\sqrt{r_{h,k}}\left|a^\top x_{k:\infty}\right|}{M} > t\right) \leq t^{-h}.$$

There exists a constant c that only depends on h such that with probability at least $1 - cn^{1-h/4} - n\delta$

$$\begin{aligned}
\mu_1(X_{k:\infty}X_{k:\infty}^\top) \leq & M^2\left(L^2 + cL^2\left(n^{1-h/4} + \sqrt{\frac{n}{r_{h,k}L^2}} + \frac{n}{r_{h,k}L^2}\right)\right), \\
\mu_n(X_{k:\infty}X_{k:\infty}^\top) \geq & M^2\left(L^{-2} - cL^2\left(n^{1-h/4} + \sqrt{\frac{n}{r_{h,k}L^2}} + \frac{n}{r_{h,k}L^2}\right)\right).
\end{aligned}$$

Proof First, note that by union bound with probability at least $1 - n\delta$ all the diagonal elements of the matrix $X_{k:\infty}X_{k:\infty}^\top$ belong to the segment $[L^{-2}M^2, L^2M^2]$. Next, take the bound on B_k from the Case 1 of Theorem 2.1 from Guédon et al. (2017) with the following choice of their parameters: $k = N$, $\tau = 1$, $\lambda = p$, $\sigma = 1 + p/4$, $t = \sqrt{n}$. Use that bound for vectors $\sqrt{r_{h,k}}x_{k:\infty}^i/M$. Note that that B_k is exactly the operator norm of the off-diagonal part of $r_{h,k}X_{k:\infty}X_{k:\infty}^\top/M^2$. ■

The quantity $r_{h,k}$ that we introduced in Theorem 4 can be interpreted as a notion of effective rank for heavy tailed distributions. Indeed, one can write

$$\sqrt{r_{h,k}} = \frac{M}{\inf\left\{\tau : \forall a \in \mathcal{S}^{p-k-1} \forall t > 0\ \mathbb{P}\left(\left|a^\top x_{k:\infty}\right|/\tau > t\right) \leq t^{-h}\right\}}.$$

— the ratio of the typical norm of the random vector to the scale of the worst case deviations of its one-dimensional projection. This is completely analogous to our usual definition of the effective rank: $r_k = \lambda_{k+1}^{-1}\sum_{i>k}\lambda_i$. Indeed, in sub-Gaussian case $\sqrt{\sum_{i>k}\lambda_i}$ is the typical value of the norm of the vector $x_{k:\infty}$, and $\sqrt{\lambda_{k+1}}$ is up to constant the largest sub-Gaussian norm of its one-dimensional projection. We see that the conditions under which the eigenvalues of $X_{k:\infty}X_{k:\infty}^\top$ are within a constant factor of each other with high probability remain the same even in the heavy-tailed case: the norm of $\|x_{k:\infty}\|$ concentrates within a constant factor of a fixed quantity, and the heavy-tailed effective rank $r_{h,k}$ should be large compared to the number n of data points.

6. Structure of the proof and role of sub-Gaussianity

6.1 Upper bound

The core of our argument is Theorem 5 given below. There are two important things to note about it: first, it only requires sub-Gaussianity and matrix A_k being positive semidefinite (which always holds with probability 1 for non-negative λ). Second, its proof decomposes very clearly into two parts: an algebraic part, which only requires A_k being PD and holds with probability 1 conditionally on this event, and a probabilistic part, where standard concentration results are directly plugged into the algebraic bounds. Because of this decomposition, it is straightforward to track how the sub-Gaussianity is used and how it can be relaxed. We provide the sketch of the proof to show these details.

Theorem 5 *There exists a (large) constant c, which only depends on σ_x, s.t. for any $k < n/c$ with probability at least $1 - ce^{-n/c}$, if the matrix A_k is PD, then*

$$\begin{aligned}
B/c \leq & \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} \left(1 + \frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2} + n\lambda_{k+1}\mu_1(A_k^{-1})\left(1 + \max(0, -\lambda)\mu_1(A_k^{-1})\right)\right) \\
& + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{1}{n^2\mu_n(A_k^{-1})^2} + \frac{\lambda_{k+1}}{n}\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\left(1 + \max(0, -\lambda)\mu_1(A_k^{-1})\right)\right), \\
V/c \leq & \frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}\frac{k}{n} + n\mu_1(A_k^{-1})^2 \sum_{i>k}\lambda_i^2.
\end{aligned}$$

Proof sketch The full proof of Theorem 5 can be found in Section I.1 of the appendix. The following is a sketch of its derivation.

Recall the following notation: for any y

$$\hat{\theta}(y) = X^\top(\lambda I_n + XX^\top)^{-1}y.$$

As explained in Section 3.2, Bartlett et al. (2020) introduced the notion of k^* for which the behaviour of the variance term in the first k^* coordinates is qualitatively different than in the rest of the coordinates. Their argument, however, relies crucially on independence of the components of the data. The main idea that allowed us to get rid of that assumption and to obtain the tight bound for the bias term was to separate the first k coordinates from the very beginning and to use some sort of uniform convergence argument in that low-dimensional subspace.

The crucial tool that allowed us to realise this idea turned out to be the following algebraic identity that we prove in Section F of the appendix:

$$\hat{\theta}(y)_{0:k} + X_{0:k}^\top A_k^{-1} X_{0:k}\hat{\theta}(y)_{0:k} = X_{0:k}^\top A_k^{-1} y.$$

This identity allows convenient access to the error in the first k coordinates (the spiked part).

The argument decomposes clearly into two parts: algebraic and probabilistic. The algebraic part is to decompose the excess risk (up to a constant multiplier) into four terms and show that the

following inequalities hold on the event that the matrix A_k is PD:
(1) Bias error in the spiked part:

$$\|\hat{\theta}(X\theta^*)_{0:k} - \theta^*_{0:k}\|_{\Sigma_{0:k}} \leq \frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})} \frac{\mu_1\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right)^{1/2}}{\mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right)} \|X_{k:\infty}\theta^*_{k:\infty}\| + \frac{\|\theta^*_{0:k}\|_{\Sigma_{0:k}^{-1}}}{\mu_n(A_k^{-1})\mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right)}.$$

(2) Variance error in the spiked part:

$$\mathbb{E}_\varepsilon \|\hat{\theta}(\varepsilon)_{0:k}\|^2_{\Sigma_{0:k}} \leq \frac{\mu_1(A_k^{-1})^2 \|X_{0:k}\Sigma_{0:k}^{-1/2}\|^2}{\mu_n(A_k^{-1})^2 \mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right)^2}.$$

(3) Variance error in the tail:

$$\mathbb{E}_\varepsilon \|\hat{\theta}(\varepsilon)_{k:\infty} - \theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} \leq \mu_1(A_k^{-1})^2 \operatorname{tr}(X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^\top).$$

(4) Bias error in the tail:

$$\begin{aligned}
&\frac{1}{3}\|\hat{\theta}(X\theta^*)_{k:\infty} - \theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} \\
\leq &\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \lambda_{k+1}\big(1 + \max(0, -\lambda)\mu_1(A_k^{-1})\big)\mu_1(A_k^{-1})\|X_{k:\infty}\theta^*_{k:\infty}\|^2 \\
+&\lambda_{k+1}\big(1 + \max(0, -\lambda)\mu_1(A_k^{-1})\big)\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2})}{\mu_k(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2})^2}\|\Sigma_{0:k}^{-1/2}\theta^*_{0:k}\|^2.
\end{aligned}$$

The probabilistic part of the argument is to control the quantities that arise in the algebraic bound with high probability. Namely, we plug in

- Concentration of k-dimensional sample covariance with n samples: w.h.p.

$$\mu_k\left(\frac{1}{n}\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right) \approx_{\sigma_x} \mu_1\left(\frac{1}{n}\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right) \approx_{\sigma_x} 1.$$

- Concentration of norm of vectors with i.i.d. components: w.h.p.

$$\frac{1}{n}\|X_{0:k}\Sigma_{0:k}^{-1/2}\|^2 \lesssim_{\sigma_x} k, \quad \frac{1}{n}\|X_{k:\infty}\Sigma_{k:\infty}^{1/2}\|^2 \lesssim_{\sigma_x} \sum_{i>k}\lambda_i^2, \quad \frac{1}{n}\|X_{k:\infty}\theta^*_{k:\infty}\|^2 \lesssim_{\sigma_x} \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}.$$

After plugging in the probabilistic bounds, the final result is obtained by a straightforward computation. ■

Note that the only probabilistic statements that are used in this proof are concentration of sample covariance in dimension k and concentration of the sum of n i.i.d. random variables. The same concentration results hold with weaker assumptions, but with larger probability. For example, under

rather weak moment assumptions only a linear in dimension number of samples is needed for the sample covariance matrix to concentrate within a constant factor of the population covariance, see Tikhomirov (2017) and references therein. It is also interesting to point out that the "uniform convergence" result that we mentioned in the beginning of this sketch is nothing but the convergence of the empirical covariance matrix $\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}/n$ to its expectation I_k, which is exactly the uniform convergence result that gives the bound in the "essentially low-dimensional" regime from Section 3.1.

Despite the fact that the bounds of Theorem 5 apply under very general assumptions, we don't expect them to be tight if the condition number of A_k is not bounded by a constant. When some oracle control of the condition number of A_k is provided, the bound becomes the following.

Corollary 6 *Fix any constants $\gamma \in [0, 1)$ and $L > 0$. There exists a constant c that only depends on σ_x, γ, L s.t. for any $k < n/c$ and $\delta < 1 - ce^{-n/c}$ under assumptions NoncritReg(k, γ) and CondNum(k, δ, L), it holds that $\rho_k > c^{-1}$, and with probability at least $1 - \delta - ce^{-n/c}$,*

$$
\begin{aligned}
B/c \leq & \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k} \lambda_i}{n} \right)^2, \\
V/c \leq & \frac{k}{n} + \frac{n \sum_{i>k} \lambda_i^2}{\left(\lambda + \sum_{i>k} \lambda_i\right)^2}.
\end{aligned}
$$

Proof sketch Assumptions *NoncritReg*(k, γ) and *CondNum*(k, δ, L) imply that all the eigenvalues of A_k are equal to $\lambda + \sum_{i>k} \lambda_i$ up to a multiplicative constant that depends on L, γ, σ_x. Plugging it into Theorem 5 gives the result. The full proof is given in Appendix I.1. ■

The sub-Gaussianity is used in Corollary 6 to ensure that $\mathrm{tr}(A_k)$ concentrates around $n\left(\lambda + \sum_{i>k} \lambda_i\right)$. Since the diagonal elements of A_k are i.i.d. random variables, the same concentration would also hold under weaker assumptions with lower but still high probability.

It is also worth mentioning that the story about "essentially high-dimensional" and "essentially low-dimensional" parts is not just an interpretation of the final result: the whole proof strategy is in accordance with it, as we explicitly separate the two parts and bound errors in them separately.

6.2 Lower bounds

Our lower bounds have a different form from the upper bounds. We show separately that they match if the condition on effective rank is satisfied. One benefit of this approach is that the lower bounds provide a different form of the same result, which allows for different analysis. We employ it in Section 7.

The lower bound for the variance term is given by the following lemma, whose proof is given in Appendix E.1:

Lemma 7 (Lower bound for the variance term) *Fix any constant $\gamma \in [0, 1)$. There exists a constant c that only depends on σ_x and γ s.t. for any $k < n/c$ under assumptions NoncritReg(k, γ) and IndepCoord w.p. at least $1 - ce^{-n/c}$*

$$
V \geq \frac{1}{cn} \sum_{i=1} \min \left\{ 1, \frac{\lambda_i^2}{\lambda_{k+1}^2 (\rho_k + 1)^2} \right\}.
$$

One can see that the assumptions under which the lower bound is proved are different from the assumptions required for the upper bound: we require independent components here. On the one hand, it means that there could be a gap between upper and lower bounds in some particular cases where one can control the condition number of A_k without independence of components. On the other hand, it means that even such strong additional assumption as independence of components does not allow the upper bounds to be improved, which suggests that those specific cases for which the bound is not tight are rare and require even stronger additional assumptions.

The most general lower bound for the bias term that we prove requires the following assumption

StableLowerEig(k, δ, L) Assume that for any $j \in \{1, 2, \dots, p\}$ with probability[7] at least $1 - \delta$

$$\mu_n(A_{-j}) \geq \mu_n(\mathbb{E}A_k)/L = \left(\sum_{i>k} \lambda_i + \lambda\right)/L,$$

and that $\lambda > -\sum_{i>k} \lambda_i$.

Then the bound is given by the following lemma, whose proof is given in Appendix E.2

Lemma 8 (Lower bound for the bias term) *Fix any constant $L > 0$. There exists c that only depends on σ_x and L s.t. for any $k \in \{1, 2, \dots, p\}$ under assumptions PriorSigns$(\bar{\theta})$ and StableLowerEig(k, δ, L) w.p. at least $1 - 2\delta - ce^{-n/c}$*

$$\mathbb{E}_{\theta^*} B \geq \frac{1}{c} \sum_i \frac{\lambda_i \bar{\theta}_i^2}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2},$$

where $\mathbb{E}_{\theta^*}$ denotes the expectation over the random draw of θ^* from the prior distribution described in assumption PriorSigns$(\bar{\theta})$.

Assumption *StableLowerEig*(k, δ, L) is formally not comparable to Assumption *CondNum*(k, δ, L), but informally if $k \geq 1$ then *StableLowerEig*(k, δ, L) is weaker: indeed, the matrix A_{-i} is obtained from the matrix A by subtracting $\lambda_i z_i^\top z_i$, while the matrix A_k is obtained from A by subtracting $\sum_{i=1}^k \lambda_i z_i^\top z_i$, i.e., the sum of k "largest" of the terms $\lambda_i z_i^\top z_i$. Therefore, the matrix A_{-i} is "larger" than A_k, and controlling its lowest singular value should be easier. The following lemma, whose proof is given in Appendix E.2, formalizes this argument under Assumption *ExchCoord*:

Lemma 9 *For any $\gamma < 1$ there exists a constant c that only depends on γ and σ_x such that if assumptions CondNum(k, δ, L), NoncritReg(k, γ) and ExchCoord are satisfied for some $L \geq 1$ and $k \in \{1, 2, \dots, p\}$, then StableLowerEig$(k, \delta + 2e^{-n/c}, cL)$ is also satisfied.*

When it comes to averaging over the prior given by the assumption *PriorSigns*$(\bar{\theta})$, it just means that it is impossible to obtain a better lower bound without some specific knowledge of how signs of components of θ^* interact with the probability distribution of the data.

7. Note that the condition on probability is separate for every j, i.e., we don't assume that events hold simultaneously for all j.

6.3 Connecting upper and lower bounds

One slight inconvenience with our approach of imposing oracle control over the spectrum of A_k via Assumption *CondNum*(k, δ, L) is the following: what if the oracle provides control for the wrong value of k? There can in principle be many values of k for which such oracle control is possible, with not all of them giving the right point where the behaviour changes from "essentially low-dimensional" to "essentially high-dimensional". As an example, consider the isotropic setting with $p \gg n$: one can exclude any number k of components such that $p - k \gg n$ and still be able to control the condition number.

First of all, in accordance with the result of Bartlett et al. (2020), the following theorem shows that the "right k" is the k that is not larger than k^*.

Theorem 10 (The lower bound is the same as the upper bound) *Denote*

$$\underline{B} := \sum_i \frac{\lambda_i |\theta_i^*|^2}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2},$$

$$\overline{B} := \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2,$$

$$\underline{V} := \frac{1}{n}\sum_i \min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\},$$

$$\overline{V} := \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}.$$

Fix constants $a > 0$ and $b > 1/n$. There exists a constant $c > 0$ that only depends on a, b, s.t. the following holds: if either $\rho_k \in (a, b)$ or $k = \min\{\kappa : \rho_\kappa > b\}$, then

$$c^{-1} \le \underline{B} \,/\, \overline{B} \le 1, \quad c^{-1} \le \underline{V} \,/\, \overline{V} \le 1.$$

Proof The proof is a rather straightforward comparison of pairs of sums term by term. It is given in Appendix I.2. ■

Secondly, if the data is sub-Gaussian, then oracle control for any $k < n$ results in tight bounds, but with worse constants. This happens because of the following lemma.

Lemma 11 (k can be taken to be k^*) *Fix any constants $\gamma \in [0, 1)$, $b > 0$, $L > 0$. Denote*

$$k^* = \min\{k : \rho_k > b\}.$$

There exist constants c, L' that only depend on σ_x, γ, b, L s.t. the following holds: suppose assumptions NoncritReg(k, γ) and CondNum(k, δ, L) hold for some $k \in [k^*, n]$. Then assumptions NoncritReg(k^*, γ) and CondNum$(k^*, \delta + ce^{-n/c}, L')$ hold too.

Proof sketch Since $k \ge k^*$, $\mu_n(A_k)$ provides a lower bound for $\mu_n(A_{k^*})$. When it comes to $\mu_1(A_{k^*})$, it can be bounded with high-probability because the data is sub-Gaussian. The full proof is given in Appendix D. ■

6.4 The role of sub-Gaussianity

As can be seen from the proof of Theorem 1, the strategy to obtain a tight bound is the following: ask the oracle to control the condition number of A_k, if that k is too large, shift it to k^*, and then apply the bound from Corollary 6. In Section 5.4 we showed that if the norm $\|x_{k:\infty}\|$ concentrates, and the effective rank $r_{h,k}$ is high enough, then the control over the condition number of A_k is possible even if we have very weak moment assumptions instead of sub-Gaussianity. Moreover, as we have discussed in the proof sketches, if we didn't shift from k to k^*, we would only need the usual concentration results such as the law of large numbers or concentration of k-dimensional empirical covariance matrix with n samples, which also hold under weak moment assumptions. Therefore, sub-Gaussianity is not essential to obtain the bound in the form given in Corollary 6, one just needs to substitute the sub-Gaussian concentration results with their heavy-tailed analogues. However it may not necessarily give a tight result unless the oracle is guaranteed to choose the appropriate k (e.g., $k = k^*$). To shift from k to k^* we also need an upper bound on $\|A_{k^*}\|$, which we derive from sub-Gaussianity. According to Section 5.4, an analogous bound is still possible under weak moment assumptions, but additional work is required: to use Theorem 4 for $k = k^*$ one would need to obtain a high-probability upper bound on $\|x_{k^*:\infty}\|$ under moment assumptions and to relate r_k which we use in definition of k^* to $r_{h,k}$, which is introduced in Theorem 4.

7. Alternative forms of the bounds and effect of increasing regularization

7.1 Alternative form of the bound and its relation to classical in-sample analysis

Theorem 10 reveals an alternative form of the bounds: when ρ_k is lower- and upper-bounded by constants or when $k = k^*$, the bounds on the bias and variance respectively become equal to the following up to a constant multiplier:

$$\tilde{B} := \sum_{i=1}^{p} \lambda_i |\theta_i^*|^2 \frac{\rho_k^2 \lambda_{k+1}^2}{\left(\rho_k \lambda_{k+1} + \lambda_i\right)^2}, \tag{6}$$

$$\tilde{V} := \frac{1}{n} \sum_{i=1}^{p} \frac{\lambda_i^2}{\left(\rho_k \lambda_{k+1} + \lambda_i\right)^2}. \tag{7}$$

These expressions closely resemble the classical expressions for the in-sample bias and variance of ridge regression. Indeed, a straightforward computation gives

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\varepsilon \|X\hat{\theta} - X\theta^*\|^2 \\
=&\frac{1}{n}\|(XX^\top(XX^\top + \lambda I_n)^{-1} - I_n)X\theta^*\|^2 + \frac{v_\varepsilon^2}{n}\|XX^\top(XX^\top + \lambda I_n)^{-1}\|_F^2 \\
=&\underbrace{\sum_{i=1}^{p} \hat{\lambda}_i \langle v_i, \theta^*\rangle^2 \frac{(\lambda/n)^2}{\left(\lambda/n + \hat{\lambda}_i\right)^2}}_{\text{in-sample bias}} + \underbrace{v_\varepsilon^2 \frac{1}{n} \sum_{i=1}^{p} \frac{\hat{\lambda}_i^2}{\left(\lambda/n + \hat{\lambda}_i\right)^2}}_{\text{in-sample variance}},
\end{aligned}$$

where $\{\hat{\lambda}_i\}_{i=1}^p$ are eigenvalues of the empirical covariance $n^{-1}X^\top X$ and $\{v_i\}_{i=1}^p$ are the corresponding eigenvectors. Recall that $\rho_k \lambda_{k+1} = \left(\lambda + \sum_{i>k} \lambda_i\right)/n$. One can see that Equations (6)–(7)

can be obtained from the classical equations for the in-sample risk by substituting the empirical eigenvalues with population eigenvalues and increasing the regularization level λ by $\sum_{i>k}\lambda_i$ — the energy of the tail of the data.

Similarly, $\tilde{B}$ has an interpretation as the bias term of ridge regression with infinite data: for $\bar{\lambda}>0$ denote $\theta^*_{\bar{\lambda}}$ to be the solution to the following "population ridge regression" problem:

$$\theta^*_{\bar{\lambda}} = \operatorname*{argmin}_{\theta}\left[\mathbb{E}\|X\theta - y\|^2 + \bar{\lambda}\|\theta\|^2\right] = \left(\Sigma + \frac{\bar{\lambda}}{n}I_p\right)^{-1}\Sigma\theta^*.$$

A straightforward computation gives

$$\|\theta - \theta^*\|^2_{\Sigma} = \sum_i \lambda_i|\theta^*_i|^2\frac{(\bar{\lambda}/n)^2}{(\lambda_i+\bar{\lambda}/n)^2},$$

which is equal to $\tilde{B}$ when $\bar{\lambda} = n\lambda_{k+1}\rho_k = \lambda + \sum_{i>k}\lambda_i$.

7.2 Dependence on λ

The alternative form of the bounds presented in Section 7.1 provides a convenient way to investigate the dependence on λ, which is cumbersome in the initial form because increasing λ may decrease k^*. This effect, however, is negligible when Equations (6)–(7) are considered. Indeed, in Appendix I.3 we show the following

Lemma 12 *Suppose $k < n/c$ for some $c>1$ and $k^*<k$. Then*

$$\lambda_{k+1}\rho_k \le \lambda_{k^*+1}\rho_{k^*} \le \lambda_{k+1}\rho_k/(1-b^{-1}c^{-1}).$$

Because of this lemma, any $k\in[k^*, n/c]$ gives the same result (up to a constant factor) in Equations (6)–(7). One can, therefore, start with some λ and the corresponding $k=k^*$ and then consider larger values of λ without decreasing k in Equations (6)–(7). The result will give sharp (up to a constant factor) bounds, which depend on λ as follows:

$$\tilde{B} = \sum_i \lambda_i|\theta^*_i|^2\frac{n^{-2}\left(\lambda+\sum_{i>k}\lambda_i\right)}{\left(n^{-1}\left(\lambda+\sum_{i>k}\lambda_i\right)+\lambda_i\right)^2},$$
$$\tilde{V} = \frac{1}{n}\sum_i \frac{\lambda_i^2}{\left(n^{-1}\left(\lambda+\sum_{i>k}\lambda_i\right)+\lambda_i\right)^2},$$

which are obtained by simply plugging in the definition of ρ_k into (6)–(7).

A particularly interesting case arises when λ is large enough that it dominates $\sum_{i>k}\lambda_i$ and all eigenvalues of A_k are equal to λ up to a constant multiplier. The corresponding result is given by the following corollary.

Corollary 13 *There is a large positive constant c that only depends on σ_x such that if*

$$\lambda > cn\lambda_{\lfloor n/c\rfloor} + 2\sum_{i>\lfloor n/c\rfloor}\lambda_i,$$

then

$$B/c \leq \sum_i \lambda_i |\theta_i^*|^2 \frac{(\lambda/n)^2}{(\lambda/n+\lambda_i)^2},$$
$$V/c \leq \frac{1}{n} \sum_i \frac{\lambda_i^2}{(\lambda/n+\lambda_i)^2}.$$

Proof The full proof is given in Appendix I.3; the following is its outline:

1. Use Lemma 3 to control the eigenvalues of $A_{\lfloor n/c \rfloor}$.
2. Use Theorem 1 to obtain the bounds for $k = k^*$.
3. Use Theorem 10 to convert the bounds into the form given in Equations (6)–(7).
4. Use Lemma 12 to substitute k^* back with $\lfloor n/c \rfloor$.
5. Since $\lambda > 2\sum_{i>k} \lambda_i$, λ/n is equal to $\rho_k \lambda_{k+1}$ up to a multiplicative constant.

■

Note that the statement of Corollary 13 does not require the notion of k^*.

7.3 Comparison with other results

As we saw in the previous section, the alternative form given by Equations (6)–(7) has milder dependence on the choice of k^* than our main bounds (4)–(5) and allows to compare to classical results for in-sample error of ridge regression. In this section we use it to compare with more recent developments: the non-asymptotic bounds in Hsu et al. (2014) and Hastie et al. (2020).

First of all, we follow Hsu et al. (2014) and introduce the following notion of effective dimension of the problem:

$$d(\bar{\lambda}) := \sum_i \frac{\lambda_i}{\bar{\lambda}+\lambda_i},$$

where $\bar{\lambda}$ is a parameter which can informally be understood as effective level of regularization. Hsu et al. (2014) provide non-asymptotic bounds for B and V in the regime when

$$n \geq c d(\lambda/n) \log(1 + d(\lambda/n)), \tag{8}$$

(see their Theorem 2).[8] The simplified version of their results given in Remark 17 gives the following bounds:[9]

$$B \leq \left(1 + \frac{c(1+d(\lambda/n))}{n}\right) \sum_i \lambda_i |\theta_i^*|^2 \frac{(\lambda/n)^2}{(\lambda/n+\lambda_i)^2},$$
$$V \leq \frac{c}{n} \sum_i \frac{\lambda_i^2}{(\lambda/n+\lambda_i)^2},$$

8. Note that in Hsu et al. (2014), the scaling of the regularization parameter is different from ours: to express their results in our terms one needs to substitute their λ by λ/n in our notation.
9. Note that under our assumptions, $\mathrm{approx}(x) = 0$, where $\mathrm{approx}(x)$ is defined in Equation (7) in Hsu et al. (2014).

where c is some constant that depends on the concentration properties of the data. This is the same as the result of Corollary 13, but with different constants. However, our Corollary 13 covers a wider range of λ if n is large enough. This follows from the following lemma, which is proven in Appendix I.3:

Lemma 14 *Suppose that $n \geq c^2 + c$ for some $c > 0$ and take*

$$\lambda = cn\lambda_{\lfloor n/c \rfloor} + 2 \sum_{i > \lfloor n/c \rfloor} \lambda_i.$$

Then

$$d(\lambda/n) \geq \frac{n}{2\max(2, (c+1)^2)}.$$

Indeed, $d(\lambda/n)$ is a decreasing function of λ, and due to Lemma 14 the range of λ for which Corollary 13 is applicable when $d(\lambda/n) = O(n)$, while Equation (8) restricts to the range $d(\lambda/n) = O(n/\log n)$.

After we posted the first preprint of this paper, the following non-asymptotic bound for the interpolating regime (i.e., $\lambda = 0$) appeared in (Hastie et al., 2020): informally

$$|V - V_S| \leq \frac{c}{n^{1/7}}, \quad |B - B_S| \leq \frac{c\|\theta^*\|^2}{n},$$

where c is a constant, V_S and B_S are defined as[10]

$$V_S := \tilde{\lambda}^{-1} \frac{\sum_i \frac{\lambda_i^2}{(1+\tilde{\lambda}^{-1}\lambda_i)^2}}{\sum_i \frac{\lambda_i}{(1+\tilde{\lambda}^{-1}\lambda_i)^2}}, \tag{9}$$

$$B_S := (1 + V_S) \sum_i \frac{\lambda_i |\theta_i^*|^2}{(1 + \tilde{\lambda}^{-1}\lambda_i)^2}, \tag{10}$$

and $\tilde{\lambda}$ is the solution to the equation $n = d(\tilde{\lambda})$. See their Definition 1 and Theorem 2 for the exact statement.[11]

Note that because of the equation for $\tilde{\lambda}$

$$\sum_i \frac{\lambda_i^2}{(\tilde{\lambda} + \lambda_i)^2} + \sum_i \frac{\tilde{\lambda}\lambda_i}{(\tilde{\lambda} + \lambda_i)^2} = \sum_i \frac{\lambda_i(\lambda_i + \tilde{\lambda})}{(\tilde{\lambda} + \lambda_i)^2} = d(\tilde{\lambda}) = n.$$

This allows us to rewrite (9)–(10) as

$$V_S := \frac{1}{1 - \frac{1}{n}\sum_i \frac{\lambda_i^2}{(\tilde{\lambda}+\lambda_i)^2}} \cdot \frac{1}{n} \sum_i \frac{\lambda_i^2}{(\tilde{\lambda} + \lambda_i)^2}, \tag{11}$$

$$B_S := (1 + V_S) \sum_i \lambda_i |\theta_i^*|^2 \frac{\tilde{\lambda}^2}{(\tilde{\lambda} + \lambda_i)^2}. \tag{12}$$

10. Here we introduce the notation $\tilde{\lambda} := (\gamma c_0)^{-1}$, where γ and c_0 are parameters used in Hastie et al. (2020).
11. Note that there is a typo in their definition of $\mathscr{V}$: a multiplicative factor of c_0 is missing.

Comparing these equations with (6)–(7) reveals that they are the same up to a constant multiplier whenever $\tilde{V} \leq 1 - 1/c$ for some constant c and $\rho_k \lambda_{k+1}$ is up to a constant equal to $\tilde{\lambda}$. In the following, we show that this is indeed the case.

Recall that these results from (Hastie et al., 2020) are for the interpolating regime, i.e., $\lambda = 0$. Let's see how $\tilde{\lambda}$ is related to $\lambda_{k+1}\rho_k$. The connection is given by the following lemma.

Lemma 15 *Suppose that $k < n/c$ and $\rho_k > c$ for some constant $c > 1$. Then*

$$\frac{\tilde{\lambda}}{\lambda_{k+1}\rho_k} \in \left(1 - \frac{1}{c}, \frac{1}{1 - \frac{1}{c}}\right).$$

Proof Denote $a = \frac{\tilde{\lambda}}{\lambda_{k+1}\rho_k}$. Then we can write

$$n = \sum_i \frac{\lambda_i}{\lambda_i + a\lambda_k\rho_k} \geq \sum_{i>k} \frac{\lambda_i}{\lambda_{k+1}(a\rho_k + 1)} = \frac{n\rho_k}{a\rho_k + 1},$$

which implies $a\rho_k + 1 \geq \rho_k$, so $a \geq 1 - 1/\rho_k > 1 - 1/c$.

For the upper bound on a we write

$$n = \sum_i \frac{\lambda_i}{\lambda_i + a\lambda_k\rho_k} \leq k + \sum_{i>k} \frac{\lambda_i}{a\lambda_{k+1}\rho_k} = k + \frac{n}{a},$$

which gives $a \leq n/(n-k) < c/(c-1)$. ■

The similarity of Equations (11)–(12) with our results should not be taken for granted, and it is actually quite surprising. As we explain in Section 9, the regime considered in Hastie et al. (2020) is significantly different, so it is rather unclear why the results would have the same form.

8. Negative regularization

The aim of this section is to find a family of regimes in which the optimal level of ridge regularization is negative. Since we are comparing different values of λ in this section, the following notation will be useful: recall that for any k

$$\rho_k(0) := \frac{1}{n\lambda_{k+1}} \sum_{i>k} \lambda_i,$$

the value of ρ_k for $\lambda = 0$. Intuitively, the components of the tail provide regularization for the first k components, and the larger ρ_k is, the more is that regularization. Thus, one could expect that if there is an abrupt jump in the sequence $\{\rho_k(0)\}_{k=0}^p$, then that additional regularization is too large and negative λ may be optimal.

As we investigate further, a jump in $\rho_k(0)$ is indeed one of the sufficient conditions for optimality of negative regularization, but not the only one: the strength of the noise and how the signal is distributed among the principal components of the data also play an important role.

We start the discussion with several informal observations. The first observation one can make is that V is a decreasing function of λ: indeed, $V = \mathrm{tr}(\Sigma^{1/2} X^\top A^{-2} X \Sigma^{1/2})$ and increasing λ increases all eigenvalues of A. Thus, negative regularization cannot help with damping the noise

compared to non-negative regularization, and the noise should be sufficiently small in order for negative regularization to be beneficial.

Now let's look at the role of the signal in the tail. It contributes to error in two ways: first — the components in the tail are not getting estimated themselves, second — the signal that comes from those components acts as additional noise for estimation of the first k components. When λ is non-negative, the error of the first type dominates the error of the second type, but negative λ can amplify the noise and result in error of the second type dominating. Therefore, the signal in the tail also needs to be sufficiently small in order for negative regularization to be optimal.

The final observation is the following: since we only compute the bounds up to a constant multiplier, the bound in Theorem 1 cannot distinguish between negative and zero regularization. To see this, consider the form of the bound given in Section 7: up to a constant factor the bound is a weighted combination in each component with weight $\lambda + \sum_{i>k} \lambda_i$, and as λ increases there is no need to change k. Now it is easy to see that for all λ in range from $-\gamma \sum_{i>k} \lambda_i$ to zero, that weight is the same up to a constant factor. Thus, negative regularization can only decrease the excess risk by more than a constant factor in the critical regime, i.e., $\lambda = -\sum_{i>k} \lambda_i + \Diamond$ where $\Diamond$ is of smaller order than $\sum_{i\geq k} \lambda_i$. To consider such λ and have A_k PD we need tight concentration of eigenvalues of $X_{k:\infty} X_{k:\infty}^\top$ around $\sum_{i>k} \lambda_i$. To ensure such tight control we restrict ourselves to the case of independent components, i.e., when Assumption *IndepCoord* is satisfied. In this case, the eigenvalues of $X_{k:\infty} X_{k:\infty}^\top$ can be bounded according to the following statement that was shown as an intermediate step in the proof of Lemma S.9 in (Bartlett et al., 2020).

Lemma 16 *Under assumption IndepCoord there exists a constant c that only depends on σ_x s.t. with probability at least $1 - ce^{-n/c}$,*

$$\mu_1(X_{k:\infty} X_{k:\infty}^\top) \leq \sum_{i>k} \lambda_i + c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k} \lambda_i^2}\right),$$

$$\mu_n(X_{k:\infty} X_{k:\infty}^\top) \geq \sum_{i>k} \lambda_i - c\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k} \lambda_i^2}\right).$$

The fluctuations $n\lambda_{k+1} + \sqrt{n\sum_{i>k} \lambda_i^2}$ will be of smaller order than $\sum_{i>k} \lambda_i$ if $\rho_k(0)$ is larger than a constant, which is shown by the following bounds:

$$n\lambda_{k+1} = \frac{1}{\rho_k(0)} \sum_{i>k} \lambda_i, \tag{13}$$

$$\sqrt{n\sum_{i>k} \lambda_i^2} \leq \sqrt{n\lambda_{k+1}\sum_{i>k} \lambda_i} = \frac{1}{\sqrt{\rho_k(0)}} \sum_{i>k} \lambda_i. \tag{14}$$

Using this lemma allows us to obtain following two lemmas. See Appendix J for the proofs.

Lemma 17 (Lower bound on the bias for any non-negative regularization) *There exist constants b, c that only depend on σ_x such that the following holds: suppose that assumptions IndepCoord and PriorSigns($\bar{\theta}$) hold. Take $k = \min\{\kappa : \rho_\kappa(0) > b\}$ and suppose that $k > 0$. Then with probability at least $1 - ce^{-n/c}$ for any $\lambda \geq 0$*

$$\mathbb{E}_{\theta^*} B \geq \frac{1}{c} \|\bar{\theta}_{0:k}\|^2_{\Sigma_{0:k}^{-1}} \frac{\left(\sum_{i>k} \lambda_i\right)^2}{n^2}.$$

Lemma 18 (Upper bound on excess risk for some negative regularization) *There exists a constant c that only depends on σ_x such that the following holds: suppose that assumptions PriorSigns($\bar{\theta}$) and IndepCoord hold and that $\rho_k(0) > c$ for some $k < n/c$. Assume also that*

$$v_\varepsilon^2 \le \frac{1}{c}\|\bar{\theta}_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \frac{\left(\sum_{i>k}\lambda_i\right)^2}{n^3\left(\sum_{i>k}\lambda_i^2\right)^2}. \tag{15}$$

Then there exists such $\lambda < 0$ that with probability at least $1 - ce^{-n/c}$

$$\mathbb{E}_{\theta^*} B + v_\varepsilon^2 V \le c\left(v_\varepsilon^2\frac{k}{n} + v_\varepsilon\|\bar{\theta}_{0:k}\|_{\Sigma^{-1}_{0:k}}\sqrt{\frac{\sum_{i>k}\lambda_i^2}{n}} + \|\bar{\theta}_{0:k}\|^2_{\Sigma^{-1}_{0:k}}\frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n} + \|\bar{\theta}_{k:\infty}\|^2_{\Sigma_{k:\infty}}\right).$$

Lemma 17 provides a lower bound on the expected (over noise and θ^*) excess risk which holds w.h.p. uniformly over all non-negative λ. Lemma 18 provides an upper bound that can be achieved by some negative λ. Combining these two lemmas gives a sufficient condition for the optimal λ to be negative, which is given by the following theorem.

Theorem 19 *There exist constants b and c that only depend on σ_x such that the following holds. Suppose that assumptions PriorSigns($\bar{\theta}$) and IndepCoord hold. Take $k = \min\{\kappa : \rho_\kappa(0) > b\}$ and suppose that $k < n/c$. The value of λ that minimizes $\mathbb{E}_{\theta^*} B + v_\varepsilon V$ will be negative with probability at least $1 - ce^{-n/c}$ if the following conditions are satisfied:*

$$\begin{aligned}
&\textit{small noise:} && v_\varepsilon^2 \le \frac{\|\bar{\theta}_{0:k}\|^2_{\Sigma^{-1}_{0:k}}}{c}\min\left(\frac{\left(\sum_{i>k}\lambda_i\right)^2}{nk}, \frac{\left(\sum_{i>k}\lambda_i\right)^4}{n^3\sum_{i>k}\lambda_i^2}\right),\\
&\textit{jump in effective rank:} && \rho_k(0) > c,\\
&\textit{small signal in the tail:} && \|\bar{\theta}_{k:\infty}\|^2_{\Sigma_{k:\infty}} \le \frac{1}{c}\|\bar{\theta}_{0:k}\|^2_{\Sigma^{-1}_{0:k}}\left(\frac{\sum_{i>k}\lambda_i}{n}\right)^2.
\end{aligned}$$

Proof It is easy to see that by taking c large enough, the conditions of Lemmas 18 and 17 are satisfied, and the upper bound in Lemma 18 becomes lower than the lower bound in Lemma 17. ■

We see that the conditions indeed align with the intuition outlined in the beginning of this section: we need small variance, small signal in the tail, and a sharp jump in effective rank. However, we do not have matching lower bounds in the critical regime when Assumption *NoncritReg*(k, γ) is not satisfied for a constant $\gamma > 1$. Thus, we don't know whether these conditions are also necessary.

9. Comparison to other works

As we mentioned in Section 1.2, recently there has been a number of papers studying population risk of interpolating solutions of linear regression, and we gave a rough split of those results into three categories there. Here we elaborate on the comparison between the approaches and results.

Results from the first category (Dobriban and Wager, 2015; Hastie et al., 2019; Wu and Xu, 2020; Richards et al., 2020) compute exact asymptotic expressions for the excess risk assuming that p/n

goes to some constant as p, n go to infinity, and that the spectral distribution of Σ converges to some limiting distribution. From the point of view of our approach, such distributions are indistinguishable from isotropic: indeed, the very existence of limiting spectral measure implies that almost all eigenvalues are within a constant factor of each other. Many of those works even assume explicitly that the spectrum of Σ is upper- and lower-bounded by two constants (Richards et al., 2020, page 7), (Wu and Xu, 2020, Assumption 1), (Hastie et al., 2019, Theorem 3). Our results don't need any asymptotic set up, and apply to $p = \infty$ with some fixed summable sequence λ_i, which has no meaningful notion of limiting distribution, and no separation from zero is needed. For example, our setup covers kernel regression with a fixed kernel and increasing number of data points. On the other hand, when all λ_i are within a constant factor of each other, our lower bounds become $B \geq \|\theta^*\|_\Sigma / c$ and $V \geq 1/c$, so the constant part of the whole signal doesn't get learned and the variance term is at least a constant, i.e., the asymptotic expressions obtained in the works from this category are all just different constants and our approach cannot distinguish them. Therefore, we answer significantly different questions: while the asymptotic work distinguishes between constant error rates, we investigate when the error can be less than a constant. The final difference with our work is rather technical but quite strong: all the works in this category assume that the coordinates of the data become independent if multiplied by the inverse square root of the covariance. This assumption stems from asymptotic random matrix theory techniques, on which these papers are based. To the best of our knowledge, it is not known how to extend these techniques beyond random matrices with independent elements. Our approach, however, does not require the coordinates to be independent.

When it comes to the second category, featurized or kernel regression (Montanari and Zhong, 2020; Ghorbani et al., 2020b; Mei and Montanari, 2019; Ghorbani et al., 2020a; Liang et al., 2020), the difference from our approach is that we do not assume any particular mechanism for data generation or how the features are constructed, but we directly make assumptions about feature vectors. Our results can in principle be applied in this setting if one computes the spectrum of the population covariance for particular features or kernels and the corresponding sub-Gaussian norms. The major difficulty that precludes such a direct comparison is that that computation is not straightforward. The works from this category operate in a more particular setting and circumvent the computation of the spectrum of Σ. On the other hand, it is not hard to trace strong similarities with our approach on the level of the proof. First of all, all the papers in this category that we are aware of assume that the data comes from a very regular distribution: either d-dimensional isotropic data with i.i.d. coordinates (Liang et al., 2020, Assumption 1), or data from the uniform distribution on the sphere (Mei and Montanari, 2019; Ghorbani et al., 2020a, abstracts), (Montanari and Zhong, 2020, Section 3.2), or data from the product of two uniform distributions on spheres (Ghorbani et al., 2020b, Section 2.1). Second, in all those papers the kernel is either spherically symmetric (Ghorbani et al., 2020b, Section 2.2), (Liang et al., 2020, Equation 4) or close to being spherically symmetric due to isotropic initialization of the neural network or isotropic choice of random features (Ghorbani et al., 2020a, Assumption 1), (Mei and Montanari, 2019, Thorem 2), (Montanari and Zhong, 2020, Section 3.2). After that, they consider the regime where n is large compared to d^α for some α (Montanari and Zhong, 2020, Assumption 1), (Ghorbani et al., 2020b, Theorem 1), (Mei and Montanari, 2019; Ghorbani et al., 2020a; Liang et al., 2020, abstracts)[12]. Finally, all those papers derive that kernel regression works effectively as ridge regression with polynomial features up to

12. In (Mei and Montanari, 2019) $\alpha = 1$.

degree α (Montanari and Zhong, 2020; Ghorbani et al., 2020a, abstracts), (Ghorbani et al., 2020b, Theorem 1), (Liang et al., 2020, Proposition 1 and Section 2.3). The only exception is Mei and Montanari (2019), who derive asymptotic expressions for excess risk when the true function is affine (i.e., a polynomial of degree 1) plus Gaussian misspecification. The connection with our results is that in such a regime (uniform distribution on the sphere, spherically symmetric kernel) polynomials are exactly the eigenfunctions of the kernel operator, which plays the role of the covariance operator, and there are $k \approx d^\alpha$ of polynomials of degree at most α. Thus, their approach is similar to ours: separate the first k eigendirections (or their approximations) and show that other directions act as regularization.

The third category is where this paper belongs, so a more concrete comparison to other results is possible. Sections 3.2 and 3.3 provide a detailed explanation of how our work generalizes the work of Bartlett et al. (2020). Kobak et al. (2020) proved that negative ridge regularization is optimal in a spiked covariance model with one spike, which is a simple particular case with $k = 1$ of our results. In Section 8, we showed that negative regularization is optimal under a rich set of covariance structures, and gave general sufficient conditions. Chinot and Lerasle (2021) obtain non-asymptotic bounds for bias and variance in the ridgeless setting. They assume Gaussian data and the existence of k^*, which means that our results apply in their setting. Our bound for the bias term is tight, so it cannot be worse than theirs by more than a constant multiplier. At the same time, their bound on the bias term can be much worse than ours: note that their bound depends on $\|\theta^*\|$, while our bound scales with $\|\theta^*\|_\Sigma$, therefore their bound can be arbitrarily close to infinity while our bound stays finite. When it comes to the variance term, the bound of Chinot and Lerasle (2021) is larger but holds with smaller probability, as they discuss when they compare their results to those in Bartlett et al. (2020). Dereziński et al. (2019) start with an arbitrary covariance matrix and construct a specific data distribution for which the approximation error $\mathbb{E}\|\hat{\theta} - \theta^*\|^2$ has an explicit expression. We provide bounds for the excess risk $\|\hat{\theta} - \theta^*\|_\Sigma^2$, so our results are not directly comparable to theirs. Dereziński et al. (2020) consider expectation of the projector on the orthogonal complement to the span of i.i.d. data with arbitrary covariance and derive tight upper and lower bounds for it with respect to Loewner order. The bias term in our setting is exactly such a projection of θ^*, but measured in $\|\cdot\|_\Sigma$. Because of this mismatch in the norm, the results of Dereziński et al. (2020) do not translate into our results directly, even if we consider the expectation of the bias term.

10. Conclusions

We studied the excess risk of ridge regression and showed how geometry of the data can influence both which part of the signal is learned and how the noise is damped. For a range of values of the regularization parameter we showed that learning can be seen as the composition of two parts: classical ridge regression in the first k components (the "essentially low-dimensional part") and learning the zero estimator in the rest of the components (the "essentially high-dimensional part"). We introduced a general assumption under which the data is "essentially high-dimensional", and provided geometric sufficient conditions for its satisfaction. Moreover, we investigated the regime in which the "essentially high-dimensional part" is too high-dimensional, and derived general sufficient conditions for negative regularization to be optimal: small noise, small energy of the "essentially low-dimensional part", but an abrupt jump in the effective rank.

On the technical side, our proof decouples cleanly into an algebraic part, which holds with probability 1 for non-negative regularization,[13] and the probabilistic part, where we plug in well-known concentration results from high-dimensional probability. This makes it easy to trace how different terms in the bound correspond to the parts of the estimator, and supports the geometric interpretation given above. One downside of this approach is that all the results are obtained with high probability w.r.t. X, and it is not clear whether they hold in expectation. Indeed, the variance term can blow up if the matrix A becomes close to degenerate, so one needs to be able to control the probability of the smallest eigenvalue of A approaching zero to obtain such a result. This problem is notoriously difficult, and we see this as an interesting topic of future research and a new motivation for studying the smallest singular values of random Gram matrices.

We provided a thorough overview of the related papers, and explained how our results are significantly different from them despite some optical similarities. Those similarities, however, are intriguing, and hint at the task of developing a unified treatment of different regimes of overparameterized linear regression as a promising direction of future work.

Acknowledgements

We gratefully acknowledge the support of the NSF through grants DMS-2023505 and DMS-2031883 and of the Simons Foundation through award #814639.

13. For the case of negative regularization we need to condition on the event that all the necessary symmetric matrices are PD.

Appendix A. Definitions and Notation

A.1 Sub-Gaussianity

A random variable z is sub-Gaussian if it has a finite sub-Gaussian norm

$$\|z\|_{\psi_2} := \inf \left\{ t > 0 : \mathbb{E} \exp(z^2/t^2) \leq 2 \right\}.$$

The sub-Gaussian norm of a random vector Z is

$$\|Z\|_{\psi_2} := \sup_{s \neq 0} \|\langle s, Z \rangle / \|s\| \rangle\|_{\psi_2}.$$

A.2 Standard mathematical objects

- $M[i, j]$ denotes the element of the matrix M which stands at the intersection of the i-th row and j-th column.
- $\|v\|$ denotes the Euclidean norm for a vector v in $\mathbb{R}^d$ for some d, .
- $\|M\|$ denotes the operator norm (i.e., maximum singular value) for a matrix M in $\mathbb{R}^{m \times n}$ for some m, n.
- $\mathrm{tr}(M)$ denotes the trace of a square matrix M.
- $\|M\|_F$ denotes the Frobenius norm for a matrix M in $\mathbb{R}^{m \times n}$, i.e., $\|M\|_F := \sqrt{\mathrm{tr}(MM^\top)}$.
- $\mathcal{S}^{p-1}$ denotes the unit sphere in $\mathbb{R}^p$, i.e., $\mathcal{S}^{p-1} = \{x \in \mathbb{R}^p : \|x\| = 1\}$.
- I_m is the $m \times m$ identity matrix.
- $\|v\|_M := \sqrt{v^\top M v}$ for any positive semidefinite (PSD) matrix $M \in \mathbb{R}^{m \times m}$ and any $v \in \mathbb{R}^m$.
- $\mu_1(M) \geq \cdots \geq \mu_m(M)$ are the eigenvalues of a symmetric matrix $M \in \mathbb{R}^{m \times m}$ in decreasing order.

A.3 Data and the learning procedure

Recall from Section 2 that

- $X \in \mathbb{R}^{n \times p}$ — a random matrix with i.i.d. centered rows.
- $y = X\theta^* + \varepsilon$ is the response vector, where $\theta^* \in \mathbb{R}^p$ is some unknown vector, and ε is noise,
- components of ε are independent and have variance v_ε,
- $\{x^i\}_{i=1}^n$ are columns of $X^\top$ (i.e., $\{x^i\}_{i=1}^n$ are our i.i.d. data points in $\mathbb{R}^p$).
- x denotes a new random draw from the data distribution, i.e., x is independent from X, ε and x has the same distribution as x^1.
- $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ is the covariance matrix of a row of X.
- $Z = X\Sigma^{-1/2}$, $\{z_i\}_{i=1}^p$ are columns of Z.
- the rows of Z are sub-Gaussian with sub-Gaussian norm at most σ_x,
- ridge regression outputs $\hat{\theta}(y) := X^\top (\lambda I_n + XX^\top)^{-1} y$.

A.4 Splitting the coordinates

For some $k < n$ we spit the coordinates into two groups: the first k components and the rest of the components. Thus we introduce the following notation. Consider integers a, b from 0 to ∞ (we always either take $a = 0$ and $b = k$ or $a = k$ and $b = \infty$).

- For any matrix $M \in \mathbb{R}^{n\times p}$ denote $M_{a:b}$ to be the matrix that is comprised of the columns of M from $a + 1$-st to b-th.
- For any vector $\eta \in \mathbb{R}^p$ denote $\eta_{a:b}$ to be the vector comprised of components of η from $a+1$-st to b-th.
- $\Sigma_{0:k} = \operatorname{diag}(\lambda_1, \dots, \lambda_k)$, and $\Sigma_{k:\infty} = \operatorname{diag}(\lambda_{k+1}, \lambda_{k+2}, \dots)$.
- $A_k = \lambda I_n + X_{k:\infty} X_{k:\infty}^\top$.
- $r_k = \frac{1}{\lambda_{k+1}} \left(\lambda + \sum_{i>k} \lambda_i\right)$.
- $\rho_k = r_k/n$.
- $\rho_k(0) = \frac{1}{n\lambda_{k+1}} \sum_{i>k} \lambda_i$.
- For any i we denote $A_{-i} = A_{-i} := X_{0:i-1} X_{0:i-1}^\top + X_{i:\infty} X_{i:\infty}^\top + \lambda I_n$.

Appendix B. Ridge regression

We are interested in evaluating the MSE of the ridge estimator. For positive regularization parameter λ that estimator is defined as

$$\begin{aligned}\hat{\theta}(y) = \hat{\theta}(y) &= \operatorname*{argmin}_\theta \left\{\|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2\right\} \\ &= \left(\lambda I_p + X^\top X\right)^{-1} X^\top y.\end{aligned}$$

In the overparametrized case (i.e., $p > n$), however, the latter expression has a singularity at zero, because the matrix $X^\top X$ does not have full rank. If $\lambda = 0$ the solution to the minimization problem above is not unique. Moreover, if $\lambda < 0$, no solution exists at all because we are minimizing a quadratic form whose matrix has negative singular values. To alleviate these issues and extend the definition of the solution to non-positive values of λ, we propose the following: since the matrix $X^\top X$ doesn't have full rank, we can apply the Sherman-Morrison-Woodbury formula:

$$\left(\lambda I_p + X^\top X\right)^{-1} = \lambda^{-1} I_p - \lambda^{-2} X^\top (I_n + \lambda^{-1} X X^\top)^{-1} X.$$

So,

$$\begin{aligned}\hat{\theta}(y) = &\left(\lambda I_p + X^\top X\right)^{-1} X^\top \\ =&\lambda^{-1}X^\top - \lambda^{-2}X^\top(I_n + \lambda^{-1}XX^\top)^{-1}XX^\top \\ =&\lambda^{-1}X^\top - \lambda^{-1}X^\top(I_n + \lambda^{-1}XX^\top)^{-1}(\lambda^{-1}XX^\top + I_n - I_n) \\ =&\lambda^{-1}X^\top(I_n + \lambda^{-1}XX^\top)^{-1} \\ =&X^\top(\lambda I_n + XX^\top)^{-1}y.\end{aligned}$$

The matrix $XX^\top$ has full rank, and the expression above is continuous in λ as long as $XX^\top + \lambda I_n$ stays PD. When $\lambda = 0$, $X^\top(\lambda I_n + XX^\top)^{-1}y$ is the minimum norm interpolating solution (the same solution that was considered in (Bartlett et al., 2020). Therefore, we use the expression

$$\hat{\theta}(y) := X^\top(\lambda I_n + XX^\top)^{-1}y$$

to define the ridge regression solution for any $\lambda > -\mu_n(XX^\top)$.

Note that $\hat{\theta}(y)$ is linear in y. Since we have $y = X\theta^* + \varepsilon$ we can also write

$$\hat{\theta}(y) = \hat{\theta}(X\theta^*) + \hat{\theta}(\varepsilon).$$

The first term is the noiseless estimate; its error gives the bias term. The second term is the estimate obtained when the signal is pure noise. It gives the variance term.

For the full MSE we have

$$\begin{aligned}\|\hat{\theta}(y) - \theta^*\|_\Sigma^2 =&\|\hat{\theta}(X\theta^*) + \hat{\theta}(\varepsilon) - \theta^*\|_\Sigma^2 \\ \le&2\|\hat{\theta}(X\theta^*) - \theta^*\|_\Sigma^2 + 2\|\hat{\theta}(\varepsilon)\|_\Sigma^2 \\ =&2(B + V_\varepsilon),\end{aligned}$$

where we introduced bias B and variance V_ε:

$$\begin{aligned}B :=& \|\hat{\theta}(X\theta^*) - \theta^*\|_\Sigma^2 = \|(I_p - X^\top(\lambda I_n + XX^\top)^{-1}X)\theta^*\|_\Sigma^2, \\ V_\varepsilon :=&\|\hat{\theta}(\varepsilon)\|_\Sigma^2 \qquad\qquad = \|X^\top(\lambda I_n + XX^\top)^{-1}\varepsilon\|_\Sigma^2.\end{aligned}$$

Finally, since V_ε is a quadratic form in ε, by Lemma 22 if the noise is sub-Gaussian, then its value is controlled by its expectation with high probability. That expectation, in its turn, scales linearly with the variance v_ε^2 of the noise. Therefore, we can decouple the effect of the noise and only study the following purified variance term:

$$\begin{aligned}V :=&\frac{1}{v_\varepsilon^2}\mathbb{E}_\varepsilon V_\varepsilon \\ =&\operatorname{tr}((\lambda I_n + XX^\top)^{-1}X\Sigma X^\top(\lambda I_n + XX^\top)^{-1}) \\ =&\operatorname{tr}(\Sigma X^\top(\lambda I_n + XX^\top)^{-2}X).\end{aligned}$$

The main aim of our work is to give sharp non-asymptotic bounds for B and V.

Appendix C. Concentration inequalities

Lemma 20 (Non-standard norms of sub-Gaussian vectors) *Suppose z is a sub-Gaussian vector in $\mathbb{R}^p$ with $\|z\|_{\psi_2} \leq \sigma$. Consider $\Sigma = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ for some positive non-increasing sequence $\{\lambda_i\}_{i=1}^p$. Then for some absolute constant c for any $t > 0$*

$$\mathbb{P}\left\{\|\Sigma^{1/2}z\|^2 > c\sigma^2\left(t\lambda_1 + \sum_i \lambda_i\right)\right\} \leq 2e^{-t/c}.$$

Proof The argument consists of two parts: first, we obtain a bound that only works well in the case when all λ_i are approximately the same. Next, we split the sequence $\{\lambda_i\}$ into pieces with approximately equal values within each piece and obtain the final result by applying the first part of the argument to each piece.

First part: Consider a $1/4$-net $\{u_j\}_{j=1}^m$ on $\mathcal{S}^{p-1}$, such that $m \leq 9^p$. Note that for any vector $v \in \mathcal{S}^{p-1}$ there exists an element u_j of that net such that $\langle v, u_j\rangle \geq 3/4 \cdot \|v\|$. Thus, we have

$$\|\Sigma^{1/2}z\| \leq \frac{4}{3}\sqrt{\lambda_1}\max_j\langle z, u_j\rangle \leq 2\sqrt{\lambda_1}\max_j\langle z, u_j\rangle.$$

Since the random variable $\langle z, u_j\rangle$ is σ-sub-Gaussian, it also holds for any $t > 0$ and some absolute constant c that

$$\begin{aligned}\mathbb{P}(|\langle z, u_j\rangle| > t) &\leq 2e^{-ct^2/\sigma^2},\\ \mathbb{P}(4\lambda_1\langle z, u_j\rangle^2 > 4\lambda_1 t\sigma^2) &\leq 2e^{-ct}.\end{aligned}$$

By multiplicity correction, we obtain

$$\mathbb{P}\left(\|\Sigma^{1/2}z\|^2 > 4\lambda_1\sigma^2 t + \frac{4\sigma^2\lambda_1\log 9}{c}p\right) \leq 2e^{-ct}.$$

We see that the random variable $\left(\|\Sigma^{1/2}z\|^2 - \frac{4\sigma^2\lambda_1\log 9}{c}p\right)_+$ has sub-exponential norm bounded by $C\sigma^2\lambda_1$.

Second part: Now, instead of applying the result that we have just obtained to the whole vector z, split it in the following way: define the sub-sequence $\{i_j\}$ in such that $i_1 = 1$, and for any $l \geq 1$ $i_{l+1} = \min\{i : \lambda_i < \lambda_{i_l}/2\}$. Denote z_l to be a sub-vector of z comprised of components from the i_l-th to $(i_{l+1} - 1)$-th. Let $\Sigma_l = \operatorname{diag}(\lambda_{i_l}, \ldots, \lambda_{i_{l+1}-1})$.

Then by the initial argument, the random variable $\left(\|\Sigma_l^{1/2}z_l\|^2 - \frac{4\sigma^2\lambda_{i_l}\log 9}{c}(i_{l+1} - i_l)\right)_+$ has sub-exponential norm bounded by $C\sigma^2\lambda_{i_l}$. Since each next λ_{i_l} is at most half of the previous, we obtain that the sum (over l) of those random variables has sub-exponential norm at most $2C\sigma^2\lambda_1$. Combining this with the fact that

$$\sum_{i=i_l}^{i_{l+1}-1}\lambda_i \geq (i_{l+1} - i_l)\lambda_{i_{l+1}-1} \geq (i_{l+1} - i_l)\lambda_{i_{l+1}}/2,$$

we obtain that for some absolute constants $c_0, c_1, \dots$ for any $t > 0$

$$\begin{aligned} 2e^{-c_0 t} \geq & \mathbb{P}\left\{ \sum_l \left(\|\Sigma_l^{1/2} z_l\|^2 - c_1 \sigma^2 \lambda_{i_l}(i_{l+1} - i_l) \right) > c_2 \sigma^2 \lambda_1 t \right\} \\ \geq & \mathbb{P}\left\{ \|\Sigma^{1/2} z\|^2 \geq c_3 \sigma^2 \sum_i \lambda_i + c_2 \sigma^2 \lambda_1 t \right\}. \end{aligned}$$

■

Lemma 21 (Concentration of the sum of squared norms) *Suppose $Z \in \mathbb{R}^{n \times p}$ is a matrix with independent isotropic sub-Gaussian rows with $\|Z[i,*]\|_{\psi_2} \leq \sigma$. Consider $\Sigma = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ for some positive non-increasing sequence $\{\lambda_i\}_{i=1}^p$. Then for some absolute constant c and any $t \in (0, n)$ with probability at least $1 - 2\exp(-ct)$,*

$$(n - \sqrt{nt}\sigma^2) \sum_{i>k} \lambda_i \leq \sum_{i=1}^n \|\Sigma_{k:\infty}^{1/2} Z_{i,k:\infty}\|^2 \leq (n + \sqrt{nt}\sigma^2) \sum_{i>k} \lambda_i.$$

Proof Since $\{Z_{i,k:\infty}\}_{i=1}^n$ are independent, isotropic and sub-Gaussian, $\|\Sigma_{k:\infty}^{1/2} Z_{i,k:\infty}\|^2$ are independent sub-exponential r.v.'s with expectation $\sum_{i>k} \lambda_i$ and sub-exponential norms bounded by $c_1 \sigma^2 \sum_{i>k} \lambda_i$. Applying Bernstein's inequality gives

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n \|\Sigma_{k:\infty}^{1/2} Z_{i,k:\infty}\|^2 - \sum_{i>k} \lambda_i \right| \geq t\sigma^2 \sum_{i>k} \lambda_i \right) \leq 2\exp\left(-c_2 \min(t, t^2) n\right).$$

Changing t to $\sqrt{t/n}$ gives the result. ■

Lemma 22 (Weakened Hanson-Wright inequality) *Suppose $M \in \mathbb{R}^{n \times n}$ is a (random) PSD matrix and $\varepsilon \in \mathbb{R}^n$ is a centered vector whose components $\{\varepsilon_i\}_{i=1}^n$ are independent and have sub-Gaussian norm at most σ. Then for some absolute constants c, C and any $t > 1$ with probability at least $1 - 2e^{-t/c}$,*

$$\varepsilon^\top M \varepsilon \leq C\sigma^2 t \operatorname{tr}(M).$$

Proof By Theorem 6.2.1 (Hanson-Wright inequality) in (Vershynin, 2018), for some absolute constant c_1 for any $t > 0$,

$$\mathbb{P}_M\left\{ |\varepsilon^\top M \varepsilon - \mathbb{E}\varepsilon^\top M \varepsilon| \geq t \right\} \leq 2\exp\left(-c_1 \min\left\{ \frac{t^2}{\|M\|_F^2 \sigma^4}, \frac{t}{\|M\|\sigma^2} \right\} \right),$$

where $\mathbb{P}_M$ denotes conditional probability given M.

Since for any i, $\mathbb{E}\varepsilon_i = 0$, and $\operatorname{Var}(\varepsilon_i) \lesssim \sigma^2$, and since M is PSD, we have

$$\mathbb{E}\varepsilon^\top M \varepsilon \leq c_2 \sigma^2 \operatorname{tr}(M).$$

Moreover, since $\|M\|_F^2 \le \operatorname{tr}(M)^2$ and $\|M\| \le \operatorname{tr}(M)$, we obtain

$$\mathbb{P}_M\left\{\varepsilon^\top M\varepsilon > \sigma^2(c_2+t)\operatorname{tr}(M)\right\} \le 2\exp\{-c_1\min(t,t^2)\}.$$

Restricting to $t > 1$ and adjusting the constants gives the result (note that since the RHS doesn't depend on M, we can replace $\mathbb{P}_M$ with $\mathbb{P}$). ■

Appendix D. Controlling the singular values

Lemma 23 (Bound on the norm of non-diagonal part of a Gram matrix) *Denote $\mathring{A}_k$ to be the matrix A_k with zeroed out diagonal elements: $\mathring{A}_k[i,j] = (1-\delta_{i,j})A_k[i,j]$. Then for some absolute constant c for any $t > 0$ with probability at least $1 - 4e^{-t/c}$,*

$$\|\mathring{A}_k\| \le c\sigma^2\sqrt{(t+n)\left(\lambda_{k+1}^2(t+n) + \sum_{i>k}\lambda_i^2\right)}.$$

Proof We follow the lines of the decoupling argument from Vershynin (2012). Consider a $1/4$-net $\{u_j\}_{j=1}^m$ on $\mathcal{S}^{n-1}$ s.t. $m \le 9^n$. Then

$$\|\mathring{A}_k\| \le 2\max_j |u_j^\top \mathring{A}_k u_j|.$$

Indeed, take $v \in \mathcal{S}^{n-1}$ to be the eigenvector of $\mathring{A}_k$ whose eigenvalue has the largest absolute value μ (i.e., $\|\mathring{A}_k\| = \mu$), and let u_j be the closest point in the net to v. Then

$$\begin{aligned}
\|v - u_j\| \le& 1/4,\\
u_j^\top v \ge& 3/4,\\
|u_j^\top \mathring{A}_k u_j| \ge& |u_j^\top \mathring{A}_k v| - |u_j^\top \mathring{A}_k(v-u_j)|\\
=& |\mu| u_j^\top v - |u_j^\top \mathring{A}_k(v-u_j)|\\
\ge& |\mu| u_j^\top v - \|u_j\|\|\mathring{A}_k\|\|v-u_j\|\\
\ge& |\mu|\left(\frac{3}{4} - \frac{1}{4}\right).
\end{aligned}$$

Denote the k-th coordinate of u_j as $u_j[k]$. Note that

$$u_j^\top \mathring{A}_k u_j = 4\mathbb{E}_T \sum_{k\in T\not\ni l} u_j[k]u_j[l]\mathring{A}_k[k,l],$$

where the expectation is taken over a uniformly chosen random subset T of $\{1,\dots,n\}$ (since $\mathring{A}_k$ has zeroed-out diagonal, we don't need to consider terms with $m = l$ which allows us to sum over

$k \in T \not\ni l$). Thus,

$$|u_j^\top \mathring{A}_k u_j| \le 4 \max_T \left| \sum_{l \in T \not\ni m} u_j[l] u_j[m] \mathring{A}_k[l,m] \right|$$
$$= 4 \max_T \left| \left\langle \sum_{l \in T} u_j[l] X_{k:\infty}[l,*], \sum_{m \notin T} u_j[m] X_{k:\infty}[m,*] \right\rangle \right|.$$

Fix j and denote

$$\xi^\top := \sum_{l \in T} u_j[l] X_{k:\infty}[l,*] \Sigma_{k:\infty}^{-1/2},$$
$$\eta^\top := \sum_{m \notin T} u_j[m] X_{k:\infty}[m,*] \Sigma_{k:\infty}^{-1/2}.$$

Note that since u_j is from the sphere, $\{X_{k:\infty}[i,*]\}_{i=1}^n$ are independent, and l, m live in disjoint subsets, the vectors ξ and η are independent sub-Gaussian with sub-Gaussian norms bounded by $C\sigma$ for some absolute constant C.

First, that means that for some absolute constant c_1 we have

$$\mathbb{P}\left\{ \left| \left\langle \Sigma^{1/2}\xi, \Sigma^{1/2}\eta \right\rangle \right| \ge t\sigma \|\Sigma\eta\| \right\} \le 2e^{-c_1 t^2}.$$

Second, by Lemma 20, for some constant c_2 for any $t > 0$

$$\mathbb{P}\left\{ \|\Sigma\eta\|^2 \ge c_2\sigma^2 \left(\lambda_{k+1}^2 t + \sum_{i>k} \lambda_i^2 \right) \right\} \le 2e^{-t/c_2}.$$

We obtain that for some absolute constant c for any $t > 0$ with probability at least $1 - 4e^{-t/c}$

$$\left| \left\langle \Sigma^{1/2}\xi, \Sigma^{1/2}\eta \right\rangle \right| < c\sigma^2 \sqrt{t \left(\lambda_{k+1}^2 t + \sum_{i>k} \lambda_i^2 \right)}.$$

Finally, making multiplicity correction for all j (there are at most 9^n of them), and all subsets T (at most 2^n), we obtain that for some absolute constant c with probability at least $1 - 4e^{-t/c}$

$$\|\mathring{A}_k\| \le c\sigma^2 \sqrt{(t+n) \left(\lambda_{k+1}^2 (t+n) + \sum_{i>k} \lambda_i^2 \right)}.$$

■

Lemma 24 *For some absolute constant c, for any $t > 0$, with probability at least $1 - 6e^{-t/c}$,*

$$\|X_{k:\infty} X_{k:\infty}^\top\| \le c\sigma_x^2 \left(\lambda_{k+1}(t+n) + \sum_{i>k} \lambda_i \right).$$

Proof Note that $\|A\| \le \max_i \|X_{i,*}\| + \|\mathring{A}\|$. Combining Lemma 20 (with multiplicity correction) and Lemma 23 gives with probability $1 - 6e^{-t/c_1}$

$$\|A\| \le c_1\sigma^2 \left((t + c_1 \log n)\lambda_1 + \sum_i \lambda_i + \sqrt{(t+n)\left(\lambda_1^2(t+n) + \sum_i \lambda_i^2\right)} \right).$$

Now note that

$$\begin{aligned}(t + c_1 \log n)\lambda_1 \le& c_1\sqrt{(t+n)\left(\lambda_1^2(t+n) + \sum_i \lambda_i^2\right)} \\ \le& c_1\sqrt{\lambda_1^2(t+n)^2 + \lambda_1(t+n)\sum_i \lambda_i} \\ \le& c_1\left(\lambda_1(t+n) + \sum_i \lambda_i\right),\end{aligned}$$

where we used $\sqrt{a^2 + ab} \le a + b$ in the last transition. Removing the dominated (up to a constant multiplier) terms gives the result.

■

Lemma 3 (Controlling $\mu_1(A_k)/\mu_n(A_k)$ under sub-Gaussianity) *For any $\gamma \in [0, 1)$ and $\sigma_x > 0$ there exists $c > 0$ that only depends on σ_x and γ such that under Assumption NoncritReg(k, γ) the following holds: for any $L \ge 1$*

- *If $\rho_k \ge L^2$ and with probability at least $(1 - \delta)^{1/n}$*

$$\lambda + \|x_{k:\infty}\|^2 \ge \frac{c}{L}\left(\lambda + \mathbb{E}\|x_{k:\infty}\|^2\right),$$

  then with probability at least $1 - \delta - ce^{-n/c}$

$$\mu_n(A_k) \ge L^{-1}\mu_1(A_k).$$

- *Suppose that it is known that with probability at least $ce^{-n/c}$ $\mu_n(A_k) \ge L^{-1}\mu_1(A_k)$. Then $\rho_k \ge \frac{1}{cL}$ and with probability at least $\left(1 - ce^{-n/c}\right)^{1/n}$*

$$\lambda + \|x_{k:\infty}\|^2 \ge \frac{1}{cL}\left(\lambda + \mathbb{E}\|x_{k:\infty}\|^2\right).$$

Proof We start with the high-probability bounds that we can derive assuming only sub-Gaussianity and independence of data vectors. By Lemma 20, for some absolute constant c and for any $t > 0$,

$$\mathbb{P}\left\{ \|X_{k:\infty}[i, *]\|^2 > c\sigma_x^2\left(t\lambda_{k+1} + \sum_{i>k}\lambda_i\right)\right\} \le 2e^{-t/c}.$$

By Lemma 23, for some absolute constant c and for any $t > 0$, with probability at least $1 - 4e^{-t/c}$,

$$\|\mathring{A}_k\| \leq c\sigma_x^2 \sqrt{(t+n)\left(\lambda_{k+1}^2(t+n) + \sum_{i>k}\lambda_i^2\right)}.$$

Since $\|A_k\| \leq \lambda + \|\mathring{A}_k\| + \max_i \|X_{k:\infty}[i,*]\|$, the above two statements imply that for any $t > 0$ with probability at least $1 - 4e^{-n/c} - 2ne^{-t/c}$,

$$\begin{aligned}
\mu_1(A_k) \leq& \lambda + c\sigma_x^2\sqrt{n\left(\lambda_{k+1}^2 n + \sum_{i>k}\lambda_i^2\right)} + c\sigma_x^2\left(t\lambda_{k+1} + \sum_{i>k}\lambda_i\right) \\
\leq& \lambda + 2c\sigma_x^2\left((t+n)\lambda_{k+1} + \sum_{i>k}\lambda_i + \sqrt{n\sum_{i>k}\lambda_i^2}\right) \\
\leq& \lambda + 3c\sigma_x^2\left((t+n)\lambda_{k+1} + \sum_{i>k}\lambda_i\right),
\end{aligned}$$

where we used the following chain of inequalities to make the last transition:

$$2\sqrt{n\sum_{i>k}\lambda_i^2} \leq 2\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} \leq n\lambda_{k+1} + \sum_{i>k}\lambda_i.$$

On the same event,

$$\mu_n(A_k) \geq \lambda + \min_i \|X_{k:\infty}[i,*]\|^2 - c\sigma_x^2\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

On the other hand, note that the sum of eigenvalues of A_k is equal to

$$\mathrm{tr}(A_k) = \lambda n + \sum_{i=1}^n \|\Sigma_{k:\infty}^{1/2} Z_{k:\infty}[i,*]^\top\|^2.$$

By Lemma 21, for some absolute constant c and any $t \in (0,n)$, with probability at least $1 - 2e^{-ct}$,

$$(n - \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i \leq \sum_{i=1}^n \|\Sigma_{k:\infty}^{1/2} Z_{k:\infty}[i,*]^\top\|^2 \leq (n + \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i.$$

On this event

$$\begin{aligned}
\mu_1(A_k) \geq& \lambda + \left(1 - \sqrt{\frac{t}{n}}\sigma_x^2\right)\sum_{i>k}\lambda_i, \\
\mu_n(A_k) \leq& \lambda + \left(1 + \sqrt{\frac{t}{n}}\sigma_x^2\right)\sum_{i>k}\lambda_i.
\end{aligned}$$

Finally, note that $\mu_1(A_k) \geq \lambda_{k+1}\|Z_{k:\infty}[*,1]\|^2 + \lambda$. By Lemma 21, for some c_3 and for any $t \in (0,n)$, with probability. at least $1-2e^{-c_3 t}$,

$$\|Z_{k:\infty}[*,1]\|^2 \geq n - \sqrt{nt}\sigma_x^2,$$

which means that

$$\mu_1(A_k) \geq \lambda + n\lambda_{k+1}\left(1 - \sqrt{\frac{t}{n}}\sigma_x^2\right).$$

Combining all those bounds together gives that there is a constant c_x that only depends on σ_x such that with probability at least $1 - c_x e^{-n/c_x}$ all the following inequalities hold simultaneously:

$$\begin{aligned}
\mu_1(A_k) \leq& \lambda + c_x\left(n\lambda_{k+1} + \sum_{i>k}\lambda_i\right),\\
\mu_1(A_k) \geq& \lambda + \frac{1}{c_x}\sum_{i>k}\lambda_i,\\
\mu_1(A_k) \geq& \lambda + \frac{1}{c_x}n\lambda_{k+1},\\
\mu_n(A_k) \geq& \lambda + \min_i \|X_{k:\infty}[i,*]\|^2 - c_x\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right),\\
\mu_n(A_k) \leq& \lambda + c_x\sum_{i>k}\lambda_i,\\
\mu_n(A_k) \leq& \lambda + \min_i \|X_{k:\infty}[i,*]\|^2.
\end{aligned}$$

In view of the bounds that we derived above, the following inequality is a sufficient condition for the statement that with probability at least $1 - c_x e^{-n/c_x}$ the condition number of A_k does not exceed L:

$$\frac{1}{L}\left(\lambda + c_x\left(n\lambda_{k+1} + \sum_{i>k}\lambda_i\right)\right) \leq \lambda + \min_i \|X_{k:\infty}[i,*]\|^2 - c_x\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

Note that for any $\zeta > 0$

$$\sqrt{n\sum_{i>k}\lambda_i^2} < 2\sqrt{n\sum_{i>k}\lambda_i^2} \leq 2\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} \leq \zeta n\lambda_{k+1} + \zeta^{-1}\sum_{i>k}\lambda_i,$$

which implies that for any ζ the following is also a sufficient condition:

$$\lambda + \min_i \|X_{k:\infty}[i,*]\|^2 \geq \lambda L^{-1} + c_x(1 + L^{-1} + \zeta)n\lambda_{k+1} + c_x(L^{-1} + \zeta^{-1})\sum_{i>k}\lambda_i.$$

Recall that $\lambda > -\gamma\sum_{i>k}\lambda_i$, so

$$\sum_{i>k}\lambda_i \leq \frac{1}{1-\gamma}\left(\lambda + \sum_{i>k}\lambda_i\right),$$

which allows us to upper bound the right-hand side of that condition. We write

$$\begin{aligned}
&\lambda L^{-1} + c_x(1 + L^{-1} + \zeta)n\lambda_{k+1} + c_x(L^{-1} + \zeta^{-1})\sum_{i>k}\lambda_i \\
\leq& L^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right) + c_x(1 + L^{-1} + \zeta)\rho_k^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right) + \frac{c_x(L^{-1} + \zeta^{-1})}{1-\gamma}\left(\lambda + \sum_{i>k}\lambda_i\right) \\
=& \left(\lambda + \sum_{i>k}\lambda_i\right)\left(L^{-1}\left(1 + c_x\rho_k^{-1} + \frac{c_x}{1-\gamma}\right) + \rho_k^{-1}\left(c_x + c_x\zeta\right) + \frac{c_x\zeta^{-1}}{1-\gamma}\right).
\end{aligned}$$

Now take $\zeta = \rho_k^{1/2}$ and a constant c that is big enough depending on γ and c_x. Then if $\rho_k > L^2 > 1$ and with probability at least $1 - \delta$,

$$\lambda + \min_i \|X_{k:\infty}[i, *]\|^2 \geq \frac{c}{L}\left(\lambda + \sum_{i>k}\lambda_i\right),$$

then with probability at least $1 - \delta - c_x e^{-n/c_x}$,

$$\mu_n(A_k) \geq L^{-1}\mu_1(A_k).$$

Note that since the rows of $X_{k:\infty}$ are i.i.d., the first condition is equivalent to that with probability at least $(1-\delta)^{1/n}$

$$\lambda + \|X_{k:\infty}[1, *]\|^2 \geq \frac{c}{L}\left(\lambda + \sum_{i>k}\lambda_i\right).$$

Now let's derive a necessary condition. Suppose it is known that with probability at least $c_x e^{-n/c_x}$ $\mu_n(A_k) \geq L^{-1}\mu_1(A_k)$. Then

$$\begin{aligned}
\lambda + \min_i \|X_{k:\infty}[i, *]\|^2 \geq& \frac{1}{L}\left(\lambda + \frac{1}{c_x}\sum_{i>k}\lambda_i\right), \\
\lambda + c_x\sum_{i>k}\lambda_i \geq& \frac{1}{L}\left(\lambda + \frac{1}{c_x}n\lambda_{k+1}\right).
\end{aligned}$$

For the first equation, we can write

$$\begin{aligned}
\lambda + \min_i \|X_{k:\infty}[i, *]\|^2 &\geq \frac{1}{L}\left(\lambda + \frac{1}{c_x}\sum_{i>k}\lambda_i\right) \\
\lambda(1 - L^{-1} + L^{-1}c_x^{-1}) + \min_i \|X_{k:\infty}[i, *]\|^2 &\geq \frac{1}{Lc_x}\left(\lambda + \sum_{i>k}\lambda_i\right), \\
\lambda + \min_i \|X_{k:\infty}[i, *]\|^2 &\geq \frac{1}{Lc_x(1 - L^{-1} + L^{-1}c_x^{-1})}\left(\lambda + \sum_{i>k}\lambda_i\right) \\
&\geq \frac{1}{Lc_x}\left(\lambda + \sum_{i>k}\lambda_i\right),
\end{aligned}$$

where we used the fact that $c_x > 1$ and $L > 1$.

When it comes to the second equation, we write

$$\begin{aligned}
\lambda + c_x \sum_{i>k} \lambda_i \geq & \frac{1}{L}\left(\lambda + \frac{1}{c_x} n\lambda_{k+1}\right), \\
(L-1)\lambda + c_x L \sum_{i>k} \lambda_i \geq & \frac{1}{c_x} n\lambda_{k+1} = \frac{1}{c_x}\rho_k^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right), \\
(L-1)\left(\lambda + \sum_{i>k}\lambda_i\right) + (c_x L - L + 1)\sum_{i>k}\lambda_i \geq & \frac{1}{c_x}\rho_k^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right) \\
\left(L - 1 + \frac{c_x L - L + 1}{1-\gamma}\right)\left(\lambda + \sum_{i>k}\lambda_i\right) \geq & \frac{1}{c_x}\rho_k^{-1}\left(\lambda + \sum_{i>k}\lambda_i\right) \\
\rho_k \geq & c_x^{-1}\left(L - 1 + \frac{c_x L - L + 1}{1-\gamma}\right)^{-1} \geq c^{-1}L^{-1},
\end{aligned}$$

where c is a large enough constant that only depends on γ and c_x. ■

Lemma 25 *Suppose assumptions NoncritReg(k,γ) and CondNum(k,δ,L) are satisfied and $\gamma < 1$. Then for some absolute constant c for any $t \in (0,n)$ with probability at least $1 - \delta - 2e^{-ct}$*

$$\frac{1}{L}\left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda + \sum_{i>k}\lambda_i\right) \leq \mu_n(A_k) \leq \mu_1(A_k) \leq L\left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda + \sum_{i>k}\lambda_i\right).$$

Moreover, if $\delta < 1 - 4e^{-ct}$ for some $t \in (0,n)$, then

$$\frac{\lambda + \sum_{i>k}\lambda_i}{n\lambda_{k+1}} \geq \frac{1 - \sigma_x^2\sqrt{t/n}}{L + \frac{\gamma}{1-\gamma} + \frac{\sqrt{t}\sigma_x^2 L}{\sqrt{n}(1-\gamma)}}.$$

Proof First of all, note that the sum of eigenvalues of A_k is equal to

$$\operatorname{tr}(A_k) = \lambda n + \sum_{i=1}^n \|\Sigma_{k:\infty}^{1/2} Z_{k:\infty}[i,*]^\top\|^2.$$

By Lemma 21 for some absolute constant c and any $t \in (0,n)$ with probability at least $1 - 2e^{-ct}$

$$(n - \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i \leq \sum_{i=1}^n \|\Sigma_{k:\infty}^{1/2} Z_{k:\infty}[i,*]^\top\|^2 \leq (n + \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i.$$

Now we know that with probability at least $1 - \delta - 2\exp(-c_2 t)$ the following two conditions hold:

$$\begin{aligned}
\mu_1(A_k) &\leq L\mu_n(A_k), \\
n\lambda + (n - \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i \leq \sum_{i=1}^n \mu_i(A_k) &\leq n\lambda + (n + \sqrt{nt}\sigma_x^2)\sum_{i>k}\lambda_i.
\end{aligned}$$

The first line of the display above implies that

$$n\mu_1(A_k)/L \le \sum_{i=1}^{n} \mu_i(A_k) \le n\mu_n(A_k) \cdot L$$

Thus, with probability at least $1 - \delta - 2\exp(-c_2 t)$,

$$\frac{\lambda}{L} + \frac{n - \sqrt{nt}\sigma_x^2}{nL} \sum_{i>k} \lambda_i \le \mu_n(A_k) \le \mu_1(A_k) \le \lambda L + \frac{(n + \sqrt{nt}\sigma_x^2)L}{n} \sum_{i>k} \lambda_i,$$

$$\frac{1}{L}\left(\lambda + \sum_i \lambda_i\right) - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}L} \sum_{i>k} \lambda_i \le \mu_n(A_k) \le \mu_1(A_k) \le L\left(\lambda + \sum_i \lambda_i\right) + \frac{\sqrt{t}\sigma_x^2 L}{\sqrt{n}} \sum_{i>k} \lambda_i.$$

Using the fact that $\sum_{i>k} \lambda_i \le \left(\lambda + \sum_{i>k} \lambda_i\right)/(1-\gamma)$, we obtain

$$\frac{1}{L}\left(\lambda + \sum_{i>k} \lambda_i\right)\left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right) \le \mu_n(A_k) \le \mu_1(A_k) \le L\left(\lambda + \sum_{i>k} \lambda_i\right)\left(1 + \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right),$$

which gives the first assertion of the lemma.

Next, note that $\mu_1(A_k) \ge \lambda_{k+1}\|Z_{k:\infty}[*,1]\|^2 + \lambda$. By Lemma 21 for some c_3 for any $t \in (0,n)$ w.p. at least $1 - 2e^{-c_3 t}$, $\|Z_{k:\infty}[*,1]\|^2 \ge n - \sqrt{nt}\sigma_x^2$, which means that if $1 - \delta - 2e^{-c_2 t} - 2e^{-c_3 t} > 0$ then with positive probability

$$\begin{aligned}
\lambda L + \frac{(n + \sqrt{nt}\sigma_x^2)L}{n} \sum_{i>k} \lambda_i &\ge \lambda_{k+1}(n - \sqrt{nt}\sigma_x^2) + \lambda, \\
\lambda(L-1) + \frac{(n + \sqrt{nt}\sigma_x^2)L}{n} \sum_{i>k} \lambda_i &\ge \lambda_{k+1}(n - \sqrt{nt}\sigma_x^2), \\
\left(\lambda + \sum_{i>k} \lambda_i\right)(L-1) + \left(1 + \frac{\sqrt{t}\sigma_x^2 L}{\sqrt{n}}\right) \sum_{i>k} \lambda_i &\ge \lambda_{k+1}(n - \sqrt{nt}\sigma_x^2), \\
\left(\lambda + \sum_{i>k} \lambda_i\right)\left(L + \frac{\gamma}{1-\gamma} + \frac{\sqrt{t}\sigma_x^2 L}{\sqrt{n}(1-\gamma)}\right) &\ge \lambda_{k+1}(n - \sqrt{nt}\sigma_x^2).
\end{aligned}$$

Taking $c_4 = \min(c_2, c_3)$ we see that if $\delta < 1 - 4e^{-c_4 t}$, then

$$\frac{\lambda + \sum_{i>k} \lambda_i}{n\lambda_{k+1}} \ge \frac{1 - \sigma_x^2\sqrt{t/n}}{L + \frac{\gamma}{1-\gamma} + \frac{\sqrt{t}\sigma_x^2 L}{\sqrt{n}(1-\gamma)}}.$$

■

Lemma 11 (k can be taken to be k^*) *Fix any constants $\gamma \in [0,1)$, $b > 0$, $L > 0$. Denote*

$$k^* = \min\{k : \rho_k > b\}.$$

There exist constants c, L' that only depend on σ_x, γ, b, L s.t. the following holds: suppose assumptions NoncritReg(k, γ) and CondNum(k, δ, L) hold for some $k \in [k^*, n]$. Then assumptions NoncritReg(k^*, γ) and CondNum$(k^*, \delta + ce^{-n/c}, L')$ hold too.

Proof First, by Lemma 25 for any $t \in (0, n)$ with probability at least $1 - \delta - 2e^{-c_1 t}$,

$$\frac{1}{L}\left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda + \sum_{i>k}\lambda_i\right) \le \mu_n(A_k) \le \mu_n(A_{k^*}).$$

Next, by Lemma 24 we know that with probability at least $1 - 6e^{-t/c_3}$,

$$\mu_1(A_{k^*}) \le c_3\sigma_x^2\left(\lambda_{k^*+1}(t+n) + \sum_{i>k^*}\lambda_i\right) + \lambda.$$

By definition of k^* and ρ_k

$$\lambda_{k^*+1}n = \rho_{k^*}^{-1}\left(\lambda + \sum_{i>k^*}\lambda_i\right) \le b^{-1}\left(\lambda + \sum_{i>k^*}\lambda_i\right).$$

Therefore,

$$\lambda + \sum_{i>k}\lambda_i = \lambda + \sum_{i>k^*}\lambda_i - \sum_{i=k^*+1}^{k}\lambda_i \ge \lambda + \sum_{i>k^*}\lambda_i - n\lambda_{k^*+1} \ge (1-b^{-1})\left(\lambda + \sum_{i>k^*}\lambda_i\right).$$

Moreover, since $\lambda > -\gamma\sum_{i>k^*}\lambda_i$,

$$\begin{aligned}
\lambda \le& \lambda + \sum_{i>k^*}\lambda_i,\\
\sum_{i>k^*}\lambda_i \le& \frac{1}{1-\gamma}\left(\lambda + \sum_{i>k^*}\lambda_i\right)\\
\lambda_{k^*+1}(t+n) \le& b^{-1}(1+t/n)\left(\lambda + \sum_{i>k^*}\lambda_i\right).
\end{aligned}$$

Thus, with probability at least $1 - \delta - 8e^{-t/c_4}$

$$\begin{aligned}
\mu_n(A_{k^*}) \ge& \frac{1}{L}\left(1 - \frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)(1-b^{-1})\left(\lambda + \sum_{i>k^*}\lambda_i\right),\\
\mu_1(A_{k^*}) \le& \left(c_3\sigma_x^2\left(\frac{1}{1-\gamma} + \frac{1}{b}\left(1 + \frac{t}{n}\right)\right) + 1\right)\left(\lambda + \sum_{i>k^*}\lambda_i\right).
\end{aligned}$$

Taking c_5 large enough (depending on L, b, σ_x and γ) and plugging in $t = n/c_5$ gives the result for $c = \max(8, c_4c_5)$ and

$$L' = \left(c_3\sigma_x^2\left(\frac{1}{1-\gamma} + \frac{1}{b}\left(1 + c_5^{-1}\right)\right) + 1\right) \div \left(\frac{1}{L}\left(1 - \frac{\sigma_x^2}{\sqrt{c_5}(1-\gamma)}\right)(1-b^{-1})\right).$$

The derivation of *NoncritReg*(k^*, γ) is obvious: indeed, assumption *NoncritReg*(k, γ) states that

$$\lambda > -\gamma \sum_{i>k} \lambda_i.$$

Since $k^* \geq k$, $\sum_{i>k} \lambda_i \leq \sum_{i>k^*} \lambda_i$, so

$$\lambda > -\gamma \sum_{i>k^*} \lambda_i,$$

which is exactly assumption *NoncritReg*(k^*, γ). ■

Appendix E. Lower bounds

A very convenient tool that we use to prove the lower bounds is the following

Lemma 26 (Lemma 9 from (Bartlett et al., 2020)) *Suppose that $\{\eta_i\}_{i=1}^p$ is a sequence of non-negative random variables, and that $\{t_i\}_{i=1}^p$ is a sequence of non-negative real numbers (at least one of which is strictly positive) such that, for some $\delta \in (0,1)$ for any $i \leq p$ with probability at least $1-\delta$, $\eta_i > t_i$. Then with probability at least $1-2\delta$,*

$$\sum_{i=1}^n \eta_i \geq \frac{1}{2} \sum_{i=1}^p t_i.$$

It turns out to be quite straightforward to express bias and variance terms as sums of non-negative series. This lemma allows us to give a separate high probability lower bound for each term in the series to obtain the high probability lower bound for the whole sum.

E.1 Variance term

The argument for lower bounding the variance term is the same as in (Bartlett et al., 2020). We repeat it here because the result in (Bartlett et al., 2020) is stated in a different form and in the ridgeless setting only.

Lemma 7 (Lower bound for the variance term) *Fix any constant $\gamma \in [0,1)$. There exists a constant c that only depends on σ_x and γ s.t. for any $k < n/c$ under assumptions NoncritReg(k,γ) and IndepCoord w.p. at least $1 - ce^{-n/c}$*

$$V \geq \frac{1}{cn} \sum_{i=1} \min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\}.$$

Proof The variance term can be written as

$$V = \operatorname{tr}\left(\Sigma X^\top A^{-2} X\right) = \sum_{i=1}^\infty \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2},$$

where z_i are columns of matrix Z (recall that $Z = X\Sigma^{-1/2}$). Note that every term in this sum is non-negative, even if A_{-i} is not PSD. Denote A_{-i+} to be the PSD square root of A_{-i}^2, i.e., the matrix

with the same eigendecomposition as A_{-i}, but with eigenvalues substituted by their absolute values. It immediately follows that

$$V \geq \sum_{i=1}^{\infty} \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i+}^{-1} z_i)^2},$$

By Cauchy-Schwartz we have

$$\|z_i\|^2 \cdot z_i^\top A_{-i}^{-2} z_i \geq (z_i^\top A_{-i+}^{-1} z_i)^2.$$

Thus,

$$V \geq \sum_{i=1}^{\infty} \frac{1}{\|z_i\|^2 \left(1+(\lambda_i z_i^\top A_{-i+}^{-1} z_i)^{-1}\right)^2}.$$

Now our goal is to lower-bound the largest eigenvalues of A_{-i+}^{-1}. Let's write

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

The idea is, as always, to separate the first k coordinates. Our initial goal is to bound the norm of $\sum_{j\neq i, j>k} \lambda_j z_j z_j^\top$. Using Lemma 24, for some absolute constant c_1 and for any $t > 0$, with probability at least $1 - 6e^{-t/c_1}$,

$$\left\| \sum_{j\neq i, j>k} \lambda_j z_j z_j^\top \right\| \leq \left\| \sum_{j>k} \lambda_j z_j z_j^\top \right\| = \|X_{k:\infty} X_{k:\infty}^\top\| \leq c_1 \sigma_x^2 \left(\lambda_{k+1}(t+n) + \sum_{i>k} \lambda_i \right)$$

The matrix $\sum_{j\neq i} \lambda_j z_j z_j^\top$ is a correction to $\sum_{j\neq i, j>k} \lambda_j z_j z_j^\top$ of rank at most k. Therefore, with probability at least $1-6e^{-t/c_1}$ the bottom k eigenvalues of $\sum_{j\neq i} \lambda_j z_j z_j^\top$ lie in the segment from 0 to $c_1\sigma_x^2\left(\lambda_{k+1}(t+n) + \sum_{i>k}\lambda_i\right)$. The matrix A_{-i} has the same eigenvalues, but with λ added to each one, so on the same event all the eigenvalues of A_{-i} are from λ to $\lambda + c_1\sigma_x^2\left(\lambda_{k+1}(t+n) + \sum_{i>k}\lambda_i\right)$. We can write

$$\begin{aligned}
&c_1\sigma_x^2 \left(\lambda_{k+1}(t+n) + \sum_{i>k} \lambda_i \right) + \lambda \\
\leq &c_1\sigma_x^2 \left(\lambda_{k+1}(t+n) + \frac{1}{1-\gamma}\left(\lambda + \sum_{i>k}\lambda_i \right) \right) + \frac{\gamma}{1-\gamma}\left(\lambda + \sum_{i>k}\lambda_i \right),
\end{aligned}$$

where we used that $\lambda > -\gamma \sum_{i>k}\lambda_i$ in the second line (for $\lambda < 0$ it implies $|\lambda| < \gamma \sum_{i>k}\lambda_i$). Moreover, for the left end of the segment we also have that either $\lambda > 0$ or

$$|\lambda| \leq \gamma \sum_{i>k} \lambda_i \leq \frac{\gamma}{1-\gamma}\left(\lambda + \sum_{i>k}\lambda_i \right).$$

Thus, for some constant c_2 which only depends on σ and γ, for any i with probability at least $1 - 6e^{-n/c_2}$, for any $j > k$

$$|\mu_j(A_i)| \leq c_2 \left(\lambda_{k+1} n + \lambda + \sum_{i>k} \lambda_i \right).$$

In words, with high probability the matrix A_{-i} has at least $n-k$ eigenvalues whose magnitude is bounded by $c_2\left(\lambda_{k+1}n+\lambda+\sum_{i>k}\lambda_i\right)$. Recall that A_{-i+} is PSD with the same magnitudes of the eienvalues. Denote $P_{i,k}$ to be the projector on the linear space spanned by the first k eigenvectors of A_{-i+}. We can now write that with probability at least $1-6e^{-n/c_2}$

$$z_i^\top A_{-i+}^{-1}z_i \geq \|(I-P_{i,k})z_i\|^2 c_2^{-1}\left(\lambda_{k+1}n+\lambda+\sum_{i>k}\lambda_i\right)^{-1}$$

Since z_i is independent of $P_{i,k}$, by Theorem 6.2.1 (Hanson-Wright inequality) in (Vershynin, 2018), for some absolute constant c_2 and for any $t>0$,

$$\mathbb{P}\left\{\left|\|P_{i,k}z_i\|^2-\mathbb{E}_{z_i}\|P_{i,k}z_i\|^2\right|\geq t\right\}\leq 2\exp\left(-c_2^{-1}\min\left\{\frac{t^2}{\sigma_x^4\|P_{i,k}^2\|_F^2},\frac{t}{\sigma_x^2\|P_{i,k}^2\|}\right\}\right).$$

Since $P_{i,k}$ is an orthogonal projector of rank k, $\|P_{i,k}^2\|_F^2=k$, $\|P_{i,k}^2\|=1$, and $\mathbb{E}_{z_i}\|P_{i,k}z_i\|^2=\mathrm{tr}(P_{i,k})=k$. Thus, w.p. at least $1-2e^{-t/c_2}$,

$$\left|\|P_{i,k}z_i\|^2-k\right|\leq\sigma_x^2\max(\sqrt{kt},t)\leq(t+\sqrt{kt})\sigma_x^2.$$

Next, by Lemma 21 for some constant c_3 and any $t\in(0,n)$ w.p. at least $1-2e^{-t/c_3}$,

$$n-\sqrt{nt}\sigma_x^2\leq\|z_i\|^2\leq n+\sqrt{nt}\sigma_x^2.$$

Take constant c_4 large enough depending on σ_x and set $t=n/c_4$. Then for any $k<n/c_5$, w.p. at least $1-10e^{-n/c_6}-\delta$,

$$z_i^\top A_{-i+}^{-1}z_i\geq\frac{n}{c_7\left(\lambda_{k+1}n+\lambda+\sum_{i>k}\lambda_i\right)},$$

where constants c_5 and c_6 depend only on σ_x and constant c_7 depends only on σ_x and γ.

Rewrite this equation as

$$(z_i^\top A_{-i+}^{-1}z_i)^{-1}\leq c_7\left(\lambda_{k+1}+\frac{1}{n}\left(\lambda+\sum_{i>k}\lambda_i\right)\right)=c_7\lambda_{k+1}(\rho_k+1),$$

where $\rho_k:=\frac{1}{n\lambda_{k+1}}\left(\lambda+\sum_{i>k}\lambda_i\right)$.

On the same event

$$\frac{1}{\|z_i\|^2\left(1+(\lambda_i z_i^\top A_{-i+}^{-1}z_i)^{-1}\right)^2}\geq\frac{1}{c_8 n\left(1+\frac{\lambda_{k+1}}{\lambda_i}(\rho_k+1)\right)^2},$$

where c_8 depends only on σ_x and γ.

Finally, by Lemma 26, we can convert lower bounds for separate non-negative terms into a lower bound on their sum: with probability at least $1-20e^{-n/c_6}$,

$$V\geq\frac{1}{8c_8 n}\sum_{i=1}^{p}\min\left\{1,\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\},$$

where we also used that $1/(a+b)^2\geq\min(a^{-2},b^{-2})/4$ for non-negative a,b. ■

E.2 Bias term

Lemma 8 (Lower bound for the bias term) *Fix any constant $L > 0$. There exists c that only depends on σ_x and L s.t. for any $k \in \{1, 2, \dots, p\}$ under assumptions PriorSigns($\bar{\theta}$) and StableLowerEig(k, δ, L) w.p. at least $1 - 2\delta - ce^{-n/c}$*

$$\mathbb{E}_{\theta^*} B \geq \frac{1}{c} \sum_i \frac{\lambda_i \bar{\theta}_i^2}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2},$$

where $\mathbb{E}_{\theta^*}$ denotes the expectation over the random draw of θ^* from the prior distribution described in assumption PriorSigns($\bar{\theta}$).

Proof

Applying Sherman-Morrison-Woodbury yields

$$\left(\lambda I_p + X^\top X\right)^{-1} = \lambda^{-1} I_p - \lambda^{-2} X^\top (I_n + \lambda^{-1} X X^\top)^{-1} X.$$

So,

$$\begin{aligned}\left(\lambda I_p + X^\top X\right)^{-1} X^\top X - I_p &= \left(\lambda I_p + X^\top X\right)^{-1} (\lambda I_p + X^\top X - \lambda I_p) - I_p \\ &= -\lambda \left(\lambda I_p + X^\top X\right)^{-1} \\ &= I_p - \lambda^{-1} X^\top (I_n + \lambda^{-1} X X^\top)^{-1} X \\ &= I_p - X^\top (\lambda I_n + X X^\top)^{-1} X.\end{aligned}$$

Thus, the bias term becomes

$$(\theta^*)^\top \left(I_p - X^\top (\lambda I_n + X X^\top)^{-1} X\right) \Sigma \left(I_p - X^\top (\lambda I_n + X X^\top)^{-1} X\right) \theta_*$$

and taking expectation over the prior kills all the off-diagonal elements, so

$$\mathbb{E}_{\theta^*} \mathcal{B} = \sum_i \left(\left(I_p - X^\top (\lambda I_n + X X^\top)^{-1} X\right) \Sigma \left(I_p - X^\top (\lambda I_n + X X^\top)^{-1} X\right)\right) [i, i] \cdot \bar{\theta}_i^2.$$

Let's compute the diagonal elements of the matrix

$$\left(I_p - X^\top (\lambda I_n + X X^\top)^{-1} X\right) \Sigma \left(I_p - X^\top (\lambda I_n + X X^\top)^{-1} X\right).$$

The i-th diagonal element is equal to the bias term for the case when $\theta^* = e_i$ — the i-th vector of the standard orthonormal basis. Note that the i-th row of $I_p - X^\top (\lambda I_n + X X^\top)^{-1} X$ is equal to $e_i - \sqrt{\lambda_i} z_i^\top (\lambda I_n + X X^\top)^{-1} X$, so the i-th diagonal element of the initial matrix is given by

$$\begin{aligned}&\sum_{j=1}^p \lambda_i \left(e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top (\lambda I_n + X X^\top)^{-1} z_j\right)^2 \\ &\lambda_i \left(1 - \lambda_i z_i^\top A^{-1} z_i\right)^2 + \sum_{j \neq i} \lambda_i \lambda_j^2 (z_i^\top A^{-1} z_j)^2.\end{aligned}$$

Recall that $A = \lambda I_n + \sum_{i=0}^{p} \lambda_i z_i z_i^\top$, $A_{-i} := A - \lambda_i z_i z_i^\top$.
First, let's use Sherman-Morrison identity to convert A in $z_i^\top A^{-1} z_i$ into A_{-i}:

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

So the diagonal element becomes

$$\frac{\lambda_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} + \sum_{j \neq i} \lambda_i \lambda_j^2 (z_i^\top A^{-1} z_j)^2 \geq \frac{\lambda_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2},$$

and thus

$$\mathbb{E}_{\theta^*} B \geq \sum_i \frac{\lambda_i \bar{\theta}_i^2}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2}.$$

Let's bound each term in that sum from below with high probability. By our assumptions, for any i with probability at least $1 - \delta$

$$\mu_n(A_{-i}) \geq \frac{1}{L} \left(\lambda + \sum_{j > k} \lambda_j \right).$$

Next,

$$\frac{\lambda_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \geq \frac{\lambda_i}{(1 + \lambda_i \mu_n(A_{-i})^{-1} \|z_i\|^2)^2},$$

and by Lemma 21 for some absolute constant c_1 for any $t \in (0, n)$ w.p. at least $1 - 2e^{-t/c_1}$ we have $\|z_i\|^2 \leq n - \sqrt{tn}\sigma_x^2 \leq n/2$, where the last transition is true if additionally $t \leq n/(4\sigma_x^4)$.

Recall that $\rho_k := \frac{\lambda + \sum_{j>k} \lambda_j}{n \lambda_{k+1}}$. We obtain by plugging $t = n/(4\sigma_x^4)$ that w.p. at least $1 - \delta - 2e^{-n/c_2}$,

$$\frac{\lambda_i \bar{\theta}_i^2}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \geq \frac{\lambda_i \bar{\theta}_i^2}{\left(1 + \frac{L \lambda_i}{2 \lambda_{k+1} \rho_k}\right)^2},$$

where c_2 only depends on σ_x.

Finally, since all the terms are non-negative and we need to obtain a lower bound on their sum, Lemma 26 gives the result.

■

Lemma 9 *For any $\gamma < 1$ there exists a constant c that only depends on γ and σ_x such that if assumptions CondNum(k, δ, L), NoncritReg(k, γ) and ExchCoord are satisfied for some $L \geq 1$ and $k \in \{1, 2, \ldots, p\}$, then StableLowerEig$(k, \delta + 2e^{-n/c}, cL)$ is also satisfied.*

Proof First of all, note that Assumption *NoncritReg*(k,γ) with $\gamma < 1$ directly implies that $\lambda + \sum_{i>k}\lambda_i \geq 0$, which is the second part of Assumption *StableLowerEig*(k,δ,L).

Next, by Lemma 25 for some absolute constant c_1 for any $t \in (0,n)$ with probability at least $1-\delta-2e^{-ct}$

$$\mu_n(A_k) \geq \frac{1}{L}\left(1-\frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda+\sum_{i>k}\lambda_i\right).$$

Taking $t = n/c_2$ where c_2 is large enough depending on γ, σ_x we get that for c large enough with probability at least $1-2e^{-n/c}$

$$\mu_n(A_k) \geq \frac{1}{cL}\left(\lambda+\sum_{i>k}\lambda_i\right).$$

Now we just need to propagate that result to A_{-i} for all i.

For $i \leq k$, we simply have $A_{-i} \succeq A_k$ with probability 1, so indeed $\forall i \leq k$

$$\mathbb{P}\left(\mu_n(A_{-i}) \geq \frac{1}{cL}\left(\lambda+\sum_{i>k}\lambda_i\right)\right) \geq \mathbb{P}\left(\mu_n(A_k) \geq \frac{1}{cL}\left(\lambda+\sum_{i>k}\lambda_i\right)\right) \geq 1-2e^{-n/c}.$$

When it comes to $i > k$, we can write

$$\begin{aligned}
A_{-i} =& \lambda I_n + \sum_{j\neq i}\lambda_j z_j z_j^\top \\
=& \lambda I_n + \sum_{j\leq k}\lambda_j z_j z_j^\top + \sum_{j>k, j\neq i}\lambda_j z_j z_j^\top \\
\succeq& \lambda I_n + \lambda_1 z_1 z_1^\top + \sum_{j>k, j\neq i}\lambda_j z_j z_j^\top \\
\succeq& \lambda I_n + \lambda_i z_1 z_1^\top + \sum_{j>k, j\neq i}\lambda_j z_j z_j^\top.
\end{aligned}$$

Now note that due to Assumption *ExchCoord*, the distribution of the matrix $\lambda I_n + \lambda_i z_1 z_1^\top + \sum_{j>k, j\neq i}\lambda_j z_j z_j^\top$ is the same as the distribution of $A_k = \lambda I_n + \sum_{j>k}\lambda_j z_j z_j^\top$. Therefore

$$\begin{aligned}
&\mathbb{P}\left(\mu_n(A_{-i}) \geq \frac{1}{cL}\left(\lambda+\sum_{i>k}\lambda_i\right)\right) \\
\geq& \mathbb{P}\left(\mu_n\left(\lambda I_n + \lambda_i z_1 z_1^\top + \sum_{j>k, j\neq i}\lambda_j z_j z_j^\top\right) \geq \frac{1}{cL}\left(\lambda+\sum_{i>k}\lambda_i\right)\right) \\
=& \mathbb{P}\left(\mu_n(A_k) \geq \frac{1}{cL}\left(\lambda+\sum_{i>k}\lambda_i\right)\right) \\
\geq& 1-2e^{-n/c},
\end{aligned}$$

which finishes the proof. ■

Appendix F. Deriving a useful identity

Motivated by the results of Bartlett et al. (2020), we split the principal directions of the covariance matrix into two parts: small dimensional and high dimensional. The main idea of our argument is to use classical machinery (like some sort of uniform convergence argument) in the small dimensional subspace. To do this we write $\hat{\theta}(y)^\top = \left[\hat{\theta}(y)_{0:k}^\top, \hat{\theta}(y)_{k:\infty}^\top\right]$ and mentally split the search process for $\hat{\theta}(y)$ into two parts: first, for any fixed $\theta_{0:k}$, optimize for $\theta_{k:\infty}$. Then only the first k coordinates are left. The result of that optimization in $\theta_{k:\infty}$ is the following identity:

$$\hat{\theta}(y)_{0:k} + X_{0:k}^\top A_k^{-1} X_{0:k} \hat{\theta}(y)_{0:k} = X_{0:k}^\top A_k^{-1} y. \tag{16}$$

The goal of this section is to derive this identity.

F.1 Derivation in the ridgeless case

In the ridgeless case we are simply dealing with projections, and $\hat{\theta}(y)$ is the minimum norm interpolating solution. Note that $\hat{\theta}(y)_{k:\infty}$ is also the minimum norm solution to the equation $X_{k:\infty}\theta_{k:\infty} = y - X_{0:k}\hat{\theta}(y)_{0:k}$, where $\theta_{k:\infty}$ is the variable. Thus, we can write

$$\hat{\theta}(y)_{k:\infty} = X_{k:\infty}^\top \left(X_{k:\infty} X_{k:\infty}^\top\right)^{-1} \left(y - X_{0:k}\hat{\theta}(y)_{0:k}\right).$$

Now we need to minimize the norm in $\hat{\theta}(y)_{0:k}$ (our choice of $\hat{\theta}(y)_{k:\infty}$ already makes the solution interpolating): we need to minimize the norm of the following vector:

$$v(\theta_{0:k}) = \left[\theta_{0:k}^\top, (y - X_{0:k}\theta_{0:k})^\top \left(X_{k:\infty} X_{k:\infty}^\top\right)^{-1} X_{k:\infty}\right]$$

As $\theta_{0:k}$ varies, this vector sweeps an affine subspace of our Hilbert space. The vector $\hat{\theta}(y)_{0:k}$ gives the minimum norm if and only if for any additional vector $\eta_{0:k}$ we have $v(\hat{\theta}(y)_{0:k}) \perp v(\hat{\theta}(y)_{0:k} + \eta_{0:k}) - v(\hat{\theta}(y)_{0:k})$. Let's write out the second vector: $\forall \eta_{0:k} \in \mathbb{R}^k$

$$v(\hat{\theta}(y)_{0:k} + \eta_{0:k}) - v(\hat{\theta}(y)_{0:k}) = \left[\eta_{0:k}^\top, -\eta_{0:k}^\top X_{0:k}^\top \left(X_{k:\infty} X_{k:\infty}^\top\right)^{-1} X_{k:\infty}\right]$$

We see that the above mentioned orthogonality for any $\eta_{0:k}$ is equivalent to the following:

$$\begin{aligned}\hat{\theta}(y)_{0:k}^\top - \left(y - X_{0:k}\hat{\theta}(y)_{0:k}\right)^\top \left(X_{k:\infty} X_{k:\infty}^\top\right)^{-1} X_{0:k} = 0,\\ \hat{\theta}(y)_{0:k} + X_{0:k}^\top A_k^{-1} X_{0:k} \hat{\theta}(y)_{0:k} = X_{0:k}^\top A_k^{-1} y,\end{aligned}$$

where we replaced $X_{k:\infty} X_{k:\infty}^\top =: A_k$.

F.2 Checking for the case of non-vanishing regularization

So, now we have $\lambda \neq 0$ and we want to prove that $\hat{\theta}(y)_{0:k} + X_{0:k}^\top A_k^{-1} X_{0:k} \hat{\theta}(y)_{0:k} = X_{0:k}^\top A_k^{-1} y$. Recall that

$$\begin{aligned}\hat{\theta}(y) =& X^\top (\lambda I_n + XX^\top)^{-1} y,\\ \hat{\theta}(y)_{0:k} =& X_{0:k}^\top (A_k + X_{0:k} X_{0:k}^\top)^{-1} y.\end{aligned}$$

This identity yields

$$\begin{aligned}&\hat{\theta}(y)_{0:k} + X_{0:k}^\top A_k^{-1} X_{0:k} \hat{\theta}(y)_{0:k} \\ =&X_{0:k}^\top (A_k + X_{0:k} X_{0:k}^\top)^{-1} y + X_{0:k}^\top A_k^{-1} X_{0:k} X_{0:k}^\top (A_k + X_{0:k} X_{0:k}^\top)^{-1} y \\ =&X_{0:k}^\top A_k^{-1} (A_k + X_{0:k} X_{0:k}^\top)(A_k + X_{0:k} X_{0:k}^\top)^{-1} y \\ =&X_{0:k}^\top A_k^{-1} y.\end{aligned}$$

Appendix G. Variance

Recall that the variance term is

$$V = \frac{1}{v_\varepsilon^2} \mathbb{E}_\varepsilon \|\hat{\theta}(\varepsilon)\|_\Sigma^2 = \frac{1}{v_\varepsilon^2} \mathbb{E}_\varepsilon \|X^\top (\lambda I_n + X X^\top)^{-1} \varepsilon\|_\Sigma^2.$$

In this section we prove the following lemma.

Lemma 27 *If for some $k < n$ the matrix A_k is PD, then*

$$V \le \frac{\mu_1(A_k^{-1})^2 \operatorname{tr}(X_{0:k} \Sigma_{0:k}^{-1} X_{0:k}^\top)}{\mu_n(A_k^{-1})^2 \mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right)^2} + \mu_1(A_k^{-1})^2 \operatorname{tr}(X_{k:\infty} \Sigma_{k:\infty} X_{k:\infty}^\top).$$

Note that the RHS of the inequality above is straightforward to estimate if one knows the spectrum of A_k. Indeed, the matrices $X_{0:k}\Sigma_{0:k}^{-1}X_{0:k}^\top$ and $X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^\top$ have i.i.d. elements on their diagonals, so their traces concentrate around expectations:

$$\operatorname{tr}(X_{0:k}\Sigma_{0:k}^{-1}X_{0:k}^\top) \sim kn \text{ and } \operatorname{tr}(X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^\top) \sim n\sum_{i>k}\lambda_i^2,$$

where we use $\sim$ informally to denote approximate equality with high probability.

When it comes to the matrix $\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}/n$, this is just a sample covariance matrix of n isotropic vectors in k-dimensional space. Since k is small compared to n, it concentrates around the identity. Thus,

$$\mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right) \sim n.$$

These computations are done rigorously in the proof of Theorem 5.

G.1 First k components

It was shown in Section F that the following identity holds (c.f. (16)):

$$X_{0:k}^\top A_k^{-1} \varepsilon = \hat{\theta}(\varepsilon)_{0:k} + X_{0:k}^\top A_k^{-1} X_{0:k} \hat{\theta}(\varepsilon)_{0:k}.$$

Multiplying the identity by $\hat{\theta}(\varepsilon)_{0:k}^\top$ from the left, and using that $\hat{\theta}(\varepsilon)_{0:k}^\top \hat{\theta}(\varepsilon)_{0:k} \ge 0$ we get

$$\hat{\theta}(\varepsilon)_{0:k}^\top X_{0:k}^\top A_k^{-1} \varepsilon \ge \hat{\theta}(\varepsilon)_{0:k}^\top X_{0:k}^\top A_k^{-1} X_{0:k} \hat{\theta}(\varepsilon)_{0:k}. \tag{17}$$

The leftmost expression is linear in $\hat{\theta}(\varepsilon)_{0:k}$, and the rightmost is quadratic. We use these expressions to bound $\|\hat{\theta}(\varepsilon)_{0:k}\|_{\Sigma_{0:k}}$.

First, we extract that norm from the quadratic part

$$\begin{aligned}\hat{\theta}(\varepsilon)_{0:k}^\top X_{0:k}^\top A_k^{-1} X_{0:k}\hat{\theta}(\varepsilon)_{0:k} \geq& \mu_n(A_k^{-1})\hat{\theta}(\varepsilon)_{0:k}^\top X_{0:k}^\top X_{0:k}\hat{\theta}(\varepsilon)_{0:k} \\ \geq& \mu_n(A_k^{-1})\|\hat{\theta}(\varepsilon)_{0:k}\|_{\Sigma_{0:k}}^2 \mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right).\end{aligned}$$

Then we can substitute (17) and apply Cauchy-Schwarz to obtain

$$\begin{aligned}\|\hat{\theta}(\varepsilon)_{0:k}\|_{\Sigma_{0:k}}^2 \mu_n(A_k^{-1})\mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right) &\leq \hat{\theta}(\varepsilon)_{0:k}^\top X_{0:k}^\top A_k^{-1} X_{0:k}\hat{\theta}(\varepsilon)_{0:k} \\ &\leq \hat{\theta}(\varepsilon)_{0:k}^\top X_{0:k}^\top A_k^{-1}\varepsilon \\ &\leq \|\hat{\theta}(\varepsilon)_{0:k}\|_{\Sigma_{0:k}}\left\|\Sigma_{0:k}^{-1/2} X_{0:k}^\top A_k^{-1}\varepsilon\right\|,\end{aligned}$$

and so

$$\|\hat{\theta}(\varepsilon)_{0:k}\|_{\Sigma_{0:k}}^2 \leq \frac{\varepsilon^\top A_k^{-1} X_{0:k}\Sigma_{0:k}^{-1} X_{0:k}^\top A_k^{-1}\varepsilon}{\mu_n(A_k^{-1})^2\mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2}.$$

Since ε is independent of X, taking expectation in ε only leaves the trace in the numerator:

$$\begin{aligned}\frac{1}{v_\varepsilon^2}\mathbb{E}_\varepsilon\|\hat{\theta}(\varepsilon)_{0:k}\|_{\Sigma_{0:k}}^2 \leq& \frac{\operatorname{tr}(A_k^{-1} X_{0:k}\Sigma_{0:k}^{-1} X_{0:k}^\top A_k^{-1})}{\mu_n(A_k^{-1})^2\mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2} \\ \leq& \frac{\mu_1(A_k^{-1})^2\operatorname{tr}(X_{0:k}\Sigma_{0:k}^{-1} X_{0:k}^\top)}{\mu_n(A_k^{-1})^2\mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2},\end{aligned}$$

where we transitioned to the second line by using the fact that $\operatorname{tr}(MM'M) \leq \mu_1(M)^2\operatorname{tr}(M')$ for PD matrices M, M'.

G.2 Components starting from $k+1$-st

The rest of the variance term is

$$\left\|\Sigma_{k:\infty}^{1/2} X_{k:\infty}^\top A^{-1}\varepsilon\right\|^2 = \varepsilon^\top A^{-1} X_{k:\infty}\Sigma_{k:\infty} X_{k:\infty}^\top A^{-1}\varepsilon.$$

Since ε is independent of X, taking expectation in ε only leaves the trace of the matrix:

$$\begin{aligned}\frac{1}{v_\varepsilon^2}\mathbb{E}_\varepsilon\left\|\Sigma_{k:\infty}^{1/2} X_{k:\infty}^\top A^{-1}\varepsilon\right\|^2 =& \operatorname{tr}(A^{-1} X_{k:\infty}\Sigma_{k:\infty} X_{k:\infty}^\top A^{-1}) \\ \leq& \mu_1(A^{-1})^2\operatorname{tr}(X_{k:\infty}\Sigma_{k:\infty} X_{k:\infty}^\top) \\ \leq& \mu_1(A_k^{-1})^2\operatorname{tr}(X_{k:\infty}\Sigma_{k:\infty} X_{k:\infty}^\top).\end{aligned}$$

Here we again used the fact that $\operatorname{tr}(MM'M) \leq \mu_1(M)^2\operatorname{tr}(M')$ for PD matrices M, M' to transition to the second line. We then used $A \succeq A_k$ to infer $\mu_1(A^{-1}) \leq \mu_1(A_k^{-1})$.

Appendix H. Bias

The bias term is given by $\|\theta^* - \hat{\theta}(X\theta^*)\|_\Sigma^2$. In this section we prove the following

Lemma 28 (Bias term) *Suppose that for some $k < n$ the matrix A_k is PD. Then there exists an absolute constant c such that*

$$\begin{aligned}
&\|\theta^* - \hat{\theta}(X\theta^*)\|_\Sigma^2/c \\
&\le \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}\frac{\mu_1\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)}{\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2}\|X_{k:\infty}\theta^*_{k:\infty}\|^2 \\
&\quad + \frac{\|\theta^*_{0:k}\|^2_{\Sigma_{0:k}^{-1}}}{\mu_n(A_k^{-1})^2\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2} \\
&\quad + \lambda_{k+1}\big(1+\max(0,-\lambda)\mu_1(A_k^{-1})\big)\mu_1(A^{-1})\|X_{k:\infty}\theta^*_{k:\infty}\|^2 \\
&\quad + \lambda_{k+1}\big(1+\max(0,-\lambda)\mu_1(A_k^{-1})\big)\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})}{\mu_k(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})^2}\|\Sigma_{0:k}^{-1/2}\theta^*_{0:k}\|^2.
\end{aligned}$$

H.1 First k components

We need to bound $\|\theta^*_{0:k} - \hat{\theta}(y)_{0:k}(\lambda, X\theta^*)\|^2_{\Sigma_{0:k}}$. By Section F, in particular identity (16), we have

$$\hat{\theta}(X\theta^*)_{0:k} + X_{0:k}^\top A_k^{-1}X_{0:k}\hat{\theta}(X\theta^*)_{0:k} = X_{0:k}^\top A_k^{-1}X\theta^*.$$

Denote the error vector as $\zeta := \hat{\theta}(X\theta^*) - \theta^*$. We can rewrite the equation above as

$$\zeta_{0:k} + X_{0:k}^\top A_k^{-1}X_{0:k}\zeta_{0:k} = X_{0:k}^\top A_k^{-1}X_{k:\infty}\theta^*_{k:\infty} - \theta^*_{0:k}.$$

Multiplying both sides by $\zeta_{0:k}^\top$ from the left and using that $\zeta_{0:k}^\top\zeta_{0:k} = \|\zeta_{0:k}\|^2 \ge 0$ we obtain

$$\zeta_{0:k}^\top X_{0:k}^\top A_k^{-1}X_{0:k}\zeta_{0:k} \le \zeta_{0:k}^\top X_{0:k}^\top A_k^{-1}X_{k:\infty}\theta^*_{k:\infty} - \zeta_{0:k}^\top\theta^*_{0:k}.$$

Next, divide and multiply by $\Sigma_{0:k}^{1/2}$ in several places:

$$\begin{aligned}
\zeta_{0:k}^\top\Sigma_{0:k}^{1/2}\Sigma_{0:k}^{-1/2}X_{0:k}^\top A_k^{-1}X_{0:k}\Sigma_{0:k}^{-1/2}\Sigma_{0:k}^{1/2}\zeta_{0:k} \le{}& \zeta_{0:k}^\top\Sigma_{0:k}^{1/2}\Sigma_{0:k}^{-1/2}X_{0:k}^\top A_k^{-1}X_{k:\infty}\theta^*_{k:\infty} \\
&- \zeta_{0:k}^\top\Sigma_{0:k}^{1/2}\Sigma_{0:k}^{-1/2}\theta^*_{0:k}.
\end{aligned}$$

Now we pull out the lowest singular values of the matrices in the LHS and largest singular values of the matrices in the RHS to obtain lower and upper bounds respectively, yielding

$$\begin{aligned}
&\|\zeta_{0:k}\|^2_{\Sigma_{0:k}}\mu_n(A_k^{-1})\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right) \\
&\le \|\zeta_{0:k}\|_{\Sigma_{0:k}}\mu_1(A_k^{-1})\sqrt{\mu_1\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)}\|X_{k:\infty}\theta^*_{k:\infty}\| \\
&+ \|\zeta_{0:k}\|_{\Sigma_{0:k}}\|\theta^*_{0:k}\|_{\Sigma_{0:k}^{-1}},
\end{aligned}$$

and so

$$\|\zeta_{0:k}\|_{\Sigma_{0:k}} \leq \frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})} \frac{\mu_1\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right)^{1/2}}{\mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right)} \|X_{k:\infty}\theta^*_{k:\infty}\| + \frac{\|\theta^*_{0:k}\|_{\Sigma_{0:k}^{-1}}}{\mu_n(A_k^{-1})\mu_k\left(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k} \Sigma_{0:k}^{-1/2}\right)}.$$

H.2 The rest of the components

Recall that the full bias term is $\|(I_p - X^\top(\lambda I_n + XX^\top)^{-1}X)\theta^*\|_\Sigma^2$ and that $A = \lambda I_n + XX^\top$. The contribution of the components of ζ, starting from the $k+1$st can be bounded as follows:

$$\begin{aligned}\|\theta^*_{k:\infty} - X_{k:\infty}^\top A^{-1} X\theta^*\|^2_{\Sigma_{k:\infty}} \\ \leq 3\left(\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|X_{k:\infty}^\top A^{-1} X_{k:\infty}\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|X_{k:\infty}^\top A^{-1} X_{0:k}\theta^*_{0:k}\|^2_{\Sigma_{k:\infty}}\right).\end{aligned}$$

First of all, let's deal with the second term:

$$\begin{aligned}\|X_{k:\infty}^\top A^{-1} X_{k:\infty}\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} =& \|\Sigma_{k:\infty}^{1/2} X_{k:\infty}^\top A^{-1} X_{k:\infty}\theta^*_{k:\infty}\|^2 \\ \leq& \|\Sigma_{k:\infty}\| \|X_{k:\infty}^\top A^{-1} X_{k:\infty}\theta^*_{k:\infty}\|^2 \\ =& \lambda_{k+1}(\theta^*_{k:\infty})^\top X_{k:\infty}^\top A^{-1} \underbrace{(A - \lambda I_n - X_{0:k}X_{0:k}^\top)}_{X_{k:\infty}X_{k:\infty}^\top} A^{-1} X_{k:\infty}\theta^*_{k:\infty} \\ \leq& \lambda_{k+1}(\theta^*_{k:\infty})^\top X_{k:\infty}^\top A^{-1}\big(A + \max(0,-\lambda) I_n\big) A^{-1} X_{k:\infty}\theta^*_{k:\infty} \\ \leq& \lambda_{k+1}\big(\mu_1(A^{-1}) + \max(0,-\lambda)\mu_1(A^{-1})^2\big)\|X_{k:\infty}\theta^*_{k:\infty}\|^2 \\ \leq& \lambda_{k+1}\big(1 + \max(0,-\lambda)\mu_1(A_k^{-1})\big)\mu_1(A_k^{-1})\|X_{k:\infty}\theta^*_{k:\infty}\|^2,\end{aligned}$$

where we used that $\mu_1(A_k^{-1}) \geq \mu_1(A^{-1})$ in the last transition.

Now, let's deal with the last term. Note that $A = A_k + X_{0:k}X_{0:k}^\top$. By the Sherman–Morrison–Woodbury formula,

$$\begin{aligned}A^{-1}X_{0:k} =& (A_k^{-1} + X_{0:k}X_{0:k}^\top)^{-1} X_{0:k} \\ =& \left(A_k^{-1} - A_k^{-1}X_{0:k}\left(I_k + X_{0:k}^\top A_k^{-1} X_{0:k}\right)^{-1} X_{0:k}^T A_k^{-1}\right) X_{0:k} \\ =& A_k^{-1}X_{0:k}\left(I_n - \left(I_k + X_{0:k}^\top A_k^{-1} X_{0:k}\right)^{-1} X_{0:k}^T A_k^{-1} X_{0:k}\right) \\ =& A_k^{-1}X_{0:k}\left(I_n - \left(I_k + X_{0:k}^\top A_k^{-1} X_{0:k}\right)^{-1}\left(I_k + X_{0:k}^T A_k^{-1} X_{0:k} - I_k\right)\right) \\ =& A_k^{-1}X_{0:k}\left(I_k + X_{0:k}^\top A_k^{-1} X_{0:k}\right)^{-1}.\end{aligned}$$

Thus,

$$
\begin{aligned}
&\|X_{k:\infty}^\top A^{-1} X_{0:k}\theta^*_{0:k}\|^2_{\Sigma_{k:\infty}}\\
=&\|X_{k:\infty}^\top A_k^{-1} X_{0:k}\left(I_k + X_{0:k}^\top A_k^{-1} X_{0:k}\right)^{-1}\theta^*_{0:k}\|^2_{\Sigma_{k:\infty}}\\
=&\|\Sigma^{1/2}_{k:\infty} X_{k:\infty}^\top A_k^{-1} X_{0:k}\Sigma_{0:k}^{-1/2}\left(\Sigma_{0:k}^{-1} + \Sigma_{0:k}^{-1/2} X_{0:k}^\top A_k^{-1} X_{0:k}\Sigma_{0:k}^{-1/2}\right)^{-1}\Sigma_{0:k}^{-1/2}\theta^*_{0:k}\|^2\\
\le&\|A_k^{-1/2} X_{k:\infty}\Sigma_{k:\infty} X_{k:\infty}^\top A_k^{-1/2}\|\mu_1(A_k^{-1/2})^2\frac{\mu_1(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})}{\mu_k(\Sigma_{0:k}^{-1/2} X_{0:k}^\top A_k^{-1} X_{0:k}\Sigma_{0:k}^{-1/2})^2}\|\Sigma_{0:k}^{-1/2}\theta^*_{0:k}\|^2\\
\le&\|\Sigma_{k:\infty}\|\|A_k^{-1/2} X_{k:\infty} X_{k:\infty}^\top A_k^{-1/2}\|\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})}{\mu_k(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})^2}\|\Sigma_{0:k}^{-1/2}\theta^*_{0:k}\|^2\\
=&\lambda_1\|I_n - \lambda A_k^{-1}\|\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})}{\mu_k(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})^2}\|\Sigma_{0:k}^{-1/2}\theta^*_{0:k}\|^2\\
\le&\lambda_1\big(1+\max(0,-\lambda)\mu_1(A_k^{-1})\big)\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})}{\mu_k(\Sigma_{0:k}^{-1/2} X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})^2}\|\Sigma_{0:k}^{-1/2}\theta^*_{0:k}\|^2,
\end{aligned}
$$

where in the last transition we used the fact that $I_n - \lambda A_k^{-1}$ is a PSD matrix with norm bounded by 1 for $\lambda > 0$.

Putting those bounds together yields the result.

Appendix I. Main results

I.1 Upper bound on the prediction MSE

Theorem 5 *There exists a (large) constant c, which only depends on σ_x, s.t. for any $k < n/c$ with probability at least $1 - ce^{-n/c}$, if the matrix A_k is PD, then*

$$
\begin{aligned}
B/c \le&\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}\left(1 + \frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2} + n\lambda_{k+1}\mu_1(A_k^{-1})\left(1+\max(0,-\lambda)\mu_1(A_k^{-1})\right)\right)\\
&+\|\theta^*_{0:k}\|^2_{\Sigma_{0:k}^{-1}}\left(\frac{1}{n^2\mu_n(A_k^{-1})^2} + \frac{\lambda_{k+1}}{n}\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\left(1+\max(0,-\lambda)\mu_1(A_k^{-1})\right)\right),\\
V/c \le&\frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}\frac{k}{n} + n\mu_1(A_k^{-1})^2\sum_{i>k}\lambda_i^2.
\end{aligned}
$$

Proof Lemmas 27 and 28 bound the bias and variance on the event that A_k is PD. Next to those lemmas we already put explanations of why those bounds are easy to assess via concentration arguments. Here we just do this rigorously.

Recall the bounds from Lemmas 27 and 28: for some absolute constant c

$$B/c \leq \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} \tag{18}$$

$$+\frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}\frac{\mu_1\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)}{\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2}\|X_{k:\infty}\theta^*_{k:\infty}\|^2 \tag{19}$$

$$+\frac{\|\theta^*_{0:k}\|^2_{\Sigma_{0:k}^{-1}}}{\mu_n(A_k^{-1})^2\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2} \tag{20}$$

$$+\lambda_{k+1}\big(1+\max(0,-\lambda)\mu_1(A_k^{-1})\big)\mu_1(A^{-1})\|X_{k:\infty}\theta^*_{k:\infty}\|^2 \tag{21}$$

$$+\lambda_{k+1}\big(1+\max(0,-\lambda)\mu_1(A_k^{-1})\big)\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})}{\mu_k(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})^2}\|\Sigma_{0:k}^{-1/2}\theta^*_{0:k}\|^2, \tag{22}$$

$$V/c \leq \frac{\mu_1(A_k^{-1})^2\,\mathrm{tr}(X_{0:k}\Sigma_{0:k}^{-1}X_{0:k}^\top)}{\mu_n(A_k^{-1})^2\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2} \tag{23}$$

$$+\mu_1(A_k^{-1})^2\,\mathrm{tr}(X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^\top), \tag{24}$$

where the first four terms correspond to the bias and the last two to the variance. By inspecting that expression one can notice that it consists of some products of simple quantities that could be assessed individually. Namely, those quantities are:

1. $\mu_1(A_k^{-1})$ and $\mu_n(A_k^{-1})$ — smallest and largest singular values of A_k. In this theorem we assume that those quantities are known or there is some oracle control over them.

2. $\mu_1\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)$ and $\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)$.

   The matrix $X_{0:k}\Sigma_{0:k}^{-1/2} \in \mathbb{R}^{k\times n}$ has n i.i.d. columns with isotropic sub-Gaussian distribution in $\mathbb{R}^k$. The matrix $\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}/n$ is the sample covariance matrix of those columns, so when $k \ll n$ it concentrates around its expectation, which is I_k. More precisely, by Theorem 5.39 in (Vershynin, 2012), for some constants c'_x, C'_x (which only depend on σ_x) for every $t > 0$ s.t. $\sqrt{n} - C'_x\sqrt{k} - \sqrt{t} > 0$, with probability $1 - 2\exp(-c'_x t)$,

$$\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right) \geq \left(\sqrt{n} - C'_x\sqrt{k} - \sqrt{t}\right)^2, \tag{25}$$

$$\mu_1\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right) \leq \left(\sqrt{n} + C'_x\sqrt{k} + \sqrt{t}\right)^2. \tag{26}$$

3. $\mathrm{tr}\left(X_{0:k}\Sigma_{0:k}^{-1}X_{0:k}^\top\right)$ and $\mathrm{tr}\left(X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^\top\right)$.

   $\mathrm{tr}\left(X_{0:k}\Sigma_{0:k}^{-1}X_{0:k}^\top\right)$ is the sum of squared norms of columns of $\Sigma_{0:k}^{-1/2}X_{0:k}^\top$, which are n i.i.d. isotropic vectors in $\mathbb{R}^k$. Analogously, $\mathrm{tr}\left(X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^\top\right)$ is the sum of squared norms of n i.i.d. sub-Gaussian vectors with covariance $\Sigma^2_{k:\infty}$. Therefore, they concentrate around their expectations by the law of large numbers. More precisely, by Lemma 21 with probability at

least $1 - 4e^{-c_2 t}$,

$$\begin{aligned}\operatorname{tr}\left(X_{0:k}\Sigma_{0:k}^{-1}X_{0:k}^\top\right) &\le (n + \sqrt{tn}\sigma_x^2)k,\\ \operatorname{tr}\left(X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^\top\right) &\le (n + \sqrt{tn}\sigma_x^2)\sum_{i>k}\lambda_i^2.\end{aligned}$$

4. $\|X_{k:\infty}\theta^*_{k:\infty}\|^2$.

   Once again, this quantity concentrates by the law of large numbers. The vector $X_{k:\infty}\theta^*_{k:\infty}/\|\theta^*_{k:\infty}\|_{\Sigma_{k:\infty}}$ has n i.i.d. centered components with unit variances and sub-Gaussian norms at most σ_x. Treating those components as sub-Gaussian vectors in $\mathbb{R}^1$, we can apply Lemma 21 to get that for any $t \in (0, n)$, with probability at least $1 - 2e^{-c_2 t}$,

$$\|X_{k:\infty}\theta^*_{k:\infty}\|^2 \le (n + \sqrt{tn}\sigma_x^2)\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}.$$

Now take constant c_4 to be large enough depending on σ_x and set $t = n/c_4$. For some constant c_5 which only depends on σ_x we get that with probability at least $1 - c_5 e^{-n/c_5}$, all the following inequalities hold at the same time:

$$\begin{aligned}\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right) &\ge n/c_5,\\ \mu_1\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right) &\le c_5 n,\\ \|X_{k:\infty}\theta^*_{k:\infty}\|^2 &\le c_5 n\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}},\\ \|X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^\top\| &\le c_5\left(\lambda_{k+1}^2 n + \sum_{i>k}\lambda_i^2\right),\\ \operatorname{tr}\left(X_{0:k}\Sigma_{0:k}^{-1}X_{0:k}^\top\right) &\le c_5 nk,\\ \operatorname{tr}\left(X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^\top\right) &\le c_5 n\sum_{i>k}\lambda_i^2.\end{aligned}$$

Next, plug these bounds into (19)–(24):

$$\frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}\frac{\mu_1\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)}{\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2}\|X_{k:\infty}\theta^*_{k:\infty}\|^2 \le c_5^3\frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}},$$

$$\frac{\|\theta^*_{0:k}\|^2_{\Sigma_{0:k}^{-1}}}{\mu_n(A_k^{-1})^2\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2} \le c_5^2\frac{\|\theta^*_{0:k}\|^2_{\Sigma_{0:k}^{-1}}}{\mu_n(A_k^{-1})^2 n^2},$$

$$\begin{aligned}\lambda_{k+1}\big(1 + \max(0, -\lambda)\mu_1(A_k^{-1})\big)\mu_1(A^{-1})\|X_{k:\infty}\theta^*_{k:\infty}\|^2 \le\\ \le c_5^2\lambda_{k+1}\left(1 + \max(0, -\lambda)\mu_1(A_k^{-1})\right)\mu_1(A_k^{-1})n\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}},\end{aligned}$$

$$\lambda_{k+1}\big(1+\max(0,-\lambda)\mu_1(A_k^{-1})\big)\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{\mu_1(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})}{\mu_k(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2})^2}\|\Sigma_{0:k}^{-1/2}\theta^*_{0:k}\|^2 \leq$$
$$\leq c_5^4\lambda_{k+1}\left(1+\max(0,-\lambda)\mu_1(A_k^{-1})\right)\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2}\frac{1}{n}\|\theta^*_{0:k}\|^2_{\Sigma_{0:k}^{-1}},$$

$$\frac{\mu_1(A_k^{-1})^2\operatorname{tr}(X_{0:k}\Sigma_{0:k}^{-1}X_{0:k}^\top)}{\mu_n(A_k^{-1})^2\mu_k\left(\Sigma_{0:k}^{-1/2}X_{0:k}^\top X_{0:k}\Sigma_{0:k}^{-1/2}\right)^2}\leq c_5^3\frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2}\frac{k}{n},$$

$$\mu_1(A_k^{-1})^2\operatorname{tr}(X_{k:\infty}\Sigma_{k:\infty}X_{k:\infty}^\top)\leq c_5\mu_1(A_k^{-1})^2 n\sum_{i>k}\lambda_i^2.$$

Putting all the terms together gives the result. ■

Corollary 6 *Fix any constants $\gamma\in[0,1)$ and $L>0$. There exists a constant c that only depends on σ_x, γ, L s.t. for any $k<n/c$ and $\delta<1-ce^{-n/c}$ under assumptions NoncritReg(k,γ) and CondNum(k,δ,L), it holds that $\rho_k>c^{-1}$, and with probability at least $1-\delta-ce^{-n/c}$,*

$$B/c\leq\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}}+\|\theta^*_{0:k}\|^2_{\Sigma_{0:k}^{-1}}\left(\frac{\lambda+\sum_{i>k}\lambda_i}{n}\right)^2,$$
$$V/c\leq\frac{k}{n}+\frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda+\sum_{i>k}\lambda_i\right)^2}.$$

Proof Almost all the work was already done in Lemma 25. It says that for some absolute constant c_1 and for any $t\in(0,n)$ with probability at least $1-\delta-2e^{-c_1 t}$,

$$\frac{1}{L}\left(1-\frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda+\sum_i\lambda_i\right)\leq\mu_n(A_k)\leq\mu_1(A_k)\leq L\left(1-\frac{\sqrt{t}\sigma_x^2}{\sqrt{n}(1-\gamma)}\right)\left(\lambda+\sum_i\lambda_i\right).$$

Moreover, if $\delta<1-4e^{-c_1 t}$, then

$$\rho_k\geq\frac{1-\sigma^2\sqrt{t/n}}{L+\frac{\gamma}{1-\gamma}+\frac{\sqrt{t}\sigma^2 L}{\sqrt{n}(1-\gamma)}}.$$

We just need to choose t, plug these bounds into the result of Theorem 5 and evaluate the result up to multiplicative constants.

First, choose constant c_2 large enough depending on L, γ, σ_x , and put $t=n/c_2$. Statements above imply that if $\delta<1-4e^{-n/(c_1c_2)}$, then for some constant c_3 which only depends on L, γ, σ_x,

with probability at least $1 - \delta - c_2 e^{-n/(c_1 c_2)}$,

$$\begin{aligned}\mu_n(A_k^{-1}) =& \mu_1(A_k)^{-1} \geq \frac{1}{c_3}\left(\lambda + \sum_i \lambda_i\right)^{-1},\\ \mu_1(A_k^{-1}) =& \mu_n(A_k)^{-1} \leq \frac{1}{c_3}\left(\lambda + \sum_i \lambda_i\right)^{-1},\\ \rho_k \geq& \frac{1}{c_3}.\end{aligned}$$

These three inequalities allow us to evaluate the result of Theorem 5: let's plug them term-by-term:

- Since $\lambda > -\gamma \sum_{i>k} \lambda_i$,

  $$\max(0, -\lambda) \leq \frac{\gamma}{1-\gamma}\left(\lambda + \sum_i \lambda_i\right).$$

  Thus,

  $$1 + \max(0, -\lambda)\mu_1(A_k^{-1}) \leq 1 + \frac{\gamma}{1-\gamma} c_3,$$

  so this term is just a constant.

- $$n\lambda_{k+1}\mu_1(A_k^{-1}) \leq c_3 n\lambda_{k+1}\left(\lambda + \sum_i \lambda_i\right) = c_3/\rho_k \leq c_3^2,$$

  so this term is also just a constant.

- $$\frac{1}{n^2 \mu_n(A_k^{-1})^2} \leq \frac{c_3^2}{n}\left(\lambda + \sum_i \lambda_i\right)^2.$$

- $$\begin{aligned}\frac{\lambda_{k+1}}{n}\frac{\mu_1(A_k^{-1})}{\mu_n(A_k^{-1})^2} \leq& \frac{c_3^3}{n^2}\cdot n\lambda_{k+1}\left(\lambda + \sum_i \lambda_i\right)\\ =& \frac{c_3^3}{n^2}\cdot \rho_k^{-1}\left(\lambda + \sum_i \lambda_i\right)^2\\ \leq& \frac{c_3^4}{n^2}\left(\lambda + \sum_i \lambda_i\right)^2.\end{aligned}$$

- $\frac{\mu_1(A_k^{-1})^2}{\mu_n(A_k^{-1})^2} \leq L^2$ — also just a constant.

- $$n\mu_1(A_k^{-1})^2 \le c_3^2 n \left(\lambda + \sum_i \lambda_i\right)^{-2}.$$

Plugging all these bounds in the statement of Theorem 5 gives the result for a large enough c. ■

I.2 Upper bound matches the lower bound

In the next theorem we show that the upper bound given in Theorem 5 matches the lower bounds from Lemmas 7 and 8 if we choose suitable k. Note that by Lemmas 25 and 11, being able to control the condition number of $A_{k'}$ for some $k' < n$ implies that we can choose a suitable k. (Note that there are choices of θ^* and Σ for which the lower bound $\overline{B}$ is larger than the upper bound of Lemma 5.4 in (Negrea et al., 2020); this seems to be because the proof of Lemma B.1 in that paper applies Lemma B.2 to a nonsymmetric matrix. This error was removed in the newer version of the same paper, which uses the results of Bartlett et al. (2020) instead.)

Theorem 10 (The lower bound is the same as the upper bound) *Denote*

$$\begin{aligned}
\underline{B} &:= \sum_i \frac{\lambda_i |\theta_i^*|^2}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2}, \\
\overline{B} &:= \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2, \\
\underline{V} &:= \frac{1}{n}\sum_i \min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\}, \\
\overline{V} &:= \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2}.
\end{aligned}$$

Fix constants $a > 0$ and $b > 1/n$. There exists a constant $c > 0$ that only depends on a, b, s.t. the following holds: if either $\rho_k \in (a, b)$ or $k = \min\{\kappa : \rho_\kappa > b\}$, then

$$c^{-1} \le \underline{B} \,/\, \overline{B} \le 1, \quad c^{-1} \le \underline{V} \,/\, \overline{V} \le 1.$$

Proof First of all, we represent

$$\begin{aligned}
\|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2 &= \sum_i \left(\mathbb{1}\{i \le k\}\frac{|\theta_i^*|^2\rho_k^2\lambda_{k+1}^2}{\lambda_i} + \mathbb{1}\{i > k\}\lambda_i|\theta_i^*|^2\right) \\
\frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\left(\lambda + \sum_{i>k}\lambda_i\right)^2} &= \sum_i \left(\mathbb{1}\{i \le k\}\frac{1}{n} + \mathbb{1}\{i > k\}\frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2}\right)
\end{aligned}$$

In the following we will bound the ratio of the sums from the statement of the theorem by bounding the ratios of the corresponding terms.

- First case: $\rho_k \in (a, b)$.

– Bias term:
  * $i \le k$:
$$\begin{aligned}&\frac{\lambda_i|\theta_i^*|^2}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2} : \frac{|\theta_i^*|^2\rho_k^2\lambda_{k+1}^2}{\lambda_i}\\ =&\frac{\lambda_i^2}{\rho_k^2\lambda_{k+1}^2\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2}\\ =&\left(1+\frac{\lambda_{k+1}\rho_k}{\lambda_i}\right)^{-2}\\ \in&\left((1+b)^{-2},1\right)\end{aligned}$$
  * $i > k$:
$$\begin{aligned}&\frac{\lambda_i|\theta_i^*|^2}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2} : \lambda_i|\theta_i^*|^2\\ =&\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^{-2}\\ \in&\left((1+a^{-1})^{-2},1\right)\end{aligned}$$
– Variance term:
  * $i \le k$:
$$\begin{aligned}&\frac{1}{n}\min\left\{1,\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\} : \frac{1}{n}\\ \in&\left((1+b)^{-2},1\right]\end{aligned}$$
  * $i > k$:
$$\begin{aligned}&\frac{1}{n}\min\left\{1,\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\} : \frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2}\\ =&\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2} : \frac{\lambda_i^2}{\lambda_{k+1}^2\rho_k^2}\\ =&\frac{\rho_k^2}{(\rho_k+1)^2}\\ \in&\left((1+a^{-1})^{-2},1\right)\end{aligned}$$

• Second case: $k = \min\{l : \rho_l > b\}$. In this case we have
$$\begin{aligned}&\rho_k \ge b,\\ &\frac{\lambda_k + n\lambda_{k+1}\rho_k}{n\lambda_k} = \frac{\lambda + \lambda_k + \sum_{i>k}\lambda_i}{n\lambda_k} = \rho_{k-1} < b,\\ \forall i \le k:\quad &\lambda_i \ge \lambda_k \ge \frac{n\lambda_{k+1}\rho_k}{nb-1} = \frac{\lambda_{k+1}\rho_k}{b} \ge \frac{\lambda_{k+1}\rho_k}{b}.\end{aligned}$$
The rest of the computation is analogous to the previous case:

– Bias term:

  * $i \le k$:

$$\begin{aligned}
&\frac{\lambda_i |\theta_i^*|^2}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2} : \frac{|\theta_i^*|^2 \rho_k^2 \lambda_{k+1}^2}{\lambda_i} \\
=&\frac{\lambda_i^2}{\rho_k^2 \lambda_{k+1}^2 \left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2} \\
=&\left(1+\frac{\lambda_{k+1}\rho_k}{\lambda_i}\right)^{-2} \\
\in&\left[(1+b)^{-2}, 1\right)
\end{aligned}$$

  * $i > k$:

$$\begin{aligned}
&\frac{\lambda_i |\theta_i^*|^2}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^2} : \lambda_i |\theta_i^*|^2 \\
=&\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)^{-2} \\
\in&\left[(1+b^{-1})^{-2}, 1\right)
\end{aligned}$$

– Variance term:

  * $i \le k$:

$$\begin{aligned}
&\frac{1}{n}\min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\} : \frac{1}{n} \\
\in&\left[\frac{\lambda_{k+1}^2\rho_k^2/b^2}{\lambda_{k+1}^2(\rho_k+1)^2}, 1\right] \\
\subseteq&\left[\frac{b^2}{(b+1)^2 b^2}, 1\right] \\
=&\left[(b+1)^{-2}, 1\right]
\end{aligned}$$

  * $i > k$:

$$\begin{aligned}
&\frac{1}{n}\min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right\} : \frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2} \\
=&\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2} : \frac{\lambda_i^2}{\lambda_{k+1}^2\rho_k^2} \\
=&\frac{\rho_k^2}{(\rho_k+1)^2} \\
\in&\left[(1+b^{-1})^{-2}, 1\right]
\end{aligned}$$

■

I.3 Alternative form of the main bound

Lemma 12 *Suppose $k < n/c$ for some $c > 1$ and $k^* < k$. Then*

$$\lambda_{k+1}\rho_k \le \lambda_{k^*+1}\rho_{k*} \le \lambda_{k+1}\rho_k/(1 - b^{-1}c^{-1}).$$

Proof

$$\begin{aligned}
\lambda_{k^*+1}\rho_{k*} =& \lambda_{k+1}\rho_k + \frac{1}{n}\sum_{i=k^*+1}^{k}\lambda_i \\
\le& \lambda_{k+1}\rho_k + \frac{k-k^*}{n}\lambda_{k^*+1} \\
=& \lambda_{k+1}\rho_k + \frac{k-k^*}{n}\frac{\lambda_{k^*+1}\rho_{k*}}{\rho_{k*}} \\
\le& \lambda_{k+1}\rho_k + \frac{\lambda_{k^*+1}\rho_{k*}}{bc},
\end{aligned}$$

where we used $k - k^* < n/c$ and $\rho_{k^*} > b$ in the last transition. Moving $\frac{\lambda_{k^*+1}\rho_{k*}}{bc}$ to the left-hand side and dividing both sides by $(1 - b^{-1}c^{-1})$ gives the result. ■

Corollary 13 *There is a large positive constant c that only depends on σ_x such that if*

$$\lambda > cn\lambda_{\lfloor n/c\rfloor} + 2\sum_{i>\lfloor n/c\rfloor}\lambda_i,$$

then

$$\begin{aligned}
B/c \le& \sum_i \lambda_i|\theta_i^*|^2\frac{(\lambda/n)^2}{(\lambda/n+\lambda_i)^2}, \\
V/c \le& \frac{1}{n}\sum_i \frac{\lambda_i^2}{(\lambda/n+\lambda_i)^2}.
\end{aligned}$$

Proof Set $\gamma = 0$ and denote c_1 to be the constant c from Lemma 3. Take $L = 2c_1$ and $b = L^2$, $a = b/2$. For such choice of γ, L, a, b denote c_2 to be the constant from Theorem 1 and take any $\tilde{k} < n/c_2$.

Take any λ s.t.

$$\lambda \ge 2\sum_{i>\tilde{k}}\lambda_i \quad \text{and} \quad \rho_{\tilde{k}} \ge L^2,$$

i.e.,

$$\lambda \ge \max\left(2\sum_{i>\tilde{k}}\lambda_i,\ L^2 n\lambda_{\tilde{k}+1} - \sum_{i>\tilde{k}}\lambda_i\right).$$

Then the conditions of the first part of Lemma 3 are satisfied with $\delta = 0$, which means that with probability $1 - c_1 e^{-n/c_1}$, $\mu_n(A_{\tilde{k}}) \ge L^{-1}\mu_1(A_{\tilde{k}})$, so the assumptions of the first part of Theorem 1

are satisfied with $\delta = c_1 e^{-n/c_1}$ and $\bar{k} = \tilde{k}$. Note also that since $\rho_{\tilde{k}} \geq L^2 = b$, then $k^* \leq \tilde{k}$. This means that with probability at least $1 - c_1 e^{-n/c_1} - c_2 e^{-n/c_2}$, for $k = k^*$,

$$\begin{aligned} B/c_2 \leq & \|\theta^*_{k:\infty}\|^2_{\Sigma_{k:\infty}} + \|\theta^*_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{\lambda + \sum_{i>k} \lambda_i}{n} \right)^2, \\ V/c_2 \leq & \frac{k}{n} + \frac{n \sum_{i>k} \lambda_i^2}{\left(\lambda + \sum_{i>k} \lambda_i\right)^2}. \end{aligned}$$

Now since $k = k^*$, by Theorem 10 there exists a large constant c_3 (that depends on b and c_2) such that on the same event,

$$\begin{aligned} B/c_3 \leq & \sum_i \lambda_i |\theta^*_i|^2 \frac{\rho_k^2 \lambda^2_{k^*+1}}{\left(\rho_{k^*} \lambda_{k^*+1} + \lambda_i\right)^2}, \\ V/c_3 \leq & \frac{1}{n} \sum_i \frac{\lambda_i^2}{\left(\rho_{k^*} \lambda_{k^*+1} + \lambda_i\right)^2}, \end{aligned}$$

where $\tilde{B}$ and $\tilde{V}$ are defined in Equations (6)–(7).

We've just cast the bounds to the alternative form, which allows us to transition from k^* to the initial value $\tilde{k}$. By Lemma 12 since $n/c_2 \geq \tilde{k} \geq k^*$ there exists a constant c_4 that depends on c_2, c_3, b such that on the same event

$$\begin{aligned} B/c_4 \leq & \sum_i \lambda_i |\theta^*_i|^2 \frac{\rho_k^2 \lambda^2_{k+1}}{\left(\rho_k \lambda_{k+1} + \lambda_i\right)^2}, \\ V/c_4 \leq & \frac{1}{n} \sum_i \frac{\lambda_i^2}{\left(\rho_k \lambda_{k+1} + \lambda_i\right)^2}. \end{aligned}$$

Finally, since $\lambda > 2 \sum_{i>k} \lambda_i$, we have

$$\lambda/n \leq \rho_k \lambda_{k+1} = \frac{1}{n} \left(\lambda + \sum_{i>k} \lambda_i \right) \leq 1.5\lambda/n.$$

Thus, on the same event

$$\begin{aligned} B/(2.25c_4) \leq & \sum_i \lambda_i |\theta^*_i|^2 \frac{(\lambda/n)^2}{\left(\lambda/n + \lambda_i\right)^2}, \\ V/(2.25c_4) \leq & \frac{1}{n} \sum_i \frac{\lambda_i^2}{\left(\lambda/n + \lambda_i\right)^2}. \end{aligned}$$

To finish the proof take $c = \max(2.25c_4, c_1 + c_2, L^2)$ and $\tilde{k} = \lfloor n/c \rfloor$.

■

Lemma 14 *Suppose that $n \geq c^2 + c$ for some $c > 0$ and take*

$$\lambda = cn\lambda_{\lfloor n/c \rfloor} + 2 \sum_{i > \lfloor n/c \rfloor} \lambda_i.$$

Then

$$d(\lambda/n) \geq \frac{n}{2\max(2, (c+1)^2)}.$$

Proof

$$d(\lambda/n) = \sum_i \frac{\lambda_i}{\lambda_i + c\lambda_{\lfloor n/c \rfloor} + \frac{2}{n}\sum_{i > \lfloor n/c \rfloor} \lambda_i}.$$

Consider two cases:

Case 1: $(1+c)\lambda_{\lfloor n/c \rfloor} \geq \frac{2}{n}\sum_{i > \lfloor n/c \rfloor} \lambda_i$. Then

$$\begin{aligned}
&\sum_i \frac{\lambda_i}{\lambda_i + c\lambda_{\lfloor n/c \rfloor} + \frac{2}{n}\sum_{i > \lfloor n/c \rfloor} \lambda_i} \\
\geq &\sum_i \frac{\lambda_i}{\lambda_i + (1+2c)\lambda_{\lfloor n/c \rfloor}} \\
\geq &\sum_{i \leq \lfloor n/c \rfloor} \frac{\lambda_i}{\lambda_i + (1+2c)\lambda_{\lfloor n/c \rfloor}} \\
\geq &\sum_{i \leq \lfloor n/c \rfloor} \frac{\lambda_i}{\lambda_i(2+2c)} \\
= &\frac{\lfloor n/c \rfloor}{2+2c}.
\end{aligned}$$

Case 2: $(1+c)\lambda_{\lfloor n/c \rfloor} < \frac{2}{n}\sum_{i > \lfloor n/c \rfloor} \lambda_i$. Then

$$\begin{aligned}
&\sum_i \frac{\lambda_i}{\lambda_i + c\lambda_{\lfloor n/c \rfloor} + \frac{2}{n}\sum_{i > \lfloor n/c \rfloor} \lambda_i} \\
\geq &\sum_{i > \lfloor n/c \rfloor} \frac{\lambda_i}{\lambda_i + c\lambda_{\lfloor n/c \rfloor} + \frac{2}{n}\sum_{i > \lfloor n/c \rfloor} \lambda_i} \\
\geq &\sum_{i > \lfloor n/c \rfloor} \frac{\lambda_i}{(1+c)\lambda_{\lfloor n/c \rfloor} + \frac{2}{n}\sum_{i > \lfloor n/c \rfloor} \lambda_i} \\
\geq &\sum_{i > \lfloor n/c \rfloor} \frac{\lambda_i}{\frac{4}{n}\sum_{i > \lfloor n/c \rfloor} \lambda_i} \\
= &\frac{n}{4}.
\end{aligned}$$

A straightforward computation shows that if $n \geq c^2 + c$ then $n/c - 1 \geq n/(c+1)$, so

$$\frac{\lfloor n/c \rfloor}{2+2c} \geq \frac{n}{2(c+1)^2},$$

which finishes the proof.

■

Appendix J. Negative regularization

Lemma 17 (Lower bound on the bias for any non-negative regularization) *There exist constants b, c that only depend on σ_x such that the following holds: suppose that assumptions IndepCoord and PriorSigns($\bar{\theta}$) hold. Take $k = \min\{\kappa : \rho_\kappa(0) > b\}$ and suppose that $k > 0$. Then with probability at least $1 - ce^{-n/c}$ for any $\lambda \geq 0$*

$$\mathbb{E}_{\theta^*} B \geq \frac{1}{c}\|\bar{\theta}_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \frac{\left(\sum_{i>k}\lambda_i\right)^2}{n^2}.$$

Proof We start exactly as in the proof of Lemma 8, where it was shown that if A_{-i} is PSD for every i (which is satisfied almost surely when $\lambda \geq 0$) then

$$\mathbb{E}_{\theta^*} B \geq \sum_i \frac{\lambda_i \bar{\theta}_i^2}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \geq \sum_i \frac{\lambda_i \bar{\theta}_i^2}{(1+\lambda_i \mu_n(A_{-i}^{-1})\|z_i\|^2)^2}. \tag{27}$$

Note that have $\mu_n(A_{-i}^{-1})$ is a decreasing function of λ with probability 1. Thus, the right-hand side of (27) is a non-decreasing function of λ with probability 1, and any lower bound for it when $\lambda = 0$ will also hold uniformly for all $\lambda \geq 0$. Thus, for the remainder of the proof, fix $\lambda = 0$.

We are going to use Lemma 16 to lower bound $\mu_n(A_{-i})$ for each i separately (we are *not* looking for a uniform bound over all i simultaneously). If $i \leq k$, then $A_{-i} \succeq A_k$ with probability 1, so we can just use Lemma 16 directly. If $i > k$, consider the following matrix:

$$X^{(i)}_{k:\infty} := [\sqrt{\lambda_{k+1}} z_{k+1}, \ldots, \sqrt{\lambda_{i-1}} z_{i-1}, \sqrt{\lambda_i} z_1, \sqrt{\lambda_{i+1}} z_{i+1}, \ldots, \sqrt{\lambda_p} z_p].$$

In words, we took matrix X, multiplied the first column by $\sqrt{\lambda_i/\lambda_1}$ (to make the variances equal to λ_i), swapped the first column with the i-th column and dropped the first k columns. The purpose of this matrix is to write the following:

$$A_{-i} = \sum_{j\neq i} \lambda_j z_j z_j^\top \succeq \lambda_i z_1 z_1^\top + \sum_{j>k, j\neq i} \lambda_j z_j z_j^\top = X^{(i)}_{k:\infty}(X^{(i)}_{k:\infty})^\top.$$

Thus, to lower bound $\mu_n(A_{-i})$ one can just lower bound $\mu_n(X^{(i)}_{k:\infty}(X^{(i)}_{k:\infty})^\top)$. This can be done by using Lemma 16 with matrix $X^{(i)}_{k:\infty}$ instead of $X_{k:\infty}$, which is valid because matrix $X^{(i)}_{k:\infty}$ satisfies exactly the same assumptions, namely the matrix $X^{(i)}_{k:\infty}\Sigma^{-1/2}_{k:\infty}$ has independent centered σ_x-sub-Gaussian elements with unit variances.

Therefore, by Lemma 16 for some constant c_1 that only depends on σ_x for any i with probability at least $1 - c_1 e^{-n/c_1}$,

$$\begin{aligned}
\mu_n(A_{-i}) \geq& \sum_{i>k}\lambda_i - c_1\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right)\\
\geq& \left(1 - c_1\rho_k(0)^{-1} - c_1\rho_k(0)^{-1/2}\right)\sum_{i>k}\lambda_i\\
=& n\lambda_{k+1}(\rho_k - c_1 - c_1\sqrt{\rho_k}),
\end{aligned}$$

where we used Equations (13) and (14). Choose a constant b large enough depending on c_1, so that $\rho_k - c_1 - c_1\sqrt{\rho_k} \geq \rho_k/c_2$ for some constant c_2 that only depends on σ_x.

By Lemma 21, for some absolute constant c_3 for any $t \in (0, n)$, w.p. at least $1 - 2e^{-t/c_3}$, we have $\|z_i\|^2 \leq n - \sqrt{tn}\sigma_x^2 \leq n/2$, provided $t \leq n/(4\sigma_x^4)$. Combining it with the previous results and taking constant c_4 large enough depending on σ_x and c_2 we get that if $\rho_k > c_4$ then for any i with probability at least $1 - c_4 e^{-n/c_4}$,

$$\frac{\lambda_i \bar{\theta}_i^2}{(1 + \lambda_i \mu_n(A_{-i}^{-1})\|z_i\|^2)^2} \geq \frac{1}{c_4} \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{n\lambda_i}{n\lambda_{k+1}\rho_k})^2} = \frac{1}{c_4} \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k})^2}.$$

Now we convert the high-probability lower bound for each term into the high-probability lower bound for the whole sum. Using Lemma 26 gives that with probability at least $1 - 2c_4 e^{-n/c_4}$,

$$\mathbb{E}_{\theta^*} B \geq \frac{1}{2c_4} \sum_i \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k})^2}.$$

Finally, by Theorem 10 there exists a constant c_5 that only depends on b s.t.

$$\sum_i \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k})^2} \geq \frac{1}{c_5} \|\theta_{0:k}\|_{\Sigma_{0:k}^{-1}}^2 \left(\frac{\sum_{i>k} \lambda_i}{n} \right)^2.$$

Therefore, setting the constant c large enough (depending on b and σ_x) gives the result. ■

Lemma 18 (Upper bound on excess risk for some negative regularization) *There exists a constant c that only depends on σ_x such that the following holds: suppose that assumptions PriorSigns($\bar{\theta}$) and IndepCoord hold and that $\rho_k(0) > c$ for some $k < n/c$. Assume also that*

$$v_\varepsilon^2 \leq \frac{1}{c} \|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}^2 \frac{\left(\sum_{i>k} \lambda_i\right)^2}{n^3 \left(\sum_{i>k} \lambda_i^2\right)^2}. \tag{15}$$

Then there exists such $\lambda < 0$ that with probability at least $1 - ce^{-n/c}$

$$\mathbb{E}_{\theta^*} B + v_\varepsilon^2 V \leq c \left(v_\varepsilon^2 \frac{k}{n} + v_\varepsilon \|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}} \sqrt{\frac{\sum_{i>k} \lambda_i^2}{n}} + \|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}^2 \frac{\lambda_{k+1} \sum_{i>k} \lambda_i}{n} + \|\bar{\theta}_{k:\infty}\|_{\Sigma_{k:\infty}}^2 \right).$$

Proof In the following $c_1, c_2, \ldots$ are constants that only depend on σ_x.

Let's introduce a new variable $\Diamond$ such that $\lambda = -\sum_{i>k} \lambda_i + \Diamond$.

By Lemma 16 with probability at least $1 - c_1 e^{-n/c_1}$,

$$\mu_1(A_k) = \lambda + \mu_1(X_{k:\infty} X_{k:\infty}^\top) \leq \Diamond + c_1 \left(n\lambda_{k+1} + \sqrt{n \sum_{i>k} \lambda_i^2} \right),$$

$$\mu_n(A_k) = \lambda + \mu_n(X_{k:\infty} X_{k:\infty}^\top) \geq \Diamond - c_1 \left(n\lambda_{k+1} + \sqrt{n \sum_{i>k} \lambda_i^2} \right).$$

Let's put

$$\sum_{i>k} \lambda_i > \lozenge > 2c_1 \left(n\lambda_{k+1} + \sqrt{n \sum_{i>k} \lambda_i^2} \right). \tag{28}$$

Note that the range for $\lozenge$ is non-empty if ρ_k is large enough according to Equations (13) and (14). On the same event we get

$$\mu_n(A_k^{-1})^{-1} = \mu_1(A_k) \leq \frac{3}{2}\lozenge, \qquad \mu_n(A_k^{-1}) \geq \frac{2}{3}\lozenge^{-1},$$
$$\mu_1(A_k^{-1})^{-1} = \mu_n(A_k) \geq \frac{1}{2}\lozenge, \qquad \mu_1(A_k^{-1}) \leq 2\lozenge^{-1}.$$

Now we are in a position to use Theorem 5. Recall that $0 < \lozenge < \sum_{i>k} \lambda_i$. Thus

$$\max(0, -\lambda) = -\lambda = \sum_{i>k} \lambda_i - \lozenge \leq \sum_{i>k} \lambda_i.$$

Note that results of Theorem 5 still apply for the case when the expectation of the bias term is taken over the prior from assumption *PriorSigns*($\bar{\theta}$). Indeed, as explained in the sketch of its proof, it decomposes very clearly into an algebraic and a stochastic part, where concentration results are applied. One can see that the only stochastic quantity that changes when the expectation over θ^* is taken is $\|X_{0:k}\theta^*_{0:k}\|^2$. To obtain the result of the theorem one needs to show that $\mathbb{E}^*_\theta \|X_{k:\infty}\theta^*_{k:\infty}\|^2 \leq \tilde{c}\|\bar{\theta}_{k:\infty}\|^2_{\Sigma_{k:\infty}}$ with probability $1 - \tilde{c}e^{-n/\tilde{c}}$ for some $\tilde{c}$ that only depends on σ_x. This is indeed the case because expectations over θ^* of the squared components of $X_{k:\infty}\theta^*_{k:\infty}$ are i.i.d. sub-exponential random variables with expectation $\|\bar{\theta}_{k:\infty}\|^2_{\Sigma_{k:\infty}}$ and sub-exponential norm bounded by $\bar{c}\|\bar{\theta}_{k:\infty}\|^2_{\Sigma_{k:\infty}}$ for a constant $\bar{c}$ that only depends on σ_x. Thus, the desired concentration result holds by the same application of Bernstein's inequality as in Lemma 21.

Thus, we can plug our bounds on eigenvalues into Theorem 5 to get that if $k < n/c_2$ then with probability at least $1 - c_1 e^{-n/c_1} - c_2 e^{-n/c_2}$,

$$\begin{aligned}
\mathbb{E}_\theta B/c_2 \leq & \|\bar{\theta}_{k:\infty}\|^2_{\Sigma_{k:\infty}} \left(1 + \frac{(2\lozenge^{-1})^2}{\left(\frac{2}{3}\lozenge^{-1}\right)^2} + n\lambda_{k+1}(2\lozenge^{-1}) \left(1 + (2\lozenge^{-1}) \sum_{i>k} \lambda_i \right) \right) \\
& + \|\bar{\theta}_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{1}{n^2 \left(\frac{2}{3}\lozenge^{-1}\right)^2} + \frac{\lambda_{k+1}}{n} \frac{(2\lozenge^{-1})}{\left(\frac{2}{3}\lozenge^{-1}\right)^2} \left(1 + (2\lozenge^{-1}) \sum_{i>k} \lambda_i \right) \right), \\
V/c_2 \leq & \frac{(2\lozenge^{-1})^2}{\left(\frac{2}{3}\lozenge^{-1}\right)^2} \frac{k}{n} + n(2\lozenge^{-1})^2 \sum_{i>k} \lambda_i^2.
\end{aligned}$$

Recall that $\lozenge < \sum_{i>k} \lambda_i$, so $1 + (2\lozenge^{-1}) \sum_{i>k} \lambda_i$ is the same as $\lozenge^{-1} \sum_{i>k} \lambda_i$ up to a constant multiplier. That is, on the same event,

$$B/c_3 \leq \|\bar{\theta}_{k:\infty}\|^2_{\Sigma_{k:\infty}} \left(1 + \frac{n\lambda_{k+1} \sum_{i>k} \lambda_i}{\lozenge^2} \right) \tag{29}$$

$$+ \|\bar{\theta}_{0:k}\|^2_{\Sigma^{-1}_{0:k}} \left(\frac{\lozenge^2}{n^2} + \frac{\lambda_{k+1} \sum_{i>k} \lambda_i}{n} \right), \tag{30}$$

$$V/c_3 \leq \frac{k}{n} + \frac{n \sum_{i>k} \lambda_i^2}{\lozenge^2}. \tag{31}$$

One can see that $\Diamond$ balances the bias in the first k components against two things: the bias in the tail and the variance. The value of $\Diamond$ that is optimal to balance the bias in the first k components and the bias in the tail is $\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i}$. As we will check further, up to a constant factor, $\Diamond$ will be in the range that we set in Equation (28). There are two cases then: the first case is when this choice of $\Diamond$ is optimal because the variance is not larger than the bias. The second case is when $\Diamond$ needs to be chosen larger than $\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i}$ to decrease the variance. So, consider two cases:

1. If the noise is small, meaning that

$$v_\varepsilon^2 \frac{n\sum_{i>k}\lambda_i^2}{n\lambda_{k+1}\sum_{i>k}\lambda_i} \leq \|\bar{\theta}_{k:\infty}\|_{\Sigma_{k:\infty}}^2 + \|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}^2 \frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n},$$

then set

$$\Diamond = a\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i}$$

for a constant a that only depends on σ_x that we will choose next. This a must be such that Equation (28) is satisfied, which means

$$\begin{aligned} a\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} &\leq \sum_{i>k}\lambda_i, \\ a\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} &\geq 2c_1\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right). \end{aligned}$$

Using $\sqrt{n\sum_{i>k}\lambda_i^2} \leq \sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i}$ we obtain that it is enough for a to satisfy

$$\begin{aligned} a &\leq \rho_k(0)^{1/2}, \\ a &\geq 2c_1\left(\rho_k(0)^{-1/2} + 1\right). \end{aligned}$$

One can see that $a = 4c_1$ satisfies this condition when $c > \max(1, 16c_1^2)$ since $\rho_k(0) > c$. Taking such an a, plugging $\Diamond$ into Equations (29)–(31), and choosing c_4 big enough depending on a, c_1, c_2, c_3, we get that with probability at least $1 - c_4 e^{-n/c_4}$,

$$\begin{aligned} B + v_\varepsilon^2 V &\leq c_4\left(\frac{k}{n}v_\varepsilon^2 + v_\varepsilon^2\frac{n\sum_{i>k}\lambda_i^2}{n\lambda_{k+1}\sum_{i>k}\lambda_i} + \|\bar{\theta}_{k:\infty}\|_{\Sigma_{k:\infty}}^2 + \|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}^2\frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n}\right) \\ &\leq 2c_4\left(\frac{k}{n}v_\varepsilon^2 + \|\bar{\theta}_{k:\infty}\|_{\Sigma_{k:\infty}}^2 + \|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}^2\frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n}\right), \end{aligned}$$

which implies the desired bound for any $c > 2c_4$.

2. If the noise is large, meaning that

$$v_\varepsilon^2 \frac{n\sum_{i>k}\lambda_i^2}{n\lambda_{k+1}\sum_{i>k}\lambda_i} > \|\bar{\theta}_{k:\infty}\|_{\Sigma_{k:\infty}}^2 + \|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}^2 \frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n}, \tag{32}$$

then set

$$\Diamond = a\sqrt{\frac{v_\varepsilon}{\|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}} n\sqrt{n\sum_{i>k}\lambda_i^2}}.$$

for a constant a that only depends on σ_x that we choose next. As in the previous case, a must be such that Equation (28) is satisfied, which means

$$\begin{aligned}
a\sqrt{\frac{v_\varepsilon}{\|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}} n\sqrt{n\sum_{i>k}\lambda_i^2}} \leq& \sum_{i>k}\lambda_i, \\
a\sqrt{\frac{v_\varepsilon}{\|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}} n\sqrt{n\sum_{i>k}\lambda_i^2}} \geq& 2c_1\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).
\end{aligned}$$

The first condition is satisfied whenever $a < \sqrt{c}$ due to Equation (15). Now consider the second condition. Because of Equation (32), we have

$$v_\varepsilon\sqrt{n\sum_{i>k}\lambda_i^2} \geq \|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}\lambda_{k+1}\sum_{i>k}\lambda_i, \tag{33}$$

$$\frac{\Diamond}{a} = \sqrt{\frac{v_\varepsilon}{\|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}} n\sqrt{n\sum_{i>k}\lambda_i^2}} \geq \sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i}. \tag{34}$$

Thus, it is enough to satisfy

$$a\sqrt{n\lambda_{k+1}\sum_{i>k}\lambda_i} \geq 2c_1\left(n\lambda_{k+1} + \sqrt{n\sum_{i>k}\lambda_i^2}\right).$$

This is exactly the same condition as in the previous case, so it can be reduced to

$$a \geq 2c_1(\rho_k(0)^{-1/2} + 1).$$

Thus, just as in the small variance case, we see that since $c > \max(1, 16c_1^2)$ then $a = 4c_1$ satisfies both conditions.

Take such an a. Before plugging $\Diamond$ into Equations (29)–(31), note the following. Because of Equation (34), we have

$$\begin{aligned}
\frac{n\lambda_{k+1}\sum_{i>k}\lambda_i}{\Diamond^2} \leq& \frac{1}{a^2}, \\
\frac{\Diamond^2}{n^2} \geq& a^2\frac{\lambda_{k+1}\sum_{i>k}\lambda_i}{n},
\end{aligned}$$

which means that if we take c_5 large enough depending on a and c_3, then Equations (29)–(31) imply

$$\begin{aligned}
B/c_5 \leq& \|\bar{\theta}_{k:\infty}\|_{\Sigma_{k:\infty}}^2 + \|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}}^2\frac{\Diamond^2}{n^2}, \\
V/c_5 \leq& \frac{k}{n} + \frac{n\sum_{i>k}\lambda_i^2}{\Diamond^2}.
\end{aligned}$$

Now plugging in the expression for $\Diamond$ gives that with probability at least $1 - c_1 e^{-n/c_1}$,

$$B + v_\varepsilon^2 V \leq c_5 \left(\frac{k}{n} v_\varepsilon^2 + (a^{-2} + a^2) v_\varepsilon \|\bar{\theta}_{0:k}\|_{\Sigma_{0:k}^{-1}} \sqrt{\frac{\sum_{i>k} \lambda_i^2}{n}} + \|\bar{\theta}_{k:\infty}\|_{\Sigma_{k:\infty}}^2 \right),$$

which implies the result for $c > \max((a^{-2} + a^2)c_5, c_1)$.

■

References

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/content/early/2020/04/22/1907378117.

P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *ArXiv*, abs/2103.09177, 2021.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, 2019a. ISSN 0027-8424. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/content/116/32/15849.

M. Belkin, D. Hsu, and J. Xu. Two models of double descent for weak features. *ArXiv*, abs/1903.07571, 2019b.

K. Bibas, Y. Fogel, and M. Feder. A new look at an old problem: A universal learning approach to linear regression. pages 2304–2308, 07 2019. doi: 10.1109/ISIT.2019.8849398.

F. Bunea, S. Strimas-Mackey, and M. H. Wegkamp. Interpolating predictors in high-dimensional factor regression. *J. Mach. Learn. Res.*, 23:10–1, 2022.

M. Celentano, T. Misiakiewicz, and A. Montanari. Minimum complexity interpolation in random features models, 2021.

C. Cheng and A. Montanari. Dimension free ridge regression, 2022.

G. Chinot and M. Lerasle. On the robustness of the minimum ℓ_2 interpolator. *ArXiv*, abs/2003.05838, 2021.

M. Dereziński, F. Liang, and M. Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. 12 2019.

M. Dereziński, F. Liang, Z. Liao, and M. W. Mahoney. Precise expressions for random projections: Low-rank approximation and randomized newton. *ArXiv*, abs/2006.10653, 2020.

E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *arXiv: Statistics Theory*, pages 247–279, 2015.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized two-layers neural networks in high dimension. *ArXiv*, abs/1904.12191, 2020a.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. When do neural networks outperform kernel methods? *ArXiv*, abs/2006.13409, 2020b.

N. Ghosh, S. Mei, and B. Yu. The three stages of learning dynamics in high-dimensional kernel methods, 2021.

O. Guédon, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann. On the interval of fluctuation of the singular values of random matrices. *Journal of the European Mathematical Society*, 19(5): 1469–1505, 2017. doi: 10.4171/jems/697.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *ArXiv*, abs/1903.08560v5, 2020.

T. J. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *ArXiv*, abs/1903.08560v4, 2019.

D. Hsu, S. Kakade, T. Zhang, S. Mannor, N. Srebro, and B. Williamson. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14, 06 2014. doi: 10.1007/s10208-014-9192-1.

D. Kobak, J. Lomond, and B. Sanchez. Optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *arXiv: Statistics Theory*, 2020.

F. Koehler, L. Zhou, D. J. Sutherland, and N. Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=FyOhThdDBM`.

T. Liang and A. Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *ArXiv*, abs/1808.00387, 2018.

T. Liang, A. Rakhlin, and X. Zhai. On the risk of minimum-norm interpolants and restricted lower isometry of kernels. *ArXiv*, abs/1908.10292, 2019.

T. Liang, A. Rakhlin, and X. Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. *ArXiv*, abs/1908.10292, 2020.

A. D. McRae, S. Karnik, M. A. Davenport, and V. Muthukumar. Harmless interpolation in regression and classification with structured features, 2021.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *ArXiv*, abs/1908.05355, 2019.

S. Mei, T. Misiakiewicz, and A. Montanari. Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration. *ArXiv*, abs/2101.10588, 2021a.

S. Mei, T. Misiakiewicz, and A. Montanari. Learning with invariances in random features and kernel models, 2021b.

T. Misiakiewicz and S. Mei. Learning with convolution and pooling operations in kernel methods, 2021.

A. Montanari and Y. Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *ArXiv*, abs/2007.12826, 2020.

V. Muthukumar, K. Vodrahalli, and A. Sahai. Harmless interpolation of noisy data in regression. *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2299–2303, 2019.

V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021. URL http://jmlr.org/papers/v22/20-603.html.

P. Nakkiran. More data can hurt for linear regression: Sample-wise double descent. *ArXiv*, abs/1912.07242, 2019.

A. Narang, V. Muthukumar, and A. Sahai. Classification and adversarial examples in an overparameterized linear model: A signal processing perspective, 2021.

J. Negrea, G. K. Dziugaite, and D. Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7263–7272. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/negrea20a.html.

D. Richards, J. Mourtada, and L. Rosasco. Asymptotics of ridge (less) regression under general source condition. *ArXiv*, abs/2006.06386, 2020.

O. Shamir. The implicit bias of benign overfitting, 2022.

K. Tikhomirov. Sample Covariance Matrices of Heavy-Tailed Distributions. *International Mathematics Research Notices*, 2018(20):6254–6289, 04 2017. ISSN 1073-7928. doi: 10.1093/imrn/rnx067. URL https://doi.org/10.1093/imrn/rnx067.

A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *ArXiv*, abs/2009.14286, 2020.

R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

D. Wu and J. Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *ArXiv*, abs/2006.05800, 2020.

J. Xu and D. J. Hsu. On the number of variables to use in principal component regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and

R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5094–5103. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8753-on-the-number-of-variables-to-use-in-principal-component-regression.pdf.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016. URL http://arxiv.org/abs/1611.03530.

L. Zhou, D. J. Sutherland, and N. Srebro. On uniform convergence and low-norm interpolation learning. *ArXiv*, abs/2006.05942, 2021.