Dimension free ridge regression

Chen Cheng* Andrea Montanari†

October 16, 2022

Abstract

Random matrix theory has become a widely useful tool in high-dimensional statistics and theoretical machine learning. However, random matrix theory is largely focused on the proportional asymptotics in which the number of columns grows proportionally to the number of rows of the data matrix. This is not always the most natural setting in statistics where columns correspond to covariates and rows to samples.

With the objective to move beyond the proportional asymptotics, we revisit ridge regression (ℓ_2-penalized least squares) on i.i.d. data $(\boldsymbol{x}_i, y_i)$, $i \le n$, where $\boldsymbol{x}_i$ is a feature vector and $y_i = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \varepsilon_i \in \mathbb{R}$ is a response. We allow the feature vector to be high-dimensional, or even infinite-dimensional, in which case it belongs to a separable Hilbert space, and assume either $\boldsymbol{z}_i := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$ to have i.i.d. entries, or to satisfy a certain convex concentration property.

Within this setting, we establish non-asymptotic bounds that approximate the bias and variance of ridge regression in terms of the bias and variance of an 'equivalent' sequence model (a regression model with diagonal design matrix). The approximation is up to multiplicative factors bounded by $(1 \pm \Delta)$ for some explicitly small Δ.

Previously, such an approximation result was known only in the proportional regime and only up to additive errors: in particular, it did not allow to characterize the behavior of the excess risk when this converges to 0. Our general theory recovers earlier results in the proportional regime (with better error rates). As a new application, we obtain a completely explicit and sharp characterization of ridge regression for Hilbert covariates with regularly varying spectrum. Finally, we analyze the overparametrized near-interpolation setting and obtain sharp 'benign overfitting' guarantees.

Contents

*Department of Statistics, Stanford University

†Department of Electrical Engineering and Department of Statistics, Stanford University; School of Mathematics, Institute for Advanced Studies, Princeton

1 Introduction

In regression modeling, we typically assume to be given data $(\boldsymbol{x}_i, y_i)$, $i \le n$ that are i.i.d. samples from a common distribution $\mathbb{P}$, with $\boldsymbol{x}_i$ a feature vector, and $y_i \in \mathbb{R}$ a scalar response. We would like to estimate a model $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$ to predict y_{new} from $\boldsymbol{x}_{\mathsf{new}}$, where $(\boldsymbol{x}_{\mathsf{new}}, y_{\mathsf{new}}) \sim \mathbb{P}$ is a new sample from the same distribution. In this paper, we will focus on linear models whereby $f(\boldsymbol{x}) = \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{x} \rangle$, and use ridge regression for the estimator $\widehat{\boldsymbol{\beta}}$. Denoting by $\boldsymbol{X}$ the matrix with rows $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, we have

$$\widehat{\boldsymbol{\beta}}_\lambda := \arg\min_{\boldsymbol{b}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda \|\boldsymbol{b}\|^2 \right\} \tag{1}$$

$$= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\mathsf{T} \boldsymbol{y} \,. \tag{2}$$

We will also be interested in the $\lambda \to 0+$ limit of this estimator which (in the overparametrized case) corresponds to the minimum norm interpolator of the data, and refer to it as 'ridgeless regression.' We will denote by $\boldsymbol{\beta} := \arg\min_{\boldsymbol{b}} \mathbb{E}\{(y - \boldsymbol{b}^{\mathsf{T}}\boldsymbol{x})^2\}$ the population regressor.

Statistical theory studies this and similar estimators in three different regimes:

1. The classical low-dimensional setting in which $\boldsymbol{x}_i, \boldsymbol{\beta} \in \mathbb{R}^d$ with d fixed and $n \to \infty$. In this regime, the empirical covariance $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}/n$ converges to the population covariance $\boldsymbol{\Sigma} := \mathbb{E}\{\boldsymbol{x}_1\boldsymbol{x}_1^{\mathsf{T}}\}$ (provided the latter exists) and $\widehat{\boldsymbol{\beta}}_\lambda$ is asymptotically normal [VdV00].

2. The (by now) classical high-dimensional regime in which $\boldsymbol{x}_i, \boldsymbol{\beta} \in \mathbb{R}^d$ with $d \gg n$ but: (i) the population covariance $\boldsymbol{\Sigma}$ is well conditioned, and (ii) the population regressor $\boldsymbol{\beta}$ is sparse. In this case it is advised to replace the ℓ_2 penalty $\|\boldsymbol{b}\|^2$ by a sparsity promoting penalty, e.g. $\|\boldsymbol{b}\|_1$ [Tib96, DET05]. In many ways, this regime is similar to the previous one, provided $n \gg s \log d$. While $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}/n$ does not concentrate, its restrictions to subsets of $O(s)$ coordinates do [CT05].

3. The proportional regime in which $n \asymp d$. In this case $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}/n$ does not concentrate, and $\widehat{\boldsymbol{\beta}}_\lambda$ is not consistent, and indeed consistent estimation is generally impossible. However, accurate characterizations of the ridge estimator and its risk can be derived using random matrix theory [Dic16, DW18, HMRT22]. Such characterizations answers the question of ε-consistency: for what sample size, and what data distributions does the ridge estimator achieves error $\mathbb{E}\{\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|^2\} \le \varepsilon$? Similar characterizations hold for other estimators such as the Lasso [BM11, MM21, CMW20], robust M-estimators [BBEKY13, EKBB+13, EK18, DM16], and so on [BKM+19, TAH18, TPT21, CM22].

Despite the wealth of fascinating technical results in this area, this state of affairs leaves open many important questions.

First, it would be important have a unified theoretical framework that does not require the statistician to decide which asymptotics to use. For instance, in order to apply sharp asymptotics in the classical or proportional regimes, it is often assumed that a given pair (n, d) is in fact an element of a sequence $(n, d(n))$ with, respectively, either $d(n) \asymp 1$, or $d(n) \asymp n$.

In practice we are given a single pair, say $(n, d) = (1000, 50)$: should we interpret this as $d \asymp 1$, $d \asymp n$, or yet another regime that is not covered by current theory (e.g., $d \asymp n^{2/3}$)?

In fact, the distinction between three types of asymptotics outlined above is rather the consequence of the technical tools used to derive them, rather than a fundamental statistical phenomenon.

Second, the restriction $d = O(n)$ (or $s = O(n)$ in sparse regression) which is implied both by the proportional and by the classical asymptotics is artificial. While this condition might seem necessary for consistency at fist sight (it might seem that at least d observations are required to estimate d parameters), as shown in [BLLT20, TB20] this is in fact not the case. Further, it is not even clear how to check in practice $d = O(n)$ for a given pair n, d.

Third, it would be important to remove the assumption of a well conditioned $\boldsymbol{\Sigma}$, and derive precise asymptotics for general covariances. We would argue that the ill-conditioned case is most important in practice, since high-dimensional data have often low-dimensional structures.

Fourth, the proportional asymptotics is somewhat un-natural from a statistical viewpoint. Most statisticians are used to think of the data distribution is fixed (in particular, d is fixed), while we sample size n increases. In a standard proportional setting, one instead assumes $n, d \to \infty$ together with $n/d \to \delta$: the data distribution changes with the sample size.

Recent progress on several of these issues was achieved in the context of ridge regression. Among others, [HMRT22] derived a characterization for bias and variance in the proportional regime that

is *non-asymptotic*, i.e. holds up to an approximation error that is explicit and vanishes for large n, d. Using a different approach, [BLLT20, TB20] obtained bounds on bias and variance that hold for arbitrary (possibly infinite) dimension d, in terms of of the decay of eigenvalue of $\mathbf{\Sigma}$. These bounds allow to demonstrate 'benign overfitting,' i.e. choices of $\mathbf{\Sigma}, \boldsymbol{\beta}$ (i.e. data distributions) such that minimum norm interpolator is consistent.

The results [HMRT22, BLLT20, TB20] have limitations. The characterization of the risk proved in [HMRT22] has sharp leading constants, but only holds for $C^{-1} \le n/d \le C$ with C a constant, and holds up to an additive error. However, this error terms can be larger than the actual excess risk when the latter vanishes. The bounds of [BLLT20, TB20], on the other hand, hold up to unspecified multiplicative constants. The proof techniques in these two sets of results are furthermore very different.

In this paper we attempt to provide a unified picture that covers these gaps, by extending the sharp characterization of ridge regression of [HMRT22] beyond the proportional regime. This will allows to recover the benign overfitting results of [BLLT20, TB20] (in several cases) with sharp constants. In doing so, we will extend random matrix theory analysis to cases with $d \gg n$ or $d = \infty$, without restrictions on the condition number of $\mathbf{\Sigma}$. In the case $d = \infty$, the feature vectors $\boldsymbol{x}_i$ are random elements in a separable Hilbert space, whose distribution is fixed (does not change with n), and whose covariance $\mathbf{\Sigma}$ is a trace class self-adjoint operator.

The rest of the paper is organized as follows. The next section describes the setting for our analysis, the main assumptions and the resulting asymptotic characterization. It also provides some intuition and connects our results to earlier work. Section 3 contains the formal statement of our general results, while Section 4 specializes our theorem to regimes of interest and develops tools to check its assumptions. Section 5 evaluates our characterization for certain choices of $\mathbf{\Sigma}, \boldsymbol{\beta}$, and compare the predictions with simulations. Finally, proof are presented in Sections 6 and 7, with most technical steps deferred to the appendices.

2 Setting and characterization

Ridge regression in Hilbert space. We consider the simple linear model

$$y_i = \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i \,, \tag{3}$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is the ground truth signal. The random features $\boldsymbol{x}_i \in \mathbb{R}^d$ and noise ε_i are independent, and the $(\boldsymbol{x}_i, \varepsilon_i)$ are i.i.d. samples with $1 \le i \le n$. We assume $\boldsymbol{x}_i, \varepsilon_i$ are mean zero with covariances $\mathrm{Cov}(\boldsymbol{x}_i) = \mathbf{\Sigma}$ and $\mathrm{Var}(\varepsilon_i) = \tau^2$. Defining the data matrix

$$\boldsymbol{X} = \begin{bmatrix} — & \boldsymbol{x}_1^{\mathsf{T}} & — \\ — & \boldsymbol{x}_2^{\mathsf{T}} & — \\ & \vdots & \\ — & \boldsymbol{x}_n^{\mathsf{T}} & — \end{bmatrix} \in \mathbb{R}^{n \times d} \,,$$

the response vector $\boldsymbol{y} = (y_1, \cdots, y_n)^{\mathsf{T}}$ and the noise vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^{\mathsf{T}}$, we can write in matrix form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \,.$$

In this paper, we assume the dimension $d \in \mathbb{Z}_{\ge 0} \cup \{\infty\}$. When $d < \infty$, we are in the usual setup of linear model with finite dimensional features. In the case $d = \infty$, we assume that the

$\boldsymbol{x}_i$'s' are i.i.d. random vectors from a real, separable Hilbert space $\mathcal{H}$. We will use $\|\boldsymbol{x}\|$ to denote the norm and $\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle$ or $\boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{x}_2$ to denote the scalar product in this space. Given a linear operator $\boldsymbol{A} : \mathcal{H} \to \mathcal{H}$, we denote by $\|\boldsymbol{A}\|$ the associated operator norm.

We will assume the covariance operator $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}]$ to be trace-class, namely

$$\mathrm{Tr}(\boldsymbol{\Sigma}) = \mathbb{E}\{\|\boldsymbol{x}_i\|^2\} < \infty\,,$$

and, without loss of generality, we also assume $\|\boldsymbol{\Sigma}\| = 1$. Recall that, without loss of generality, one can always assume $\mathcal{H}$ to be $\ell_2 := \{\boldsymbol{x} = (x_1, x_2, \cdots,) : \sum_{i=1}^{\infty} x_i^2 < \infty\}$ [Bré11].

For an estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{y})$ we define the excess risk as

$$\mathscr{R}_{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{x}_{\mathsf{new}}, \boldsymbol{y}}\left[(\boldsymbol{x}_{\mathsf{new}}^{\mathsf{T}}\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_{\mathsf{new}}^{\mathsf{T}}\boldsymbol{\beta})^2 \mid \boldsymbol{X}\right] = \mathbb{E}_{\boldsymbol{y}}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}^2 \mid \boldsymbol{X}\right],$$

where $\boldsymbol{x}_{\mathsf{new}}$ is an independent copy of $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ and $\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}}^2 := \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{x}$. We will also refer to this as the 'test error' or the 'generalization error' (although the latter is actually given by the difference between $\mathscr{R}_{\boldsymbol{X}}$ and ts empirical version.) Let us emphasize that in this definition, $\mathscr{R}_{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta})$ is a random quantity because it depends on the data $\boldsymbol{X}$: however, as we will prove, it concentrates around a non-random value.

The generalization error admits a variance-bias decomposition $\mathscr{R}_{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}) = \mathscr{V}_{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}) + \mathscr{B}_{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta})$, with

$$\mathscr{V}_{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}) = \mathrm{Tr}\left(\boldsymbol{\Sigma}\mathrm{Cov}(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X})\right), \qquad \mathscr{B}_{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}}; \boldsymbol{\beta}) = \left\|\mathbb{E}_{\boldsymbol{y}}[\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}] - \boldsymbol{\beta}\right\|_{\boldsymbol{\Sigma}}^2.$$

For ridge regression, we can write explicit forms of variance and bias:

$$\mathscr{V}_{\boldsymbol{X}}(\lambda) = \tau^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I})^{-2}\right), \tag{4a}$$

$$\mathscr{B}_{\boldsymbol{X}}(\lambda) = \lambda^2 \langle \boldsymbol{\beta}, (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\beta}\rangle\,. \tag{4b}$$

Assumptions on the covariates distribution. We impose the following assumptions on the covariates $\boldsymbol{x}_i$ throughout the paper.

Assumption 1. *We assume $\mathbb{E}[\boldsymbol{x}_i] = \boldsymbol{0}$, $\boldsymbol{\Sigma} := \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}]$ is a trace class operator: $\mathrm{Tr}(\boldsymbol{\Sigma}) < \infty$ and (without loss of generality) $\|\boldsymbol{\Sigma}\| = 1$. We denote its eigenvalues by $1 = \sigma_1 \geq \sigma_2 \geq \cdots$ in non-increasing order. We assume $\|\boldsymbol{\beta}\|_{\boldsymbol{\Sigma}^{-1}} := \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}\| < \infty$.*

We further assume $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$ where the following hold.

I. *There exist $\mathsf{d}_{\boldsymbol{\Sigma}} := \mathsf{d}_{\boldsymbol{\Sigma}}(n) \geq n$ such that, for all $1 \leq k \leq \min\{n, d\}$*

$$\sum_{l=k}^{d} \sigma_l \leq \mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\,.$$

II. *There exist $\mathsf{C}_{\boldsymbol{x}} > 0$, such that one of the following condition holds:*

(*a*) ***Independent sub-Gaussian coordinates**: $\boldsymbol{z}_i$ has independent but not necessarily identically distributed coordinates with uniformly bounded sub-Gaussian norm. Namely: each coordinate z_{ij} of $\boldsymbol{z}_i$ satifies $\mathbb{E}[z_{ij}] = 0$, $\mathrm{Var}(z_{ij}) = 1$ and $\|z_{ij}\|_{\psi_2} := \sup_{p \geq 1} p^{-\frac{1}{2}} \left(\mathbb{E}\left[|z_{ij}|^p\right]\right)^{\frac{1}{p}} \leq \mathsf{C}_{\boldsymbol{x}}$.*

(*b*) ***Convex concentration**: allowing $\boldsymbol{z}_i$ to have dependent coordinates, the following holds for any $1-$Lipschitz convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$, and for every $t > 0$*

$$\mathbb{P}\left(|\varphi(\boldsymbol{z}_i) - \mathbb{E}\varphi(\boldsymbol{z}_i)| \geq t\right) \leq 2\exp\left(-t^2/\mathsf{C}_{\boldsymbol{x}}^2\right).$$

The technical motivation for assumption II is to establish concentration of quadratic forms of $\boldsymbol{z}_i$, via Hanson-Wright inequality. We notice that the convex concentration property is implied by any of the following. (*i*) By Talagrand inequality, convex concentration holds for random vectors $\boldsymbol{z}_i$ with independent bounded entries [BLM13, Theorem 7.12]. (*ii*) By Herbst's argument, concentration of Lipschitz functions (and hence in particular convex concentration) holds for random vectors $\boldsymbol{z}_i$ that satisfy a log-Sobolev inequality [BGL$^+$14, Proposition 5.4.1]. (*iii*) Finally, as a special case of the last point, vectors $\boldsymbol{z}_i$ with strongly log-concave probability density function satisfy this condition [BGL$^+$14, Corollary 5.7.2].

The form of Hanson-Wright inequality that we will use is given below.

Lemma 2.1 (Hanson-Wright inequality [Ada15, RV13])**.** *Suppose $\boldsymbol{x} \in \mathbb{R}^d$ is a random copy of the features vector $\boldsymbol{x}_i$ satisfying Assumption 1. Then there exists a universal constant $c_0 > 0$ such that, for any matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ with $\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}^{\frac{1}{2}}\| < \infty$, we have*

$$\mathbb{P}\left(\left|\boldsymbol{x}^{\mathsf{T}}\boldsymbol{M}\boldsymbol{x} - \mathrm{Tr}\left(\boldsymbol{\Sigma}\boldsymbol{M}\right)\right| \geq t\right) \leq 2\exp\left\{-c_0\,\min\left(\frac{t^2}{\mathsf{C}_{\boldsymbol{x}}^4\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}^{\frac{1}{2}}\|_F^2}, \frac{t}{\mathsf{C}_{\boldsymbol{x}}^2\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Sigma}^{\frac{1}{2}}\|}\right)\right\}.$$

Remark 2.1. The results of [Ada15, RV13] are stated for finite d. However, the inequality also holds for $d = \infty$ on the Hilbert space ℓ_2 by a standard approximation argument. Namely, one can project the vector $\boldsymbol{x}$ on the span of the top k-eigenvectors of $\boldsymbol{\Sigma}$, establish concentration, and take $k \to \infty$ at the end.

Effective variance and bias. An important observation of [HMRT22] is that variance $\mathscr{V}_{\boldsymbol{X}}$ and bias $\mathscr{B}_{\boldsymbol{X}}$ concentrate around some non-random quantities, that can be interpreted in terms of an 'effective' regression problem. While [HMRT22] proves such characterization in the proportional regime $n \asymp d$, here we will extend its validity and prove stronger guarantees.

Define the effective regularization $\lambda_\star$ as the unique non-negative solution of

$$n - \frac{\lambda}{\lambda_\star} = \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-1}\right), \tag{5}$$

we that define the effective variance and bias as

$$\mathsf{V}_n(\lambda) := \frac{\tau^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)}, \tag{6}$$

$$\mathsf{B}_n(\lambda) := \frac{\lambda_\star^2 \langle \boldsymbol{\beta}, (\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}\rangle}{1 - n^{-1}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)}, \tag{7}$$

$$\mathsf{R}_n(\lambda) := \mathsf{B}_n(\lambda) + \mathsf{V}_n(\lambda). \tag{8}$$

Our main result —stated in the next section— will establish dimension-free guarantees of the form

$$\mathscr{V}_{\boldsymbol{X}} = (1 + o_n(1))\mathsf{V}_n \quad \mathscr{B}_{\boldsymbol{X}} = (1 + o_n(1)))\mathsf{B}_n \tag{9}$$

where the term These improve over earlier work in two important directions. First, they are *dimension free*, and in particular do not assume $n \asymp d$. Second, they provide *multiplicative approximations*, and hence retain their utility when the risk is small.

Bounds, interpretation, benign overfitting. Before stating our formal results relating $\mathscr{V}_{\boldsymbol{X}}$ to V_n and $\mathscr{B}_{\boldsymbol{X}}$ to B_n, it is useful to develop some intuition about the expressions (6), (7) and their immediate consequences. Note that, by Eq. (5), we necessarily have

$$\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right) < \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right) \le n\,. \tag{10}$$

If we assume that inequality between the first and last term holds with a constant multiplicative factor, i.e. $\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right) \le n(1-c_\star^{-1})$ for some constant $c_\star \in (0,\infty)$, then we get

$$\mathsf{V}_n(\lambda) \le \frac{c_\star \tau^2}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right)\,, \tag{11}$$

$$\mathsf{B}_n(\lambda) \le c_\star \lambda_\star^2 \langle \boldsymbol{\beta}, (\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}\rangle\,. \tag{12}$$

Comparing these bounds with the bias and variance of general ridge regression in Eqs. (4a), (4a), we observe that the right hand sides are (modulo the factor $c_\star$) the bias and variance of a modified ridge regression in which:

- The design matrix is non-random and given by $\boldsymbol{\Sigma}^{1/2}$ instead of $\boldsymbol{X}$.
- The regularization parameter is $\lambda_\star$ instead of λ.
- The noise level is $\tau/\sqrt{n}$ instead of τ.

Even more explicit expressions can be obtained by writing the right-hand side of Eqs. (11), (12) in the basis that diagonalizes $\boldsymbol{\Sigma}$ as in the next proposition. A proof of this statement is in Appendix A.

Proposition 2.2. *Assume* $\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right) \le n(1-c_\star^{-1})$, *for* $c_\star \in (1,\infty)$. *Let* $\boldsymbol{\Sigma} := \sum_{i\ge 1}\sigma_i \boldsymbol{v}_i \boldsymbol{v}_i^{\mathsf{T}}$ *be the eigendecomposition of of* $\boldsymbol{\Sigma}$, *and denote by* $\boldsymbol{\beta}_{\le k} := \sum_{i\le k}\langle \boldsymbol{\beta}, \boldsymbol{v}_i\rangle \boldsymbol{v}_i$ *the orthogonal projection of* $\boldsymbol{\beta}$ *onto the span of* $\boldsymbol{v}_1,\dots,\boldsymbol{v}_k$, *and by* $\boldsymbol{\beta}_{>k} := \boldsymbol{\beta} - \boldsymbol{\beta}_{\le k}$ *its complement. Finally, let* $k_\star := \max\{k\,:\,\sigma_k \ge \lambda_\star\}$, *and define the tail effective rank parameters by*

$$r_q(k) := \sum_{\ell>k}\left(\frac{\sigma_\ell}{\sigma_{k+1}}\right)^q, \quad \overline{r}(k) := \frac{r_1(k)^2}{r_2(k)}\,. \tag{13}$$

Then, defining $b_k := \sigma_k/\sigma_{k+1}$, *we have* $2n \ge k_\star + r_1(k_\star)/b_{k_\star}$ *and*

$$\mathsf{V}_n(\lambda) \le c_\star \tau^2\left(\frac{k_\star}{n} + \frac{r_2(k_\star)}{n}\right) \le c_\star \tau^2\left(\frac{k_\star}{n} + \frac{4b_{k_\star}^2 n}{\overline{r}(k_\star)}\right), \tag{14}$$

$$\mathsf{B}_n(\lambda) \le c_\star\left(\sigma_{k_\star}^2 \|\boldsymbol{\beta}_{\le k_\star}\|_{\boldsymbol{\Sigma}^{-1}}^2 + \|\boldsymbol{\beta}_{>k_\star}\|_{\boldsymbol{\Sigma}}^2\right). \tag{15}$$

(We notice that if the singular values σ_k do not decay faster than exponentially, then b_k is of order one.) While these are only bounds on the theoretical characterization $\mathsf{B}_n(\lambda)$, $\mathsf{V}_n(\lambda)$ for bias and variance, our main resuls (Theorem 1and Theorem 3) will allow to transfer them to the actual bias and variance $\mathscr{B}_{\boldsymbol{X}}(\lambda)$, $\mathscr{V}_{\boldsymbol{X}}(\lambda)$ (modulo additional error terms).

Remark 2.2. These bounds (more precisely, the bounds on $\mathscr{B}_{\boldsymbol{X}}(\lambda)$, $\mathscr{V}_{\boldsymbol{X}}(\lambda)$ that follow from these and Theorem 1) are closely related to the ones in [BLLT20, TB20], see in particular [TB20, Theorem 1]. It is worth pointing out two important differences. *First,* the bounds in Eqs. (14), (15) are somewhat more precise/explicit: there is no unspecified constant factor[1], no dependence on the condition number of $\sigma_1/\sigma_{k_\star}$, and no multiplicative factor depending on the probability. *Second,* Eqs. (14), (15) are only proved for the specific value of $k_\star$ defined there.

[1]The factor $c_\star$ is explicit and, if useful, can be replaced by the original expression.

Remark 2.3. The bounds of Eqs. (14), (15) allow to characterize settings in which the excess test error (as predicted by our theory) vanishes. Indeed, for $\mathsf{V}_n(\lambda)$ to vanish, it is sufficient that $k_\star/n \to 0$ and $\overline{r}(k_\star)/n \to \infty$. A simple sufficient condition for $\mathsf{B}_n(\lambda) \to 0$ is that $\boldsymbol{\beta} \in \text{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$ with $\sigma_k/\sigma_{k_\star} \to \infty$.

We will discuss special examples in Section 4, and show how our general results allow to derive more precise estimates of the risk in those cases.

Equivalent sequence model. The discussion above relies on the assumption $\text{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right) \le n(1-c_\star^{-1})$, which implies the simple bounds (11), (12). However the interpretation in terms of a modified ridge regression problem holds for the exact formulas of Eqs. (6), (7). This interpretation was developed in the context of earlier work on the proportional asymptotics [DJM13], but it is useful to spell it out here for the present context.

In the modified model, we observe $\boldsymbol{y}^s$ that is related to $\boldsymbol{\beta}$ according to

$$\boldsymbol{y}^s = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta} + \frac{\omega}{\sqrt{n}}\boldsymbol{g}\,, \quad \boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_d)\,, \tag{16}$$

Without loss of generality, we can work in the basis in which $\boldsymbol{\Sigma}$ is diagonal, and therefore rewrite the above as $y_i^s = \sigma_i^{1/2}\beta_i + (\omega/\sqrt{n})g_i$, which coincides with the definition of the classical sequence model [Tsy09].

We use ridge regression at regularization level $\lambda_\star$ as defined in Eq. (5):

$$\widehat{\boldsymbol{\beta}}^s_\lambda := \text{argmin}_{\boldsymbol{b}}\big\{\|\boldsymbol{y}^s - \boldsymbol{\Sigma}^{1/2}\boldsymbol{b}\|^2 + \lambda_\star\|\boldsymbol{b}\|^2\big\}\,. \tag{17}$$

Finally, choose the noise level ω to be the unique positive solution of

$$\omega^2 = \tau^2 + \mathbb{E}_{\boldsymbol{g}}\big\{\|\widehat{\boldsymbol{\beta}}^s_\lambda - \boldsymbol{\beta}\|^2_{\boldsymbol{\Sigma}}\big\}\,. \tag{18}$$

Then our theoretical prediction for the excess test error $\mathsf{R}_n(\lambda)$ coincides with the excess test error of the sequence model:

$$\mathsf{R}_n(\lambda) = \mathbb{E}_{\boldsymbol{g}}\big\{\|\widehat{\boldsymbol{\beta}}^s_\lambda - \boldsymbol{\beta}\|^2_{\boldsymbol{\Sigma}}\big\}\,. \tag{19}$$

Summarizing, the predicted test error for the original model is equal to the test error in the sequence model, albeit at a different value of the ridge regularization parameter and of the noise level. Needless to say, studying the sequence model is significantly simpler than the original model (3).

3 Statement of main results

Big-Oh notation. For two functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ (where $\boldsymbol{x}$ can be a scalar or a vector), we write $f(\boldsymbol{x}) = \mathcal{O}_{\boldsymbol{\alpha}}(g(\boldsymbol{x}))$ if there exists a constant $\mathsf{C}_{\boldsymbol{\alpha}}$ depending only on the value of $\boldsymbol{\alpha}$ (also $\boldsymbol{\alpha}$ can be either a scalar or a vector) such that $|f(\boldsymbol{x})| \le \mathsf{C}_{\boldsymbol{\alpha}}|g(\boldsymbol{x})|$ for all $\boldsymbol{x}$. In particular, if the constant is universal we write $f(\boldsymbol{x}) = \mathcal{O}(g(\boldsymbol{x}))$. Similarly, we write $f(\boldsymbol{x}) = \Omega_{\boldsymbol{\alpha}}(g(\boldsymbol{x}))$ if $|f(\boldsymbol{x})| \ge \mathsf{C}_{\boldsymbol{\alpha}}|g(\boldsymbol{x})|$ for all $\boldsymbol{x}$ and some constant $\mathsf{C}_{\boldsymbol{\alpha}} > 0$. Finally, we write $f(\boldsymbol{x}) = \Theta_{\boldsymbol{\alpha}}(g(\boldsymbol{x}))$ if we have both $f(\boldsymbol{x}) = \mathcal{O}_{\boldsymbol{\alpha}}(g(\boldsymbol{x}))$ and $f(\boldsymbol{x}) = \Omega_{\boldsymbol{\alpha}}(g(\boldsymbol{x}))$.

We wil state three theorems: the first one for ridge regression with positive regularization $\lambda > 0$ (Theorem 1), and the other two for the ridgeless case $\lambda = 0+$ (Theorem 2 for the overparametrized regime, and Theorem 3 for the underparametrized one). Our approximation guarantees will depend on the pair $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$ through the following three quantities (in the case $\lambda = 0+$, these quantities will be modified as described below):

1. The ratio between effective dimension and regularization parameter:

$$\chi_n(\lambda) := 1 + \frac{\sigma_{\lfloor \eta n \rfloor} \mathsf{d}_{\mathbf{\Sigma}} \log^2(\mathsf{d}_{\mathbf{\Sigma}})}{\lambda} \,. \tag{20}$$

   Here η is a constant that only depends on $\mathsf{C}_{\boldsymbol{x}}$, and hence we will leave it implicit.

2. The ratio between regularization and effective regularization

$$\kappa := \min\left(\frac{\lambda}{n\lambda_\star}; 1 - \frac{\lambda}{n\lambda_\star}\right) > 0 \,. \tag{21}$$

3. For a positive semi-definite operator $\boldsymbol{Q}$, define the modified population resolvent:

$$\mathscr{R}_0(\mu_0, \mu; \boldsymbol{Q}) := \mathrm{Tr}\left(\mathbf{\Sigma}^{\frac{1}{2}} \boldsymbol{Q} \mathbf{\Sigma}^{\frac{1}{2}} (\mu_0 \boldsymbol{I} + \mu \mathbf{\Sigma})^{-1}\right) . \tag{22}$$

   Letting $\boldsymbol{\beta} = \mathbf{\Sigma}^{1/2}\boldsymbol{\theta}$, $\|\boldsymbol{\theta}\| < \infty$, we consider the ratio

$$\rho(\lambda) := \frac{\mathscr{R}_0(\lambda_\star, 1; \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} / \|\boldsymbol{\theta}\|^2)}{\mathscr{R}_0(\lambda_\star, 1; \boldsymbol{I})} \in (0, 1] \,. \tag{23}$$

We next present our master theorem for ridge regression: its proof is postponed to Section 6.

Theorem 1 (Ridge regression)**.** *Under Assumption 1, for any positive integers k and D, there exist constants $\eta = \eta(\mathsf{C}_{\boldsymbol{x}}) \in (0, 1/2)$ and $\mathsf{C} = \mathsf{C}(\mathsf{C}_{\boldsymbol{x}}, D) > 0$ such that the following hold. Define $\chi_n(\lambda), \kappa, \rho(\lambda)$ as above (with $\eta = \eta(\mathsf{C}_{\boldsymbol{x}})$ in Eq. (20)).*

If it holds that

$$\chi_n(\lambda)^3 \log^2 n \le \mathsf{C} n \kappa^{4.5} \,, \qquad n^{-2D+1} = \mathcal{O}\left(\sqrt{\frac{\kappa^3 \log^2 n}{\max\{n, \lambda\}}}\right) ,$$

then for all $n = \Omega_{k,D}(1)$, with probability $1 - \mathcal{O}_k(n^{-D+1})$ we have:

1. ***Variance approximation.***

$$|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| = \mathcal{O}_{k, \mathsf{C}_{\boldsymbol{x}}, D}\left(\frac{\chi_n(\lambda)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}}\right) \cdot \mathsf{V}_n(\lambda) \,.$$

2. ***Bias approximation.*** *If we additionally have $\chi_n(\lambda)^3 \log^2 n \le \mathsf{C} n \kappa^{4.5} \sqrt{\rho(\lambda)}$ and $\lambda k n^{-\frac{1}{k}} \le n\kappa/2$, for all $n = \Omega_{k,D}(1)$, we have*

$$|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| = \mathcal{O}_{k, \mathsf{C}_{\boldsymbol{x}}, D}\left(\frac{\lambda_\star(\lambda)^{k+1}}{n\kappa^3} + \frac{\chi_n(\lambda)^3 \log^2 n}{\sqrt{\rho(\lambda)} n^{1-\frac{1}{k}} \kappa^{8.5}}\right) \cdot \mathsf{B}_n(\lambda) \,.$$

Remark 3.1. The condition $\|\boldsymbol{\beta}\|_{\mathbf{\Sigma}^{-1}} < \infty$ in Assumption 1 amounts to requiring that the coefficients of $\boldsymbol{\beta}$ in the basis of eigenvectors $\boldsymbol{v}_i$ of $\mathbf{\Sigma}$ decay fast enough. Namely, it is equivalent to $\sum_i \langle \boldsymbol{v}_i, \boldsymbol{\beta}\rangle^2 / \sigma_i < \infty$. This condition appears to be a proof artifact and it would be interesting to relax it.

As mentioned above, the conditions on the isotropic random vectors $\boldsymbol{z}_i$ in Assumption 1 are mainly imposed to be able to apply Hanson-Wright inequality (Lemma 2.1). It is an interesting research question to analyze ridge regression for covariates which do not satisfy this inequality.

We next consider the ridgeless limit for in the overparametrized case: recall that $\widehat{\boldsymbol{\beta}}_\lambda$ coincides in this case with the minimum norm interpolator. In this case we need to modify the quantities defined above to measure the quality of our approximation. We begin by noting that Eq. (5) makes perfect sense in the case $\lambda = 0$ and we have $\lim_{\lambda\downarrow 0}\lambda_\star(\lambda) = \lambda_\star(0) > 0$. We then use the following definitions.

1. We replace $\chi_n(\lambda)$ of Eq. (20) by:

$$\chi_n'(\kappa) := 1 + \frac{\sigma_{\lfloor \eta n\rfloor}\mathsf{d}_{\boldsymbol{\Sigma}}\log^2(\mathsf{d}_{\boldsymbol{\Sigma}})}{\kappa n\lambda_\star(0)}\,, \tag{24}$$

   where κ will be introduced in the theorem statement.

2. The quantity $\rho(\lambda)$ defined in Eq. (23) has a well defined limit as $\lambda\downarrow 0$, given by

$$\rho(0) := \frac{\mathscr{R}_0(\lambda_\star(0), 1; \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}/\|\boldsymbol{\theta}\|^2)}{\mathscr{R}_0(\lambda_\star(0), 1; \boldsymbol{I})} \in (0, 1]\,.$$

3. Finally we define

$$\mathsf{C}_{\boldsymbol{\Sigma}} := 1 - \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star(0)\boldsymbol{I})^{-2}\right) \in (0, 1)\,.$$

Before giving the statement, we introduce a piece of terminology. We say that A happens on the event E with probability at least $1 - \Delta$ if $\mathbb{P}(A^c \text{ and } E) \le \Delta$ (and, as a consequence, $\mathbb{P}(A) \ge 1 - \Delta - \mathbb{P}(E^c)$).

Theorem 2 (Ridgeless regression in the overparameterized regime)**.** *Suppose Assumption 1 holds with $n < d$. Further assume $\sigma_n > 0$, and let s_{min} be the minimum nonzero eigenvalue of the sample covariance $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}/n$. For any positive integers k and D, there exist constants $\eta = \eta(\mathsf{C}_{\boldsymbol{x}}) \in (0, 1/2)$ and $\mathsf{C}_1 = \mathsf{C}_1(\mathsf{C}_{\boldsymbol{x}}, D) > 0$, $\mathsf{C}_i = \mathsf{C}_i(k, \mathsf{C}_{\boldsymbol{x}}, D) > 0$, $i \in \{2, 3\}$, such that the following hold. Define $\chi_n'(\kappa)$, $\rho(0)$, $\mathsf{C}_{\boldsymbol{\Sigma}}$ as above.*

Let $\kappa > 0$ be such that the following hold

$$\kappa \le \mathsf{C}_{\boldsymbol{\Sigma}}^2/8\,, \qquad \chi_n'(\kappa)^3\log^2 n \le \mathsf{C}_1 n\kappa^{4.5}\,, \qquad n^{-2D+1} = \mathcal{O}\left(\sqrt{\frac{\kappa^3\log^2 n}{\max\{n, \lambda\}}}\right)\,.$$

Then, on the event $\{s_{\mathsf{min}} \ge 8\lambda_\star(0)\kappa\}$, the following hold with probability $1 - \mathcal{O}_k(n^{-D+1})$:

1. ***Variance approximation.*** *If in addition $\chi_n'(\kappa)^3\log^2 n \le \mathsf{C}_2 n^{1-\frac{1}{k}}\kappa^{9.5}$, then*

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| = \mathcal{O}_{k,\mathsf{C}_{\boldsymbol{x}},D}\left(\kappa\cdot\left(\frac{\lambda_\star(0)}{s_{\mathsf{min}}} + \frac{1}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}\right) + \frac{\chi_n'(\kappa)^3\log^2 n}{n^{1-\frac{1}{k}}\kappa^{9.5}}\right)\cdot\mathsf{V}_n(0)\,.$$

2. ***Bias approximation.*** *If in addition $\chi_n'(\kappa)^3\log^2 n \le \mathsf{C}_1 n\kappa^{4.5}\sqrt{\rho(0)}$, $\lambda_\star(0)kn^{-\frac{1}{k}} \le 1/4$ and*

$$\frac{\lambda_\star(0)^{k+1}}{n\kappa^3} + \frac{\chi_n'(\kappa)^3\log^2 n}{\sqrt{\rho(0)}n^{1-\frac{1}{k}}\kappa^{8.5}} \le \mathsf{C}_3\,,$$

   then, for $n = \Omega_{k,D}(1)$,

   $|\mathscr{B}_{\boldsymbol{X}}(0) - \mathsf{B}_n(0)|$

$$= \mathcal{O}_{k,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}+\frac{\lambda_\star(0)^{k+1}}{n\kappa^3}+\frac{\chi_n'(\kappa)^3\log^2 n}{\sqrt{\rho(0)}n^{1-\frac{1}{k}}\kappa^{8.5}}\right)\cdot \mathsf{B}_n(0)$$
$$+\min\left\{\mathcal{O}\left(\frac{\kappa\lambda_\star(0)\,\|\boldsymbol{\beta}\|^2}{s_{\mathsf{min}}}\right), \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa^2\lambda_\star(0)^2\chi_n'(\kappa)^2)\,\|\boldsymbol{\theta}_{\le n}\|^2+\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa\lambda_\star(0)\chi_n'(\kappa))\,\|\boldsymbol{\beta}_{>n}\|^2\right\}.$$

The proof of this theorem is presented in Appendix D.

Remark 3.2. Our approach to proving Theorem 2 consists in reducing the ridgeless case $\lambda = 0+$ to the case $\lambda > 0$, and appealing to Theorem 1. For instance, when controlling the variance, we will use triangular inequality

$$|\mathscr{V}_{\boldsymbol{X}}(0)-\mathsf{V}_n(0)| \le |\mathscr{V}_{\boldsymbol{X}}(\lambda)-\mathsf{V}_n(\lambda)|+|\mathscr{V}_{\boldsymbol{X}}(0)-\mathscr{V}_{\boldsymbol{X}}(\lambda)|+|\mathsf{V}_n(0)-\mathsf{V}_n(\lambda)|\,.$$

We then use Theorem 1 to bound the first term by a quantity that diverges as $\lambda \downarrow 0$, and the main technical challenge is in bounding the other two terms by a quantity that vanishes faster than any polynomial as $\lambda \downarrow 0$.

Remark 3.3. In Theorem 2 we use the (random) minimum nonzero eigenvalue s_{min} of the sample covariance $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}/n$. To apply the theorem, we need to choose κ such that $\{s_{\mathsf{min}} \ge 8\lambda_\star(0)\kappa\}$ holds with high probability, and therefore we need a lower bound on s_{min} that holds with high probability. In this paper, we will provide such lower bonds in two cases: (i) proportional regime and (ii) bounded varying spectrum, cf. Section 4. In proportional regime, $s_{\mathsf{min}} = \Omega_{|d/n-1|^{-1}}(1)$ by Bai-Yin law; and for for bounded varying spectrum, $s_{\mathsf{min}} = \Omega(\sigma_n)$ (cf. Lemma G.1).

Beyond the two cases in the paper, it would be interesting to apply Theorem 2 with results lower bounding s_{min} for other examples (e.g. for the kernel random matrices [HLCH19]).

In the underparameterized regime $d < n$, we have $\lim_{\lambda\downarrow 0}\lambda_\star(\lambda) = 0$ and therefore the previous bounds do not apply. In this case, we trivially have $\mathscr{B}_{\boldsymbol{X}}(0) = \mathsf{B}_n(0) = 0$. The proof for the variance approximation requires a different proof, which is presented in Appendix E.

Theorem 3 (Ridgeless regression in the underparameterized regime)**.** *Suppose Assumption 1 holds with $n > d$, and further assume*

$$\mathsf{C}_{\boldsymbol{\Sigma}} = \min\left(\frac{d}{n}, 1-\frac{d}{n}\right) \in (0,1)\,.$$

For any positive integers k and D, there exist constants $\eta = \eta(\mathsf{C}_{\boldsymbol{x}}) > 0$ $\mathsf{C}_1 = \mathsf{C}_1(\mathsf{C}_{\boldsymbol{x}}, D) > 0$, $\mathsf{C}_2 = \mathsf{C}_2(k, \mathsf{C}_{\boldsymbol{x}}, D) > 0$, such that the following hold.

If $\boldsymbol{X}$ has rank d and s_{min} is the minimum eigenvalue of the sample covariance $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}/n$, then the following hold:

1. ***Variance approximation.*** *Let ε be such that*

$$\varepsilon \le \mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d/4\,, \qquad n^{-2D+1} = \mathcal{O}\left(\sqrt{\frac{\mathsf{C}_{\boldsymbol{\Sigma}}^3\log^2 n}{\max\{n,\lambda\}}}\right),$$
$$\chi_n(\varepsilon n)^3\log^2 n \le \mathsf{C}_1 n\mathsf{C}_{\boldsymbol{\Sigma}}^{4.5}\,, \qquad \chi_n(\varepsilon n)^3\log^2 n \le \mathsf{C}_2 n^{1-\frac{1}{k}}\mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}\,.$$

   Then, on the event $\{s_{\mathsf{min}} \ge 2\varepsilon\}$, with probability $1-\mathcal{O}_k(n^{-D+1})$:

$$|\mathscr{V}_{\boldsymbol{X}}(0)-\mathsf{V}_n(0)| = \mathcal{O}_{k,\mathsf{C}_{\boldsymbol{x}},D}\left(\varepsilon\cdot\left(\frac{1}{s_{\mathsf{min}}}+\frac{1}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d}\right)+\frac{\chi_n(\varepsilon n)^3\log^2 n}{n^{1-\frac{1}{k}}\mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}}\right)\cdot \mathsf{V}_n(0)\,.$$

2. ***Bias approximation.*** $\mathscr{B}_{\boldsymbol{X}}(0) = \mathsf{B}_n(0) = 0$ *(this holds deterministically on the event* $\mathrm{rank}(\boldsymbol{X}) = d$*).*

4 Applications

4.1 Proportional regime

As a first application, we revisit the proportional regime that is defined by the following condition.

Assumption 2. *There exists a constant $M > 1$ such that $M^{-1} \le d/n \le M$ and $\sigma_d \ge M^{-1}$.*

This case is well studied and is not the main motivation of the present paper, but it is nevertheless important to compare our results to earlier work. We refer the reader to [Dic16, ASS20, DW18, WX20, RMR21] for background.

Among others, the results of [HMRT22] are more directly comparable to ours because they establish nonasymptotic bounds comparing variance and bias to the effective variance and bias of Eqs. (6) and (7), for both ridge and ridgeless regression. The proofs of [HMRT22] build on recent advances in random matrix theory, and in particular the anisotropic local law of [KY17].

Here we apply Theorems 1, 2 and 3 to the proportional regime. We note that, under assumption 2, the minimum eigenvalue of $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ is, with high probability, of order n. In order for the ridge regularization to have a non-trivial effect, we need to choose $\lambda \asymp n$ as well, cf. (4a) and (4b). We will therefore assume λ/n bounded above and below (there is no loss of generality in using the same constant as in Eq. (2)). We will address the case $\lambda = 0+$ in a separate statement below.

Proposition 4.1. *Let Assumptions 1 and 2 hold, and further assume $\lambda/n \in [1/M, M]$. Then for any positive integers k and D, if $n = \Omega_{k,M,\mathsf{C}_{\boldsymbol{x}},D}(1)$, with probability $1 - \mathcal{O}_k(n^{-D+1})$ we have*

$$|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| = \mathcal{O}_{k,M,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\log^8 n}{n^{1-\frac{1}{k}}}\right) \cdot \mathsf{V}_n(\lambda)\,,$$
$$|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| = \mathcal{O}_{k,M,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\log^8 n}{n^{\frac{1}{2}-\frac{1}{k}}}\right) \cdot \mathsf{B}_n(\lambda)\,.$$

The proof of this result is presented in Appendix F.

We note that the rates $\mathcal{O}(n^{-1})$ and $\mathcal{O}(n^{-1/2})$ are optimal for variance and bias approximation—corresponding to fluctuations of the average law and local law for the resolvent [AEK+14, KY17]. The most direct comparison is with [HMRT22, Theorem 5]: let us point out two ways in which the present result improves over the earlier [HMRT22].

- In [HMRT22], the rate for variance approximation of ridge regression is $\mathcal{O}(n^{-1/2})$, while here we obtain the faster rate $\mathcal{O}(n^{-1})$.
- Error terms in [HMRT22] are additive, while Proposition 4.1 provides multiplicative error terms: the quality of approximation does not deteriorate in the interesting case in which bias and variance become small.

Note that [HMRT22] informally claimed that $n^{-1/2}$ is the optimal rate in the above estimates. While this is correct for the bias, for the variance Proposition 4.1 yields a faster rate. As related phenomenon arises for linear eigenvalue statistics of random matrices (i.e. statistics of the form $n^{-1}\sum_{i=1}^n \varphi(\lambda_i)$). While naively such statistics would have normal deviations of order $n^{-1/2}$, the actual deviations are of order n^{-1} because of eigenvalues correlations [LP09].

We finally consider the ridgeless case.

Proposition 4.2. *Let Assumptions 1 and 2 hold for $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$, where $\boldsymbol{z}_i$ has i.i.d. sub-Gaussian coordinates.*

1. ***Overparameterized regime.*** *If additionally* $d/n \geq 1 + M^{-1}$*, then for all* $n = \Omega_{M,\mathsf{C}_{\boldsymbol{x}},D}(1)$*, with probability* $1 - \mathcal{O}(n^{-D+1})$ *we have*

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| = \mathcal{O}_{M,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/14}\right) \cdot \mathsf{V}_n(0)\,,$$
$$|\mathscr{B}_{\boldsymbol{X}}(0) - \mathsf{B}_n(0)| = \mathcal{O}_{M,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/28}\right) \cdot \mathsf{B}_n(0)\,.$$

2. ***Underparameterized regime.*** *If additionally* $d/n \leq 1 - M^{-1}$*, then for all* $n = \Omega_{M,\mathsf{C}_{\boldsymbol{x}},D}(1)$*, with probability* $1 - \mathcal{O}(n^{-D+1})$ *we have*

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| = \mathcal{O}_{M,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/5}\right) \cdot \mathsf{V}_n(0)\,,$$
$$\mathscr{B}_{\boldsymbol{X}}(0) = \mathsf{B}_n(0) = 0\,.$$

We do not expect the exponent 1/14, 1/28, 1/5 in this statement to be tight. However, as in the positive λ case, also in this case the error is multiplicative.

4.2 Bounded varying spectrum

We next consider the highly overparametrized case $d \gg n$. Overparametrized ridge (or minimum norm) regression attracted significant attention recently because of the realization that many deep learning models are overparametrized and overfit the training data. This connection is reviewed in [BMR21, Bel21].

Here we consider covariate vectors $\boldsymbol{x}_i$ taking values in a general Hilbert space with $d = \infty$, under Assumption 1 on the covariates distribution. This is most closely related to [BLLT20, TB20], and [KZSS21]. The last paper derives refined upper bounds using Gaussian width techniques, but is limited to the case of Gaussian covariates and, as for earlier results, is only accurate up to constant factors.

We will show that our general theory yields excess risk estimates that are accurate up to $1+o_n(1)$ multiplicative errors. We impose the following condition on the spectrum of $\boldsymbol{\Sigma}$.

Assumption 3 (Bounded varying spectrum)**.** *There exists a monotone decreasing function* $\psi : (0,1] \to [1,\infty)$ *with* $\lim_{\delta\downarrow 0}\psi(\delta) = \infty$*, such that* $\sigma_{\lfloor \delta i\rfloor}/\sigma_i \leq \psi(\delta)$ *for all* $\delta \in (0,1]$*,* $i \in \mathbb{N}$ *and* $\delta i \geq 1$*.*

Recall that, by definition, for any $j \leq i$, $\sigma_j/\sigma_i \geq 1$. The bounded varying condition requires that, if i, j diverge proportionally, then the eigenvalue ratio σ_j/σ_i stays bounded. Note that this assumption is equivalent to $\sup_{i\geq 1}\sigma_{\lfloor \delta i\rfloor}/\sigma_i < \infty$ for every $\delta \in (0,1]$, which is in turn equivalent to

$$\limsup_{i\to\infty} \frac{\sigma_{\lfloor \delta i\rfloor}}{\sigma_i} < \infty\,. \tag{25}$$

As special case, Assumption 3 holds if the sorted eigenvalues $(\sigma_1, \sigma_2, \cdots)$ forms a so-called *regularly varying sequence*, namely for any $\delta \in (0,\infty)$,

$$\lim_{i\to\infty} \frac{\sigma_{\lfloor \delta i\rfloor}}{\sigma_i} = \psi(\delta)\,,$$

where $\psi(\delta)$ is positive and finite for any δ. In other words, in the regularly varying case, the ratio σ_j/σ_i converges when i, j diverge proportionally.

A special case of regularly varying spectrum is given by Zipf's law whereby $\sigma_i = i^{-\alpha}$ for some $\alpha > 1$ (in this case $\psi(\delta) = \delta^{-\alpha}$). Regularly varying functions were characterized by Karamata

[Kar33] (for functions on the positive real line), and by Galambos and Seneta [GS73] (for the sequences, i.e. functions defined on the naturals). Namely all such sequences take the form

$$\sigma_i = i^{-\alpha} a_i \exp\Big\{ \sum_{j=1}^{i} b_j/j \Big\},$$

where a_i are arbitrary and converge to a positive limit as $i \to \infty$ and $b_i \to 0$.

It is easy to see that Assumption 3 holds beyond the case of regularly varying sequences. Consider for instance $\sigma_i = 3^{-s}$ for all $2^s \le i < 2^{s+1}$, $s = 0, 1, \cdots$.

Applying Theorems 1 and 2 to $\boldsymbol{\Sigma}$ with bounded varying spectrum, we obtain the following result, whose proofs are detailed in Appendix G.

Proposition 4.3. *Let Assumptions 1 and 3 hold. For any constants $M > 0$, $\gamma \in (0, 1/3)$, and positive integers k, D the following holds. If $\mathsf{d}_{\boldsymbol{\Sigma}} \le M n^{1+\gamma}$ and $\lambda \in [n\lambda_\star(0)/M, n\lambda_\star(0)M]$, then for $n = \Omega_{k,M,\psi,\gamma,\mathsf{C}_{\boldsymbol{x}},D}(1)$, with probability $1 - \mathcal{O}_k(n^{-D+1})$*

$$|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| = \mathcal{O}_{k,M,\psi,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{(\mathsf{d}_{\boldsymbol{\Sigma}}/n)^3 \log^8 n}{n^{1-\frac{1}{k}}} \right) \cdot \mathsf{V}_n(\lambda)\,.$$

If additionally $\mathsf{d}_{\boldsymbol{\Sigma}} = \mathcal{O}_{M,\psi,\mathsf{C}_{\boldsymbol{x}}}(n^{1+\gamma}\,(\rho(\lambda))^{1/6})$ $\lambda \in [n\lambda_\star(0)/M, n\lambda_\star(0)M]$ and $\lambda_\star(0) = \mathcal{O}(1)$ (cf. Theorem 1 for the function ρ), with the same probability we have

$$|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| = \mathcal{O}_{k,M,\psi,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{(\mathsf{d}_{\boldsymbol{\Sigma}}/n)^3 \log^8 n}{\sqrt{\rho(\lambda)} n^{1-\frac{1}{k}}} \right) \cdot \mathsf{B}_n(\lambda)\,.$$

Applying Theorem 2, we have the following conclusion for ridgeless regression.

Proposition 4.4. *Let Assumptions 1 and 3 hold. Suppose $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\| < \infty$. If we have $\lambda_\star(0)/\sigma_n = \mathcal{O}(\log^{\mathcal{O}(1)} n)$ and $\mathsf{d}_{\boldsymbol{\Sigma}}(n) = \mathcal{O}(n \log^{\mathcal{O}(1)} n))$, for any $n = \Omega_{\psi,\mathsf{C}_{\boldsymbol{x}},D}(1)$, it holds with probability $1 - \mathcal{O}(n^{-D+1})$ that*

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| = \mathcal{O}_{\psi,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/15} \right) \cdot \mathsf{V}_n(0)\,.$$

Remark 4.1. The assumptions $\lambda_\star(0)/\sigma_n = \mathcal{O}(\log^{\mathcal{O}(1)} n)$ and $\mathsf{d}_{\boldsymbol{\Sigma}} = \mathcal{O}(n \log^{\mathcal{O}(1)} n)$ are primarily introduced to simplify the form of the statement. These two conditions can be relaxed to $lambdaefct(0)/\sigma_n = \mathcal{O}(n^{\bar\gamma})$ and $\mathsf{d}_{\boldsymbol{\Sigma}} = \mathcal{O}(n^{1+\bar\gamma})$ for a sufficiently small $\bar\gamma$, but we do not pursue this generalization here.

Remark 4.2. It is possible to apply the upper/lower bounds on the bias of Theorem 2 to prove bounds on the bias in the setting of Proposition 4.4. However the resulting error term is larger than $(\sigma_n^2 \|\boldsymbol{\theta}_{\le n}\|^2 + \sigma_n \|\boldsymbol{\beta}_{>n}\|^2)$, which is the size of the upper bound on $\mathsf{B}_n(0)$ in Proposition 2.2.

In order to illustrate the accuracy of our general framework, we apply Proposition 4.3 to derive sharp asymptotics for bias and variance in a number cases. In each of the case below, we scale the regularization parameter λ as $\lambda = \lambda_0(n) \cdot \nu$ for a certain explicit function $\lambda_0(n)$. The scaling $\lambda_0(n)$ is chosen so that the bias and variance retain a non-trivial dependence on ν for large n. We expect that the excess risk achieved by optimal regularization is also covered by this scaling (up to negligible corrections), but do not prove it formally here.

Theorem 4. *Let Assumption 1 hold. Then, for a fixed constant $\nu > 0$ and any positive integer D, the following events hold with probability $1 - \mathcal{O}(n^{-D})$ (the $o_n(1)$ errors may depend on D):*

1. ***Regularly varying spectrum with*** $\alpha > 1$**.** *Assume* $(\sigma_i)_{i\geq 1}$ *is a regularly varying sequence with exponent* $\alpha > 1$*. As a consequence,* $\sigma_i = i^{-\alpha} a_i \exp\left\{\sum_{j=1}^{i} b_j/j\right\}$ *with* a_i *converging to a positive limit and* $b_i \to 0$*. Define* $\mathsf{c}_\star = \mathsf{c}_\star(\nu) > 0$ *as the unique positive solution of*

$$1 = \nu \mathsf{c}_\star^{-1} + \frac{\pi/\alpha}{\sin(\pi/\alpha)} \mathsf{c}_\star^{-1/\alpha} .$$

Then we have

$$\lambda_\star(\nu n^{1-\alpha}) = \mathsf{c}_\star \sigma_n (1 + o_n(1)) , \tag{26}$$

$$\mathscr{V}_{\boldsymbol{X}}(\nu n^{1-\alpha}) = \frac{\tau^2 (1 - \nu \mathsf{c}_\star^{-1})(\alpha - 1)}{1 + \nu \mathsf{c}_\star^{-1}(\alpha - 1)} (1 + o_n(1)) . \tag{27}$$

Let $F_{\boldsymbol{\beta}}(x) = \sum_{k=1}^{\lfloor nx \rfloor} \langle \boldsymbol{\beta}, \boldsymbol{v}_k \rangle^2$*. If additionally* $\boldsymbol{\beta}$ *satisfies the following "polynomial-decay" property: for some* $0 < \theta \leq 1$ *that*

$$\int_0^\infty x^\alpha \, \mathrm{d}F_{\boldsymbol{\beta}}(x) = \mathcal{O}\left(n^{1-\theta} \int_0^\infty x^\alpha \, (1 + \mathsf{c}_\star x^\alpha)^{-1} \, \mathrm{d}F_{\boldsymbol{\beta}}(x) \right) ,$$

we further have

$$\mathscr{B}_{\boldsymbol{X}}(\nu n^{1-\alpha}) = \frac{\sigma_n \mathsf{c}_\star^2 \alpha}{1 + \nu \mathsf{c}_\star^{-1}(\alpha - 1)} \int_0^\infty \frac{x^\alpha}{(1 + \mathsf{c}_\star x^\alpha)^2} \mathrm{d}F_{\boldsymbol{\beta}}(x) \big(1 + o_n(1)\big) . \tag{28}$$

2. ***Regularly varying spectrum with*** $\alpha = 1$**.** *Next consider the case* $\sigma_i = i^{-1} a_i (1 + \log i)^{-\alpha'}$ *for some* $\alpha' > 1$ *with* a_i *converging to a positive limit. Define* $\mathsf{c}_\star = \mathsf{c}_\star(\nu) > 0$ *as*

$$\mathsf{c}_\star = \nu + \frac{1}{\alpha' - 1} .$$

We have

$$\lambda_\star(\nu \log^{1-\alpha'} n) = \mathsf{c}_\star \sigma_n \log n (1 + o_n(1)) , \tag{29}$$

$$\mathscr{V}_{\boldsymbol{X}}(\nu \log^{1-\alpha'} n) = \frac{\tau^2}{\mathsf{c}_\star \log n} \big(1 + o_n(1)\big) . \tag{30}$$

Let $F_{\boldsymbol{\beta}}(x) = \sum_{k=1}^{\lfloor (n/\log n) x \rfloor} \langle \boldsymbol{\beta}, \boldsymbol{v}_k \rangle^2$*. If additionally* $\boldsymbol{\beta}$ *satisfies the following "rapid-decay" property: for some* $0 < \theta \leq 1$ *that*

$$\int_0^\infty x \, \mathrm{d}F_{\boldsymbol{\beta}}(x) = \mathcal{O}\left(n^{1-\theta} \int_0^\infty x \, (1 + \mathsf{c}_\star x)^{-1} \, \mathrm{d}F_{\boldsymbol{\beta}}(x) \right) .$$

then we further have

$$\mathscr{B}_{\boldsymbol{X}}(\nu \log^{1-\alpha'} n) = \mathsf{c}_\star^2 \sigma_n \log n \int_0^\infty \frac{x}{(1 + \mathsf{c}_\star x)^2} \, \mathrm{d}F_{\boldsymbol{\beta}}(x) \big(1 + o_n(1)\big) . \tag{31}$$

3. ***A non-regularly varying spectrum.*** $\sigma_i = p^{-s}$ *for all* $q^s \leq i < q^{s+1}$*, with* $1 < q < p$ *and* $s = 0, 1, \ldots$ *Define* $s_\star$ *such that* $q^{s_\star} \leq n < q^{s_\star + 1}$*, and for positive integer* r *the following decreasing function in* $t > 0$*,*

$$G_{p,q,r}(t) = \sum_{k=-\infty}^{\infty} \frac{q^k}{(1 + t p^k)^r} .$$

Let $\rho_\star = n/(q^{s_\star+1}-q^{s_\star}) \in [1/(q-1), q/(q-1))$. Then there exists a unique solution $\mathsf{c}_\star = \mathsf{c}_\star(\nu)$ to the following equation

$$1 = \nu \mathsf{c}_\star^{-1} + \rho_\star^{-1} \cdot G_{p,q,1}(\mathsf{c}_\star)\,.$$

Then we have

$$\lambda_\star(\nu n p^{-s_\star}) = \mathsf{c}_\star \sigma_n \left(1 + o_n(1)\right), \tag{32}$$

$$\mathscr{V}_{\boldsymbol{X}}(\nu n p^{-s_\star}) = \frac{G_{p,q,2}(\mathsf{c}_\star)\tau^2}{\rho_\star - G_{p,q,2}(\mathsf{c}_\star)}\big(1 + o_n(1)\big)\,. \tag{33}$$

Let $F_{\boldsymbol{\beta}}(x) = \sum_{k=1}^{q^{\lceil x\rceil}-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_k\rangle^2$. If additionally $\boldsymbol{\beta}$ satisfies the following "rapid-decay" property: for some $0 < \theta \le 1$ that

$$\int_0^\infty p^{x-s_\star}\,\mathrm{d}F_{\boldsymbol{\beta}}(x) = \mathcal{O}\left(n^{1-\theta}\int_0^\infty p^{x-s_\star}(1+\mathsf{c}_\star p^{x-s_\star})^{-1}\,\mathrm{d}F_{\boldsymbol{\beta}}(x)\right),$$

we further have

$$\mathscr{B}_{\boldsymbol{X}}(\nu n p^{-s_\star}) = \frac{\mathsf{c}_\star^2\sigma_n}{1-\rho_\star^{-1}G_{p,q,2}(\mathsf{c}_\star)}\int_0^\infty \frac{p^{x-s_\star}}{(1+\mathsf{c}_\star p^{x-s_\star})^2}\,\mathrm{d}F_{\boldsymbol{\beta}}(x)\,\big(1+o_n(1)\big)\,. \tag{34}$$

The proof of this theorem is presented in Appendix H.

Remark 4.3. In the case of a regularly varying spectrum with $\alpha > 1$, the bias vanishes with the sample size as $n^{-\alpha+o(1)}$ but the variance stays bounded away from zero as long as $\tau > 0$, cf. Eq. (27). In other words in this case overfitting is not benign and Theorem 4 quantifies precisely this claim.

On the other hand, in the case $\alpha = 1$, both bias and variance vanish for large n, an therefore we achieve benign overfitting. We must emphasize however that the variance decay is very slow, namely $\mathscr{V}_{\boldsymbol{X}}(\lambda) \asymp (\log n)^{-1}$, and hence the decay of the excess risk is at least as slow.

5 Numerical illustrations

In this section we evaluate numerically the theoretical prediction for variance and bias, cf. Eqs. (6), (7) and compare them with the results of numerical simulations with synthetic data. We carry out the simulations in the ridgeless limit $\lambda = 0+$ (corresponding to min-norm interpolation). This case is interesting because it is not covered by some of our theorems. Our numerical experiments suggest that the theoretical predictions of Eqs. (6), (7) hold in a broader domain of validity than the one that we are able to control rigorously.

We use Gaussian covariates $\boldsymbol{x}_i$. By rotational invariance, we can limit ourselves to diagonal covariance $\boldsymbol{\Sigma}$. We will consider two eigenvalue structures:

(*I*) **Regularly varying with $\alpha > 1$.** This is defined by $\sigma_i = i^{-\alpha}$ for all $i \ge 1$. This fits within the first case of Theorem 4.

(*II*) **Regularly varying with $\alpha = 1$.** This model is defined by $\sigma_i = i^{-1}(1+\log i)^{-\alpha'}$, with $\alpha' > 1$. This fits within the second case of Theorem 4.

In all numerical experiments, we generate data according to the model (3) with a true parameters vector $\boldsymbol{\beta}$ concentrated on the top $d_0 = 100$ eigenvectors of $\boldsymbol{\Sigma}$. More precisely, we will use $\boldsymbol{\beta} = (1, 1, \dots, 1, 0, 0, \dots)$ where $\|\boldsymbol{\beta}\|_0 = d_0 = 100$.

In Figure 1, we plot our theoretical predictions V_n, B_n, R_n for variance, bias and as a function of the sample size n, for the two models (I) and (II) defined above. We use $\lambda = 0+$. In each case, we consider several values of the exponents α, α' that control the decay of eigenvalues of $\boldsymbol{\Sigma}$.

In Figure 2, we plot the same quantities at fixed sample size $n = 500$ and vary the regularization parameter λ. A few facts emerge from these figures:

- For both models, the bias of the minimum norm interpolator is a decreasing function of the sample size n, and appears to vanish as $n \to \infty$, see second row of Figure 1.
- In contrast, the variance exhibits a strikingly different behavior in the two covariance models, see first row of Figure 1. For model (I) (polynomial eigenvalue decay, with exponent $\alpha > 1$), the variance increases with n, and eventually stabilizes to a limit value. For model (II) (exponent $\alpha = 1$), the variance decreases with n, and appears to vanish, albeit very slowly, as $n \to \infty$.
- As a consequence of these points, the excess test error of minimum norm interpolation vanishes with sample size in model (II) but does not vanish in model (I). This behavior (and the one at previous points) is precisely quantified by Theorem 4 for $\lambda > 0$.
- Finally the dependence of bias and variance on λ is the expected one. As λ increases, bias increases but variance decreases. However, the balance between these two factors is non-trivial:
  - For the slowest eigenvalue decay (large α in model (I) or large α' in model (II)), the optimal λ is strictly positive.
  - On the other hand, for the fastest eigenvalue decay, the optimal λ vanishes. In these case interpolation is superior to ridge regression: we need to overfit to achieve the best test error.

The above discussion is based on evaluating the theoretical formulas for bias and variance, as given in Eqs. (6), (7). While our main result, Theorems 1, 2 guarantee that these formulas are accurate, it is important how accurate they are at small or moderate n, and whether random deviations modify the picture.

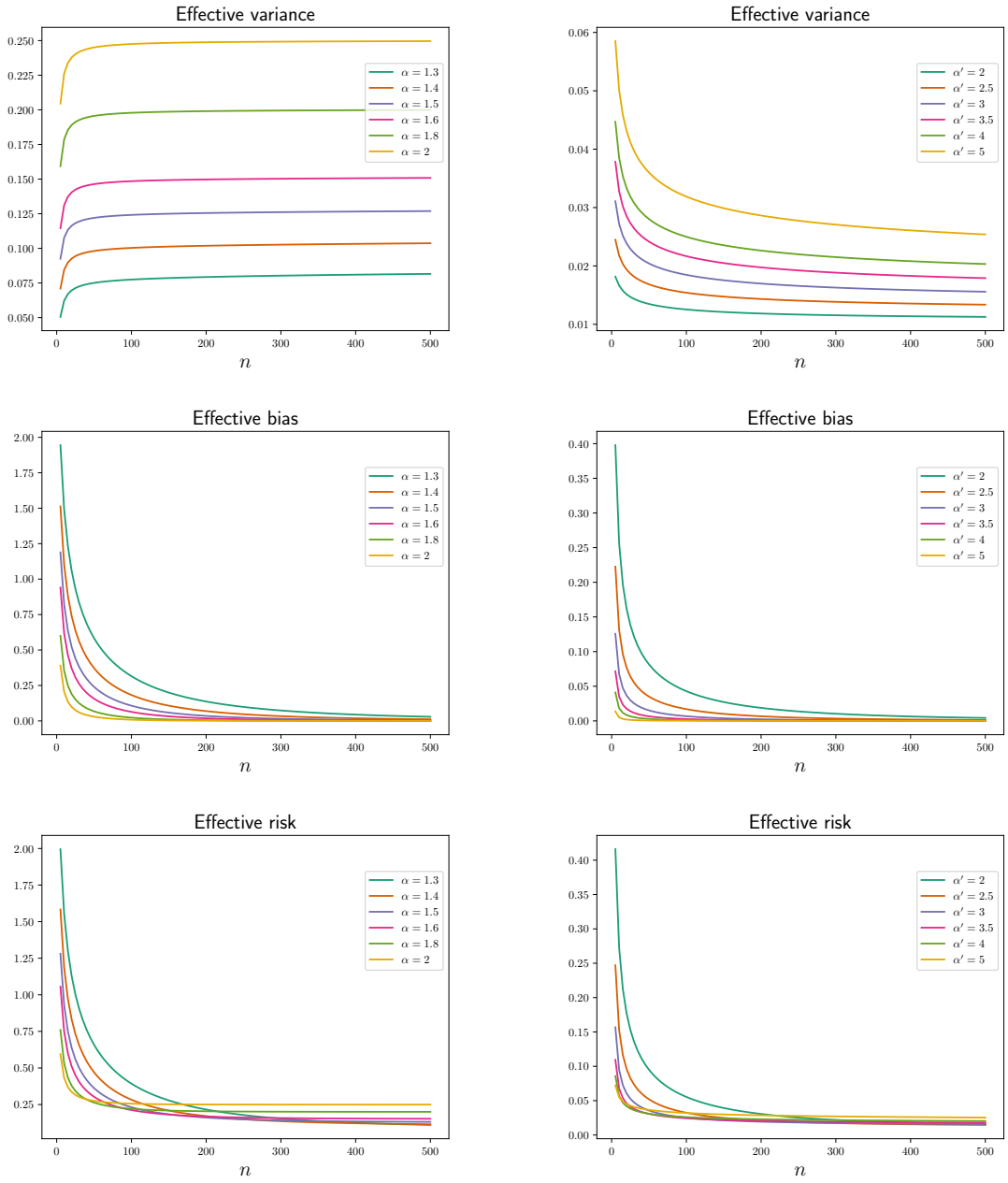
In Figure 3 we plot numerical simulations corroborating that $\mathscr{V}_{\boldsymbol{X}}, \mathscr{B}_{\boldsymbol{X}}$ do concentrate around $\mathsf{V}_n, \mathsf{B}_n$ in models (I) and (II). As mentioned above, the predictions $\mathsf{V}_n, \mathsf{B}_n$ appear to be accurate beyond what is guaranteed by Theorem 2, and the error appears to be a $(1 + o_n(1))$ multiplicative factor.

6 Proof of Theorem 1

Let $\mathcal{F}_k := \sigma(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k)$ be the σ-field generated by the first k data points for $1 \le k \le n$, and $\mathcal{F}_0$ the trivial σ-field. We then have $\mathscr{V}_{\boldsymbol{X}}, \mathscr{B}_{\boldsymbol{X}} \in \mathcal{F}_n$ and $\mathsf{V}_n, \mathsf{B}_n \in \mathcal{F}_0$. Extending the previous notation of $\mathscr{R}_0$ in Eq. (22) to $\mathscr{R}_k$, we let

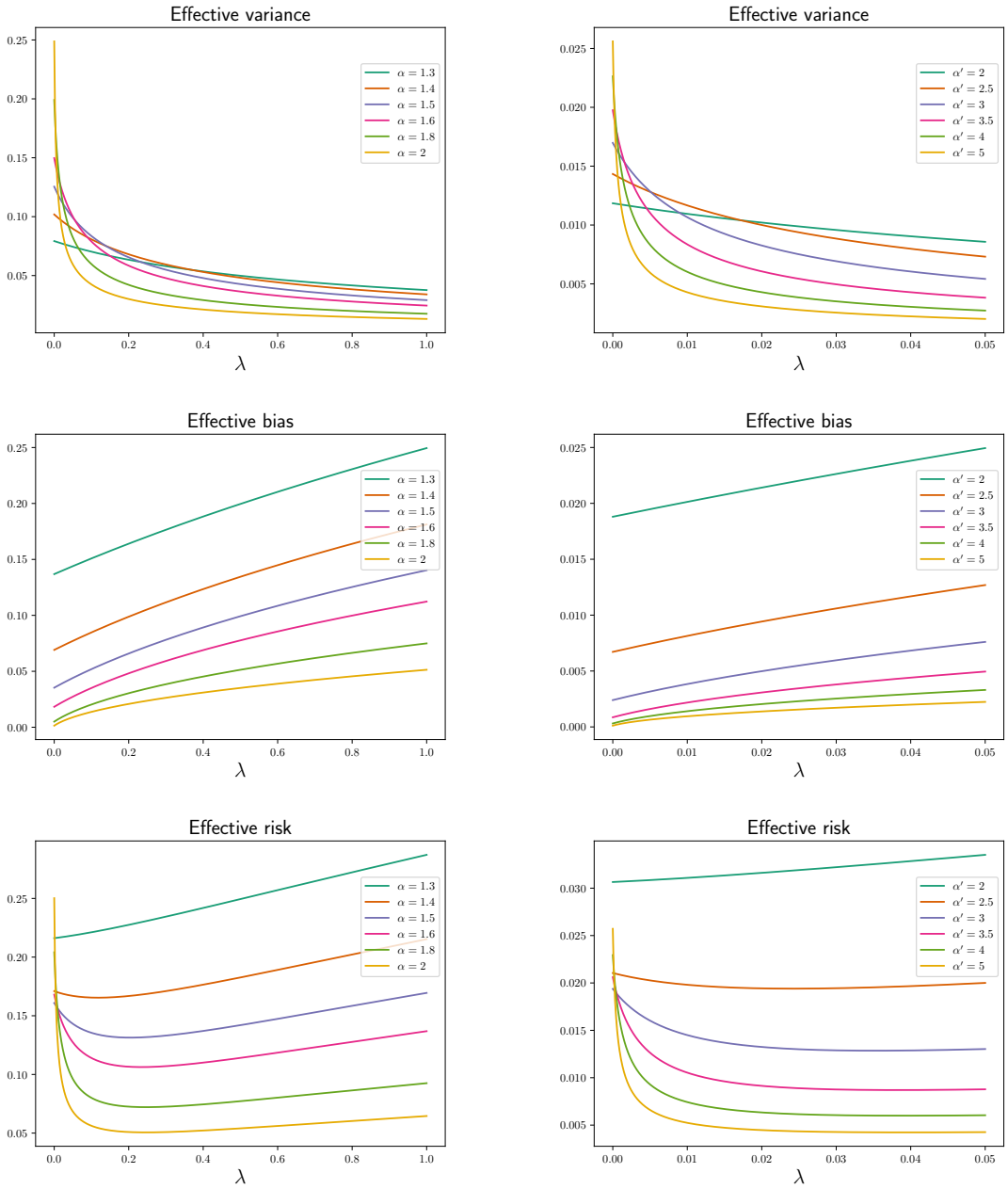
$$\mathscr{R}_k(\mu_0, \mu; \boldsymbol{Q}) = \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{Q} \boldsymbol{\Sigma}^{\frac{1}{2}} (\mu_0 \boldsymbol{I} + \mu \boldsymbol{\Sigma} + \boldsymbol{X}_k^{\mathsf{T}} \boldsymbol{X}_k)^{-1}\right), \qquad \mathscr{F}_k(\mu_0, \mu; \boldsymbol{Q}) = \mu_0 \mathscr{R}_k(\mu_0, \mu; \boldsymbol{Q}), \quad (35)$$

where $\mu_0 > 0, \mu \ge 0$, $\boldsymbol{Q}$ is a p.s.d. matrix with bounded spectral norm, and $\boldsymbol{X}_k = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k]^{\mathsf{T}} \in \mathbb{R}^{k \times d}$ is the partial data matrix comprising the first k rows of $\boldsymbol{X}$. By convention we set $\boldsymbol{X}_0^{\mathsf{T}} \boldsymbol{X}_0 := \boldsymbol{0}$



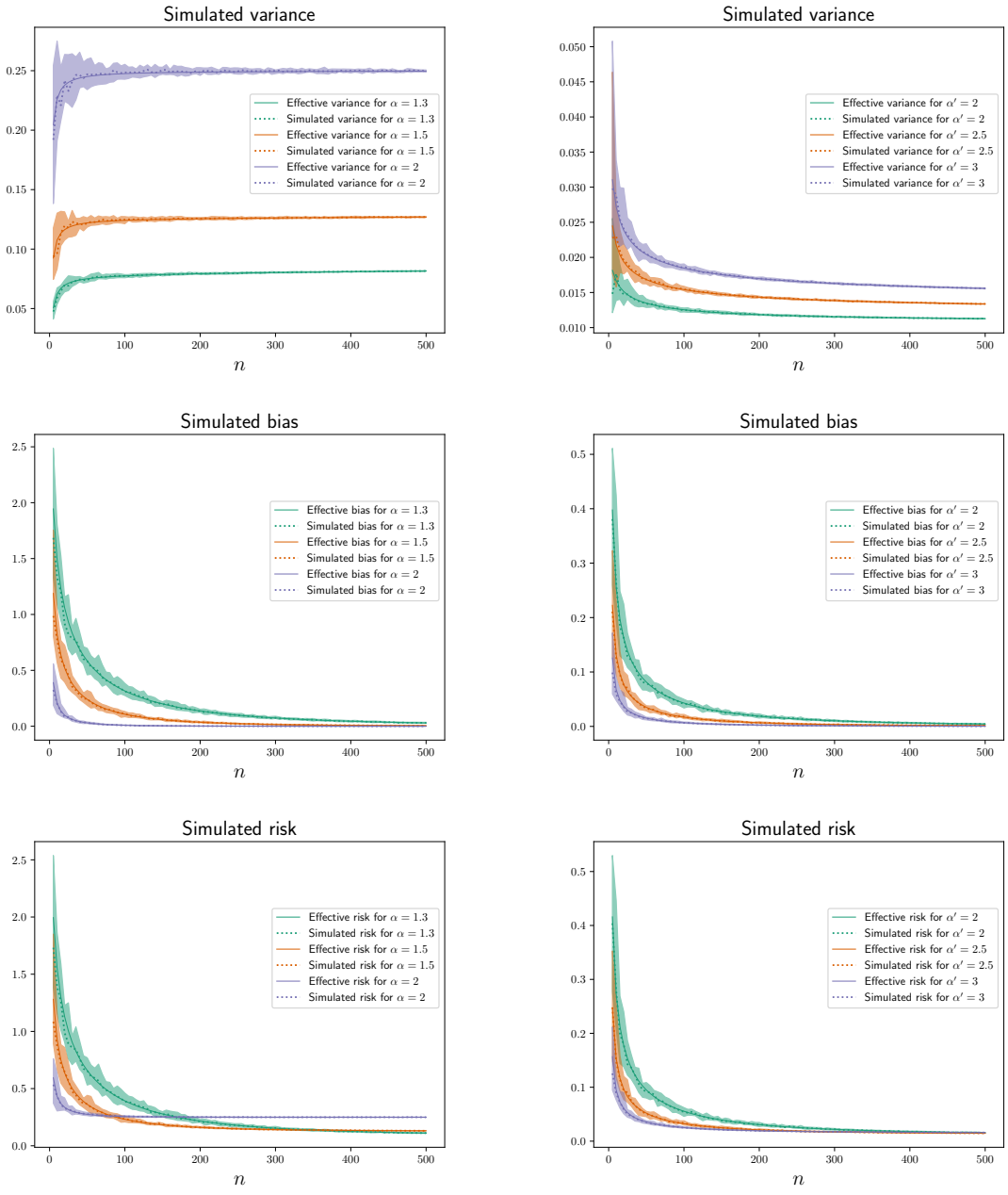
Model (I): $\sigma_i = i^{-\alpha}$. Model (II): $\sigma_i = i^{-1}(1 + \log i)^{-\alpha'}$.

Figure 1: Effective variance, bias, and risk of minimum norm interpolation (a.k.a. ridgeless regression) for two covariance structures defined as models (I) and (II) (power law decay of the eigenvalues with exponents $\alpha > 1$ and $\alpha = 1$), as a function of the sample size n. In model (I) we let the noise level to be $\tau = 0.5$, and in (II) we take $\tau = 0.2$.



Model (I): $\sigma_i = i^{-\alpha}$. Model (II): $\sigma_i = i^{-1}(1+\log i)^{-\alpha'}$.

Figure 2: Effective variance, bias, and risk of minimum norm interpolation for two covariance structures defined as models (I) and (II). Here we fix $n = 500$ and vary the regularization parameter. In model (I) we let the noise size to be $\tau = 0.5$, and in (II) we take $\tau = 0.2$.



Model (I): $\sigma_i = i^{-\alpha}$. Model (II): $\sigma_i = i^{-1}(1 + \log i)^{-\alpha'}$.

Figure 3: Simulation results for variance, bias, and risk of minimum norm interpolation for two covariance structures defined as models (I) and (II). In model (I) we let the noise size to be $\tau = 0.5$, and in (II) we take $\tau = 0.2$. For each n, we run 20 independent trials and take the median for the dotted lines. We also show the shaded areas between 10% and 90% quantiles.

when $k = 0$. An immediate consequence is that $\mathscr{R}_k, \mathscr{F}_k \in \mathcal{F}_k$. Define $\mu_\star := \mu_\star(\lambda, \mu)$ as the unique solution on of the following equation on (μ, ∞)

$$\mu_\star = \mu + \frac{n}{1 + \mathscr{R}_0(\lambda, \mu_\star; \boldsymbol{I})}\,. \tag{36}$$

For $\mu = 0$, this equation reduces to Eq. (5), via the change of variables $\mu_\star = \lambda/\lambda_\star$. For $\mu > 0$ existence and uniqueness follows by a similar argument to the case $\mu = 0$. Indeed, setting $\xi := (\mu_\star - \mu)^{-1}$, the equation is equivalent to $n\xi = 1 + \mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{A} + \xi\boldsymbol{\Sigma}))$, where $\boldsymbol{A} := \lambda\boldsymbol{I} + \mu\boldsymbol{\Sigma}$. Existence and uniqueness follow since the left-hand side is monotone increasing and the right-hand side monotone decreasing in ξ.

In order to quantify the approximation errors $|\mathscr{V}_{\boldsymbol{X}} - \mathsf{V}_n|$ and $|\mathscr{B}_{\boldsymbol{X}} - \mathsf{B}_n|$, we will apply the following lemma (Lemma 6.1), which expresses the bias and variance $\mathscr{B}_{\boldsymbol{X}}$, $\mathscr{V}_{\boldsymbol{X}}$ in terms of derivatives of $\mathscr{F}_n$ and $\mathscr{F}_0$ w.r.t. λ and μ.

Lemma 6.1. *For any $\lambda > 0, \mu \geq 0$, the quantity $\mu_\star > \mu$ is uniquely determined and we have*

$$\mathscr{V}_{\boldsymbol{X}}(\lambda) = \tau^2 \cdot \frac{\partial}{\partial\lambda}\mathscr{F}_n(\lambda, \mu; \boldsymbol{I})\Big|_{\mu=0}\,, \qquad \mathscr{B}_{\boldsymbol{X}}(\lambda) = -\lambda \cdot \frac{\partial}{\partial\mu}\mathscr{F}_n(\lambda, \mu; \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}})\Big|_{\mu=0}\,;$$
$$\mathsf{V}_n(\lambda) = \tau^2 \cdot \frac{\partial}{\partial\lambda}\mathscr{F}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{I})\Big|_{\mu=0}\,, \qquad \mathsf{B}_n(\lambda) = -\lambda \cdot \frac{\partial}{\partial\mu}\mathscr{F}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}})\Big|_{\mu=0}\,,$$

and $\mu_\star(\lambda, 0) = \lambda/\lambda_\star$.

Our proof strategy proceeds in four parts: (I) We show that —due to the regularity properties of $\mathscr{F}_0$ and $\mathscr{F}_n$— a bound on $|\mathscr{F}_0 - \mathscr{F}_n|$ implies a bound on the difference of their derivatives, and hence (via Lemma 6.1) on the error in approximating bias and variance; (II) We prove a bound on $|\mathscr{F}_0 - \mathscr{F}_n|$ interpolating between $\mathscr{F}_0$ and $\mathscr{F}_n$ by adding one row at the time to $\boldsymbol{X}$; (III), (IV) We apply these general bounds respectively to controlling variance and bias.

Recall that we defined $\boldsymbol{\theta} := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}$, and assumed $\|\boldsymbol{\theta}\| < \infty$. By homogeneity, we can and will assume $\|\boldsymbol{\theta}\| = 1$ throughout the proof.

Part I: Reduction to function values approximation. The following lemma reduces controlling the difference of derivatives of $\mathscr{F}_0$ and $\mathscr{F}_n$ to the less arduous task of bounding the difference in function values. Its proof is presented in Appendix B.2.

Lemma 6.2. *For any fixed $k \in \mathbb{N}, \delta \in \mathbb{R}_{\geq 0}$ and a $(k+1)$-times continuously differentiable function $f(t)$ on $[0, k\delta]$, we have*

$$\big|f'(0)\big| \leq \mathcal{O}_k\left(\frac{\max_{0\leq j\leq k}|f(j\delta)|}{\delta} + \sup_{t\in[0,k\delta]}|f^{(k+1)}(t)| \cdot \delta^k\right).$$

With the help of Lemmas 6.1 and 6.2, for any $\delta \in \mathbb{R}_{\geq 0}$, we can upper bound the variance and bias approximations by

$$|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| = \mathcal{O}_k\Bigg(\tau^2 \cdot \max_{0\leq j\leq k}\frac{|\mathscr{F}_n(\lambda + j\delta, 0; \boldsymbol{I}) - \mathscr{F}_0(\lambda + j\delta, \mu_\star(\lambda + j\delta, 0); \boldsymbol{I})|}{\delta}$$
$$+ \tau^2\delta^k \cdot \sup_{\lambda'\in[\lambda,\lambda+k\delta]}\left|\frac{\partial^{k+1}}{\partial\lambda'^{k+1}}\mathscr{F}_n(\lambda', 0; \boldsymbol{I}) - \frac{\partial^{k+1}}{\partial\lambda'^{k+1}}\mathscr{F}_0(\lambda', \mu_\star(\lambda', 0); \boldsymbol{I})\right|\Bigg), \tag{37}$$

and

$$|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| = \mathcal{O}_k\Bigg(\lambda \cdot \max_{0\le j\le k} \frac{\left|\mathscr{F}_n(\lambda, j\delta; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}) - \mathscr{F}_0(\lambda, \mu_\star(\lambda, j\delta); \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\right|}{\delta} + \lambda\delta^k \cdot \sup_{\mu'\in[0,k\delta]} \left|\frac{\partial^{k+1}}{\partial {\mu'}^{k+1}}\mathscr{F}_n(\lambda, \mu'; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}) - \frac{\partial^{k+1}}{\partial {\mu'}^{k+1}}\mathscr{F}_0(\lambda, \mu_\star(\lambda, \mu'); \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\right|\Bigg). \tag{38}$$

Before passing to bounding errors in function values, we provide upper bounds for higher order derivatives in Eqs. (37) and (38). Bounding the derivatives of $\mathscr{F}_n$ is easier as we can easily write an explicit formula for the k-th derivative for any k. (The proof of this lemma is presented in Appendix B.3).

Lemma 6.3. *For any fixed $k \in \mathbb{N}$, we have for all $\lambda > 0$ and $\mu \ge 0$,*

$$\left|\frac{\partial^k}{\partial\lambda^k}\mathscr{F}_n(\lambda, 0; \boldsymbol{I})\right| = \mathcal{O}_k\left(\frac{\mathscr{F}_n(\lambda, 0; \boldsymbol{I})}{\lambda^k}\right), \qquad \left|\frac{\partial^k}{\partial\mu^k}\mathscr{F}_n(\lambda, \mu; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\right| = \mathcal{O}_k\left(\frac{\mathscr{F}_n(\lambda, \mu; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\lambda^k}\right).$$

Computing higher order derivatives of $\mathscr{F}_0$ is less straightforward because $\mathscr{F}_0$ depends on $\mu_\star$ which itself depends implicitly depending on (λ, μ). We postpone this proof to Appendix B.4.

Lemma 6.4. *Let Eq. (21) hold. Then, for any fixed $k \in \mathbb{N}$, we have for all $\lambda > 0$,*

$$\left|\frac{\partial^k}{\partial\lambda^k}\mathscr{F}_0(\lambda, \mu_\star(\lambda, 0); \boldsymbol{I})\right| = \mathcal{O}_k\left(\frac{\mathscr{F}_0(\lambda, \mu_\star(\lambda, 0); \boldsymbol{I})}{\lambda^k\kappa^{2k}}\right),$$

and for all μ such that $0 \le \mu \le \mu_\star(\lambda, \mu)/2$,

$$\left|\frac{\partial^k}{\partial\mu^k}\mathscr{F}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\right| = \mathcal{O}_k\left(\frac{\mathscr{F}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\mu_\star(\lambda, \mu)^k\kappa^{2k}}\right).$$

Part II: Bounding errors in function values. We next proceed to bounding $|\mathscr{F}_n(\lambda, \mu; \boldsymbol{Q}) - \mathscr{F}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q})|$ for a p.s.d. matrix $\boldsymbol{Q}$, which appears in Eqs. (37) and (38). Recall that $\mathscr{F}_i(\lambda, \mu; \boldsymbol{Q}) = \lambda\mathscr{R}_i(\lambda, \mu; \boldsymbol{Q})$.

The next theorem bounds $|\mathscr{R}_n(\lambda, \mu; \boldsymbol{Q}) - \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q})|$ and is the most importan technical step in the proof of our main theorems. Its proof is outlined in Section 7, with several technical lemmas deferred to the appendices

Theorem 5. *Introduce the shorthand $\mathsf{R}_0(\boldsymbol{Q}) := \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q})$. Under Assumption 1, for any $\lambda > 0, \mu \ge 0$, p.s.d. matrix $\boldsymbol{Q}$ with $\|\boldsymbol{Q}\| = 1$ and positive integer D, there exists constants $\eta = \eta(\mathsf{C}_{\boldsymbol{x}}) \in (0, 1/2)$, $\mathsf{C}_\alpha = \mathsf{C}_\alpha(\mathsf{C}_{\boldsymbol{x}}, D) > 0$, $\mathsf{C}_\beta = \mathsf{C}_\beta(\mathsf{C}_{\boldsymbol{x}}, D) > 0$ and $\mathsf{C}_\gamma = \mathsf{C}_\gamma(\mathsf{C}_{\boldsymbol{x}}, D)$ such that for*

$$\gamma := \min\left\{\frac{2}{n}\left(1 + \frac{\mathsf{C}_\gamma \mathsf{d}_{\boldsymbol{\Sigma}}\sigma_{\lfloor\eta n\rfloor}\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)}{\lambda}\right) + \frac{2}{\mu_\star(\lambda,\mu)}, \frac{1}{\lambda}\right\},$$

$$\alpha_1 := \mathsf{C}_\alpha\log n\cdot\sqrt{\gamma\mathsf{R}_0(\boldsymbol{I})},$$

$$\alpha_2 := \mathsf{C}_\alpha\log n\cdot\sqrt{\gamma^3\mathsf{R}_0(\boldsymbol{Q})},$$

$$\beta_1 := \mathsf{C}_\beta\left(\sqrt{n\log n}\cdot\frac{\alpha_1\gamma\mathsf{R}_0(\boldsymbol{Q}) + \alpha_2(1+\mathsf{R}_0(\boldsymbol{I}))}{1+\mathsf{R}_0(\boldsymbol{I})^2} + n\cdot\left\{\frac{\gamma^2\mathsf{R}_0(\boldsymbol{Q}) + \alpha_1\alpha_2}{1+\mathsf{R}_0(\boldsymbol{I})^2} + \frac{\alpha_1^2\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^3}\right\} + \frac{\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})}\right),$$

$$\beta_2 := \frac{\mathsf{C}_\beta n \beta_1}{1 + \mathsf{R}_0(\boldsymbol{I})^2},$$

if $\alpha_1 \le \mathsf{R}_0(\boldsymbol{I})/8$, $\beta_1 \le \mathsf{R}_0(\boldsymbol{Q})/64$, $\gamma\beta_2(1+\mathsf{R}_0(\boldsymbol{I})) \le 1/64$ *and* $n^{-D} = \mathcal{O}(\alpha_1/(1+\mathsf{R}_0(\boldsymbol{I})))$, *for all* $n = \Omega_D(1)$ *with probability* $1 - \mathcal{O}(n^{-D+1})$ *we have*

$$|\mathscr{R}_n(\lambda,\mu;\boldsymbol{Q}) - \mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{Q})| = \mathcal{O}\left(\gamma\beta_2\left(1+\mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{I})\right)\mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{Q}) + \beta_1\right).$$

Let us emphasize that this theorem holds under weaker assumptions than Theorem 1, but the error bounds it provides are quite implicit. We can obtain more explicit bounds by imposing the assumptions of Theorem 1. We first define the generalized version of $\rho(\lambda)$ in Eq. (23) for any p.s.d. matrix $\boldsymbol{Q}$ as

$$\rho(\lambda) := \frac{\mathscr{R}_0(\lambda_\star, 1; \boldsymbol{Q}/\|\boldsymbol{Q}\|)}{\mathscr{R}_0(\lambda_\star, 1; \boldsymbol{I})} \in (0,1]. \tag{39}$$

The proof of this corollary is given in Appendix C.6.

Corollary 6.5. *Under Assumption 1, for any positive integers* k, D *and p.s.d. matrix* $\boldsymbol{Q}$ *with* $\|\boldsymbol{Q}\| = 1$, *there exist constants* $\eta = \eta(\mathsf{C}_{\boldsymbol{x}}) \in (0,1/2)$ *and* $\mathsf{C} = \mathsf{C}(\mathsf{C}_{\boldsymbol{x}}, D) > 0$, *such that the following hold. Define* $\chi_n(\lambda), \kappa, \rho(\lambda)$ *as per Eqs. (20), (21), (39) (those quantities are defined for* $\mu = 0$*). If it holds that* $\mu_\star(\lambda,\mu) \le (1-\kappa/2)^{-1}\mu_\star(\lambda,0)$, *and*

$$\chi_n(\lambda)^3 \log^2 n \le \mathsf{C} n \kappa^{4.5}\sqrt{\rho(\lambda)}, \qquad n^{-2D+1} = \mathcal{O}\left(\sqrt{\frac{\kappa^3 \log^2 n}{\max\{n,\lambda\}}}\right),$$

we then have for all $n = \Omega_D(1)$ *with probability* $1 - \mathcal{O}(n^{-D+1})$ *that*

$$|\mathscr{R}_n(\lambda,\mu;\boldsymbol{Q}) - \mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{Q})| \le \mathcal{E}_n \cdot \mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{Q}),$$

where

$$\mathcal{E}_n = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\chi_n(\lambda)^3\log^2 n}{n\kappa^{6.5}} \cdot \sqrt{\frac{\mathsf{R}_0(\boldsymbol{I})}{\mathsf{R}_0(\boldsymbol{Q})}}\right). \tag{40}$$

To further simplify the assumption $\mu_\star(\lambda,\mu) \le (1-\kappa/2)^{-1}\mu_\star(\lambda,0)$ in Corollary 6.5, the next lemma will be helpful. We defer its proof to Appendix B.5.

Lemma 6.6. *For any fixed* $\lambda > 0$, *the function* $\mu_\star(\lambda,\mu)$ *is increasing in* μ *for all* $\mu \ge 0$. *Assuming Eq. (21), if* $0 \le \mu \le n\kappa^3/2$, *then*

$$\mu_\star(\lambda,\mu) \le (1-\kappa/2)^{-1}\mu_\star(\lambda,0).$$

Part III: Approximation error for variance. We are now ready to combine our results in Part I and Part II to obtain approximation errors $|\mathscr{V}_{\boldsymbol{X}} - \mathsf{V}_n|$ and $|\mathscr{B}_{\boldsymbol{X}} - \mathsf{B}_n|$. For the variance, we want to take δ in Eq. (37) such that $k\delta \le \lambda$. In this case, $[\lambda, \lambda + k\delta] \subset [\lambda, 2\lambda]$. Note that $\lambda_\star(\lambda)$ is an increasing function of λ. Further, by

$$n - \frac{\lambda}{\lambda_\star} = \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-1}\right),$$

we know $\lambda \mapsto \mu_\star(\lambda, 0) = \lambda/\lambda_\star$ is an increasing function. Therefore $\lambda/\lambda_\star(\lambda) \le 2\lambda/\lambda_\star(2\lambda)$, which implies $\lambda_\star(\lambda) \le \lambda_\star(2\lambda) \le 2\lambda_\star(\lambda)$. For λ that satisfies Eq. (21), this guarantees that for any $\lambda' \in [\lambda, 2\lambda]$,

$$\frac{\lambda'}{n\lambda_\star(\lambda')} \ge \frac{\lambda}{n\lambda_\star(\lambda)} \ge \kappa\,,$$

and

$$\begin{aligned}1 - \frac{\lambda'}{n\lambda_\star(\lambda')} &= \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star(\lambda')\boldsymbol{I})^{-1}\right) \ge \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + 2\lambda_\star(\lambda)\boldsymbol{I})^{-1}\right) \ge \frac{1}{2}\cdot\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star(\lambda)\boldsymbol{I})^{-1}\right) \\ &= \frac{1}{2}\left(1 - \frac{\lambda}{n\lambda_\star(\lambda)}\right) \ge \kappa/2\,.\end{aligned}$$

Hence, for any $\lambda' \in [\lambda, 2\lambda]$, Eq. (21) still holds but with constant $\kappa' \ge \kappa/2$. Therefore, we can apply Corollary 6.5 for any $\lambda' \in [\lambda, 2\lambda]$ for $\boldsymbol{Q} = \boldsymbol{I}$ and $\mu = 0$, provided the following conditions hold

$$\chi_n(\lambda')^3 \log^2 n \le \mathsf{C} n(\kappa/2)^{4.5}\sqrt{\rho(\lambda')} = \mathsf{C} n(\kappa/2)^{4.5}\,,$$

where the last equality used the fact that $\rho(\lambda') = 1$ when $\boldsymbol{Q} = \boldsymbol{I}$. Finally, setting $\mathsf{C}' := 2^{-4.5}\mathsf{C}$ and using the fact that $\chi_n(\lambda')$ is decreasing in λ, it suffices to require

$$\chi_n(\lambda)^3 \log^2 n \le \mathsf{C}' n\kappa^{4.5}\,,$$

which holds by the theorem's assumptions.

Hence, we can now apply Corollary 6.5 with $\boldsymbol{Q} = \boldsymbol{I}$, and it follows that with probability $1 - \mathcal{O}_k(n^{-D+1})$,

$$\begin{aligned}\max_{0\le j\le k}\frac{|\mathscr{F}_n(\lambda + j\delta, 0; \boldsymbol{I}) - \mathscr{F}_0(\lambda + j\delta, \mu_\star(\lambda + j\delta, 0); \boldsymbol{I})|}{\delta} &\le \frac{2\lambda\mathcal{E}_n}{\delta}\cdot\max_{0\le j\le k}\mathscr{R}_0(\lambda + j\delta, \mu_\star(\lambda + j\delta, 0); \boldsymbol{I}) \\ &\le \frac{2\mathcal{E}_n}{\delta}\mathscr{F}_0(\lambda, \mu_\star(\lambda, 0); \boldsymbol{I})\,,\end{aligned}$$

where in the last inequality we use that $\mathscr{R}_0(\lambda, \mu_\star(\lambda, 0); \boldsymbol{I}) = n/\mu_\star(\lambda, 0) - 1$ is a decreasing function in λ as $\mu_\star(\lambda, 0)$ is increasing in λ. Next by Lemmas 6.3 and 6.4 we obtain

$$\begin{aligned}&\sup_{\lambda'\in[\lambda,\lambda+k\delta]}\left|\frac{\partial^{k+1}}{\partial\lambda'^{k+1}}\mathscr{F}_n(\lambda', 0; \boldsymbol{I}) - \frac{\partial^{k+1}}{\partial\lambda'^{k+1}}\mathscr{F}_0(\lambda', \mu_\star(\lambda', 0); \boldsymbol{I})\right| \\ &= \mathcal{O}_k\left(\sup_{\lambda'\in[\lambda,2\lambda]}\frac{\mathscr{F}_n(\lambda', 0; \boldsymbol{I}) + \mathscr{F}_0(\lambda', \mu_\star(\lambda', 0); \boldsymbol{I})}{\lambda'^{k+1}\kappa^{2k+2}}\right) \\ &= \mathcal{O}_k\left(\sup_{\lambda'\in[\lambda,2\lambda]}\frac{\lambda'\mathscr{R}_n(\lambda', 0; \boldsymbol{I}) + \lambda'\mathscr{R}_0(\lambda', \mu_\star(\lambda', 0); \boldsymbol{I})}{\lambda^{k+1}\kappa^{2k+2}}\right) \\ &\overset{\text{(i)}}{=} \mathcal{O}_k\left(\frac{\lambda\mathscr{R}_n(\lambda, 0; \boldsymbol{I}) + \lambda\mathscr{R}_0(\lambda, \mu_\star(\lambda, 0); \boldsymbol{I})}{\lambda^{k+1}\kappa^{2k+2}}\right) \\ &\overset{\text{(ii)}}{=} \mathcal{O}_k\left(\frac{(1 + \mathcal{E}_n)\mathscr{F}_0(\lambda, \mu_\star(\lambda, 0); \boldsymbol{I})}{\lambda^{k+1}\kappa^{2k+2}}\right)\,,\end{aligned}$$

where in (i) we use again that $\mathscr{R}_n(\lambda, \mu; \boldsymbol{I})$ and $\mathscr{R}_0(\lambda, \mu_\star(\lambda, 0); \boldsymbol{I})$ are decreasing in λ and in (ii) we apply Corollary 6.5. Substituting the above displays into Eq. (37), we have

$$|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| = \mathcal{O}_k\left(\left(\frac{\mathcal{E}_n}{\delta} + \frac{(1 + \mathcal{E}_n)\delta^k}{\lambda^{k+1}\kappa^{2k+2}}\right)\cdot\tau^2\mathscr{F}_0(\lambda, \mu_\star(\lambda, 0); \boldsymbol{I})\right)\,. \tag{41}$$

Finally, we use the fact that

$$\begin{aligned}\mathscr{F}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{I}) &= \lambda\mathrm{Tr}\left(\boldsymbol{\Sigma}\left(\frac{\lambda}{\lambda_\star}\boldsymbol{\Sigma}+\lambda\boldsymbol{I}\right)^{-1}\right) = \lambda_\star\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)\\ &= \lambda\cdot\frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)} \le \lambda\cdot\frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)}\\ &\le \lambda\cdot\frac{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)}\cdot\frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}\\ &= \lambda\cdot\frac{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)}\cdot\frac{\partial}{\partial\lambda}\mathscr{F}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{I})\Big|_{\mu=0}\,,\end{aligned}$$

and by Eq. (21),

$$\frac{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)} \le \frac{n}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)} = \frac{n\lambda_\star}{\lambda} \le \kappa^{-1}\,.$$

We therefore have, by Lemma 6.1, $\tau^2\mathscr{F}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{I}) \le \lambda\kappa^{-1}\mathsf{V}_n(\lambda)$. Substituting in Eq. (41), we obtain

$$|\mathscr{V}_{\boldsymbol{X}}(\lambda)-\mathsf{V}_n(\lambda)| = \mathcal{O}_k\left(\frac{\lambda\mathcal{E}_n}{\delta\kappa}\cdot+\frac{(1+\mathcal{E}_n)\delta^k}{\lambda^k\kappa^{2k+3}}\right)\cdot\mathsf{V}_n(\lambda)\,.$$

By setting $\delta = \lambda\kappa^2 n^{-1/k}$, the condition $\delta k \le \lambda$ is satisfied for all $n = \Omega_k(1)$, which completes the proof for variance approximation with

$$|\mathscr{V}_{\boldsymbol{X}}(\lambda)-\mathsf{V}_n(\lambda)| = \mathcal{O}_k\left(\mathcal{E}_n\cdot n^{-\frac{1}{k}}\kappa^{-3}+n^{-1}\kappa^{-3}\right)\cdot\mathsf{V}_n(\lambda) = \mathcal{O}_{k,\mathsf{C}_{\mathfrak{a}},D}\left(\frac{\chi_n(\lambda)^3\log^2 n}{n^{1-\frac{1}{k}}\kappa^{9.5}}\right)\cdot\mathsf{V}_n(\lambda)\,,$$

where we use $\chi_n(\lambda)\ge 1$ in the final bound.

Part IV: Approximation error for bias. Note that all the terms on the right-hand side of Eq. (38) are evaluated at the same value of λ. Hence, Eq. (21) applies to each of these terms. We claim that the assumptions of Corollary 6.5 apply to all of these terms, provided the following conditions hold

$$\mu_\star(\lambda,k\delta) \le (1-\kappa/2)^{-1}\mu_\star(\lambda,0)\,, \tag{42}$$

$$\chi_n(\lambda)^3\log^2 n \le \mathsf{C}n\kappa^{4.5}\sqrt{\rho(\lambda)}\,. \tag{43}$$

Indeed, condition (42) implies $\mu_\star(\lambda,\mu)\le(1-\kappa/2)^{-1}\mu_\star(\lambda,0)$ for all $\mu\in[0,k\delta]$ since $\mu\mapsto\mu_\star(\lambda,\mu)$ is monotone decreasing; finally, condition (43) is independent of μ.

then we can apply Lemmas 6.3 and 6.4 and invoke Corollary 6.5 with $\boldsymbol{Q}=\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}$. To be specific, by Corollary 6.5, we have with probability $1-\mathcal{O}_k(n^{-D+1})$,

$$\begin{aligned}\max_{0\le j\le k}\frac{\left|\mathscr{F}_n(\lambda,j\delta;\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})-\mathscr{F}_0(\lambda,\mu_\star(\lambda,j\delta);\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\right|}{\delta} &\le \frac{\mathcal{E}_n}{\delta}\max_{j\in\{0,\dots,k\}}\mathscr{F}_0(\lambda,\mu_\star(\lambda,j\delta);\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\\ &\le \frac{\mathcal{E}_n}{\delta}\mathscr{F}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\,,\end{aligned}$$

as $\mathscr{F}_0(\lambda, \mu_\star(\lambda,\mu); \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})$ decreases with μ. By Lemmas 6.3 and 6.4 we obtain

$$\begin{aligned}
&\sup_{\mu'\in[0,k\delta]} \left| \frac{\partial^{k+1}}{\partial {\mu'}^{k+1}} \mathscr{F}_n(\lambda,\mu';\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}) - \frac{\partial^{k+1}}{\partial {\mu'}^{k+1}} \mathscr{F}_0(\lambda,\mu_\star(\lambda,\mu');\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}) \right| \\
&= \mathcal{O}_k\left(\sup_{\mu'\in[0,k\delta]} \frac{\mathscr{F}_n(\lambda,\mu';\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\lambda^{k+1}} + \sup_{\mu'\in[0,k\delta]} \frac{\mathscr{F}_0(\lambda,\mu_\star(\lambda,\mu');\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\mu_\star(\lambda,\mu')^{k+1}\kappa^{2k+2}} \right) \\
&\overset{\text{(i)}}{=} \mathcal{O}_k\left(\frac{\mathscr{F}_n(\lambda,0;\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\lambda^{k+1}} + \frac{\mathscr{F}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\mu_\star(\lambda,0)^{k+1}\kappa^{2k+2}} \right) \\
&\overset{\text{(ii)}}{=} \mathcal{O}_k\left(\frac{(1+\mathcal{E}_n+\lambda_\star^{k+1}\kappa^{-2k-2})\mathscr{F}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\lambda^{k+1}} \right),
\end{aligned}$$

where in the bound (i) we use the fact that $\mu_\star(\lambda,\mu)$ is increasing in μ (cf. Lemma 6.6) and $\mathscr{F}_k(\lambda,\mu;\boldsymbol{Q})$ is decreasing in μ when $\mu \geq 0$; in (ii) we use that $\mu_\star(\lambda,0) = \lambda/\lambda_\star$. Combining the calculations above, we have from Eq. (38)

$$|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| = \mathcal{O}_k\left(\left(\frac{\lambda\mathcal{E}_n}{\delta} + \frac{\delta^k\left(1+\mathcal{E}_n+\lambda_\star^{k+1}\kappa^{-2k-2}\right)}{\lambda^k} \right) \cdot \mathscr{F}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}) \right).$$

Then we make use of the following bound

$$\begin{aligned}
\mathscr{F}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}) &= \lambda\mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac12}\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{\frac12}\left(\frac{\lambda}{\lambda_\star}\boldsymbol{\Sigma}+\lambda\boldsymbol{I}\right)^{-1} \right) = \lambda_\star\mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac12}\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{\frac12}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1} \right) \\
&= \lambda\cdot\frac{\boldsymbol{\theta}^\mathsf{T}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\boldsymbol{\Sigma}\boldsymbol{\theta}}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)} \leq \frac{\lambda}{n\lambda_\star}\cdot\frac{\lambda_\star^2\boldsymbol{\theta}^\mathsf{T}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\boldsymbol{\Sigma}^2\boldsymbol{\theta}}{1-n^{-1}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)} \\
&= \frac{\lambda}{n\lambda_\star}\cdot\frac{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)}\cdot\frac{\lambda_\star^2\boldsymbol{\theta}^\mathsf{T}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\boldsymbol{\Sigma}^2\boldsymbol{\theta}}{1-n^{-1}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)} \\
&= \frac{\lambda}{n\lambda_\star}\cdot\frac{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)}\cdot\mathsf{B}_n(\lambda),
\end{aligned}$$

where in the last line we used the definition of $\mathsf{B}_n(\lambda)$ in Eq. (7). By Eq. (21), we have

$$\frac{\lambda}{n\lambda_\star}\cdot\frac{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)} = 1-\frac1n\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right) \leq 1,$$

which reduces the approximation bound for bias to

$$|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| = \mathcal{O}_k\left(\frac{\lambda\mathcal{E}_n}{\delta} + \frac{\delta^k\left(1+\mathcal{E}_n+\lambda_\star^{k+1}\kappa^{-2k-2}\right)}{\lambda^k} \right)\cdot\mathsf{B}_n(\lambda).$$

We again take $\delta = \lambda\kappa^2 n^{-\frac1k}$ and the bound becomes

$$|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| = \mathcal{O}_{k,\mathsf{c}_{\boldsymbol{x}},D}\left(\frac{\lambda_\star^{k+1}}{n\kappa^2} + \frac{\chi_n(\lambda)^3\log^2 n}{n^{1-\frac1k}\kappa^{8.5}}\cdot\sqrt{\frac{\mathsf{R}_0(\boldsymbol{I})}{\mathsf{R}_0(\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}} \right)\cdot\mathsf{B}_n(\lambda).$$

This bounds hold under the conditions (42) to (43), which are implied by the following:

$$\mu_\star(\lambda, k\lambda\kappa^2 n^{-\frac1k}) \leq (1-\kappa/2)^{-1}\mu_\star(\lambda,0).$$

$$\chi_n(\lambda)^3 \log^2 n \le \mathsf{C} n\kappa^{4.5}\sqrt{\rho(\lambda)}\,.$$

For the first condition, we invoke Lemma 6.6 to obtain a sufficient requirement $\lambda k n^{-\frac{1}{k}} \le n\kappa/2$. For the last condition, it suffices to have $\chi_n(\lambda)^3 \log^2 n \le \mathsf{C}' n\kappa^{4.5}\sqrt{\rho(\lambda)}$ for the same C' defined in Part III.

7 Proof of Theorem 5

Part I: The iterative sequence. The proof is based on the following interpolating construction. We will construct a sequence of random variables $\mu_i \in \mathcal{F}_{i-1}$ for $i = 0, 1, \cdots, n+1$ (where, by convention, $\mathcal{F}_{-1} = \mathcal{F}_0$ is the trivial σ-algebra) such that, defining

$$\mathsf{R}_i(\boldsymbol{Q}) := \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I} + \mu_i\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)^{-1}\right) = \mathscr{R}_i(\lambda, \mu_i; \boldsymbol{Q})\,, \tag{44}$$

we obtain that $\mathsf{R}_i(\boldsymbol{Q})$ is approximately a martingale and, as a consequence, $\mathsf{R}_0(\boldsymbol{Q}) \approx \mathsf{R}_n(\boldsymbol{Q})$. We will further have $\mu_0 = \mu_\star(\lambda, \mu)$ and $\mu_n \approx \mu$, asd therefore we obtain the desired claim $\mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q}) \approx \mathscr{R}_n(\lambda, \mu; \boldsymbol{Q})$.

Before formally defining the sequence $\{\mu_0, \cdots, \mu_{n+1}\}$, we introduce some helpful notations. We first define the matrices $\boldsymbol{A}_i, \boldsymbol{B}_i \in \mathcal{F}_i$ for $0 \le i \le n$ as

$$\boldsymbol{A}_i := \boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I} + \mu_i\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\,, \tag{45a}$$

$$\boldsymbol{B}_i := \boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I} + \mu_{i+1}\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\,. \tag{45b}$$

Then we can write $\mathsf{R}_i(\boldsymbol{Q}) = \mathrm{Tr}\,(\boldsymbol{Q}\boldsymbol{A}_i)$. Similarly we define another sequence of functions by $\mathsf{S}_i(\boldsymbol{Q}) := \mathrm{Tr}\,(\boldsymbol{Q}\boldsymbol{B}_i)$.

Now we are ready to define the sequence $\mu_i \in \mathcal{F}_{i-1}$. We set the initial value $\mu_0 = \mu_\star(\lambda, \mu) \in \mathcal{F}_{-1}$ and thus $\mathsf{R}_0(\boldsymbol{Q}) = \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q})$. The sequence $(\mu_j)_{j\ge 1}$ is iteratively determined through the following equation

$$\mu_{i+1} = \mu_i - \frac{1}{1 + \mathsf{S}_i(\boldsymbol{I})} = \mu_i - \frac{1}{1 + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I} + \mu_{i+1}\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right)} \qquad \text{s.t. } \boldsymbol{B}_i \succ 0\,. \tag{46}$$

It is evident that if the solution μ_{i+1} exists and is unique (almost surely with respect to the random choice of μ_i), since $\mu_i \in \mathcal{F}_{i-1}$ and $\boldsymbol{X}_i \in \mathcal{F}_i$, it follows that $\mu_{i+1} \in \mathcal{F}_i$. The next lemma shows that the iteration via (46) is indeed well-defined. Its proof is in Appendix C.1.

Lemma 7.1. *There exists a unique strictly decreasing sequence $\mu_0 > \mu_1 > \mu_2 > \cdots > \mu_n > \mu_{n+1}$ satisfying the update rule (46).*

Part II: Approximation to a martingale. We next explain what is the rationale for the iterative definition of Eq. (46), and how it will help us prove the theorem claim.

Since we want to upper bound $|\mathsf{R}_n(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})|$, it makes sense to compute the difference $\mathsf{R}_i(\boldsymbol{Q}) - \mathsf{R}_{i-1}(\boldsymbol{Q})$,

$$\begin{aligned}\mathsf{R}_i(\boldsymbol{Q}) - \mathsf{R}_{i-1}(\boldsymbol{Q}) &= (\mathsf{R}_i(\boldsymbol{Q}) - \mathsf{S}_{i-1}(\boldsymbol{Q})) + (\mathsf{S}_{i-1}(\boldsymbol{Q}) - \mathsf{R}_{i-1}(\boldsymbol{Q})) \\ &= \underbrace{\mathrm{Tr}\,(\boldsymbol{Q}\,(\boldsymbol{A}_i - \boldsymbol{B}_{i-1}))}_{\text{(I)}} + \underbrace{\mathrm{Tr}\,(\boldsymbol{Q}\,(\boldsymbol{B}_{i-1} - \boldsymbol{A}_{i-1}))}_{\text{(II)}}\,.\end{aligned} \tag{47}$$

Using rgw definitions in Eqs. (45a) and (45b), we can further expand (I) by Sherman-Morrison formula

$$\begin{aligned}
\boldsymbol{A}_i - \boldsymbol{B}_{i-1} &= \boldsymbol{\Sigma}^{\frac{1}{2}} \left\{ \left(\lambda \boldsymbol{I} + \mu_i \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} + \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \right)^{-1} - \left(\lambda \boldsymbol{I} + \mu_i \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} \right\} \boldsymbol{\Sigma}^{\frac{1}{2}} \\
&= - \frac{\boldsymbol{\Sigma}^{\frac{1}{2}} \left(\lambda \boldsymbol{I} + \mu_i \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \left(\lambda \boldsymbol{I} + \mu_i \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}}}{1 + \boldsymbol{x}_i^\mathsf{T} \left(\lambda \boldsymbol{I} + \mu_i \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} \boldsymbol{x}_i} \\
&= - \frac{\boldsymbol{\Sigma}^{\frac{1}{2}} \left(\lambda \boldsymbol{I} + \mu_i \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{z}_i \boldsymbol{z}_i^\mathsf{T} \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\lambda \boldsymbol{I} + \mu_i \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}}}{1 + \boldsymbol{z}_i^\mathsf{T} \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\lambda \boldsymbol{I} + \mu_i \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{z}_i} \\
&= - \frac{\boldsymbol{B}_{i-1} \boldsymbol{z}_i \boldsymbol{z}_i^\mathsf{T} \boldsymbol{B}_{i-1}}{1 + \boldsymbol{z}_i^\mathsf{T} \boldsymbol{B}_{i-1} \boldsymbol{z}_i} ,
\end{aligned}$$

and thus write

$$\text{(I)} = - \frac{\operatorname{Tr}\left(\boldsymbol{Q} \boldsymbol{B}_{i-1} \boldsymbol{z}_i \boldsymbol{z}_i^\mathsf{T} \boldsymbol{B}_{i-1} \right)}{1 + \boldsymbol{z}_i^\mathsf{T} \boldsymbol{B}_{i-1} \boldsymbol{z}_i} .$$

We can also compute (II) by noting that

$$\begin{aligned}
\boldsymbol{B}_{i-1} - \boldsymbol{A}_{i-1} &= \boldsymbol{\Sigma}^{\frac{1}{2}} \left\{ \left(\lambda \boldsymbol{I} + \mu_i \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} - \left(\lambda \boldsymbol{I} + \mu_{i-1} \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} \right\} \boldsymbol{\Sigma}^{\frac{1}{2}} \\
&= \boldsymbol{\Sigma}^{\frac{1}{2}} \left\{ \left(\lambda \boldsymbol{I} + \mu_i \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} \cdot (\mu_{i-1} - \mu_i) \boldsymbol{\Sigma} \cdot \left(\lambda \boldsymbol{I} + \mu_{i-1} \boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^\mathsf{T} \boldsymbol{X}_{i-1} \right)^{-1} \right\} \boldsymbol{\Sigma}^{\frac{1}{2}} ,
\end{aligned}$$

and therefore

$$\text{(II)} = (\mu_{i-1} - \mu_i) \cdot \operatorname{Tr}\left(\boldsymbol{Q} \boldsymbol{B}_{i-1} \boldsymbol{A}_{i-1} \right) .$$

It is now clear what is the motivation for defining μ_{i+1} as per Eq. (46). We hope to have (I)+(II) ≈ 0. Under the approximation $\operatorname{Tr}\left(\boldsymbol{Q} \boldsymbol{B}_{i-1} \boldsymbol{z}_i \boldsymbol{z}_i^\mathsf{T} \boldsymbol{B}_{i-1} \right) \approx \mathbb{E}\left[\operatorname{Tr}\left(\boldsymbol{Q} \boldsymbol{B}_{i-1} \boldsymbol{z}_i \boldsymbol{z}_i^\mathsf{T} \boldsymbol{B}_{i-1} \right) \mid \mathcal{F}_{i-1} \right] = \operatorname{Tr}\left(\boldsymbol{Q} \boldsymbol{B}_{i-1}^2 \right) \approx \operatorname{Tr}\left(\boldsymbol{Q} \boldsymbol{B}_{i-1} \boldsymbol{A}_{i-1} \right)$, this is achieved when

$$\mu_i - \mu_{i-1} = - \frac{1}{1 + \mathbb{E}\left[\boldsymbol{z}_i^\mathsf{T} \boldsymbol{B}_{i-1} \boldsymbol{z}_i \mid \mathcal{F}_{i-1} \right]} = - \frac{1}{1 + \mathsf{S}_{i-1}(\boldsymbol{I})} ,$$

which recovers the iteration in Eq. (46).

Part III: Proof via stopping times. We next make the previous argument rigorous. For any scalars $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma > 0$ (in what follows, we'll use the notation $\Delta := (\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma)$) we consider the events

$$E_i(\boldsymbol{Q}) := \left\{ \left| \boldsymbol{z}_i^\mathsf{T} \boldsymbol{B}_{i-1} \boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I}) \right| \le \alpha_1, \left| \boldsymbol{z}_i^\mathsf{T} \boldsymbol{B}_{i-1} \boldsymbol{Q} \boldsymbol{B}_{i-1} \boldsymbol{z}_i - \operatorname{Tr}\left(\boldsymbol{Q} \boldsymbol{B}_{i-1}^2 \right) \right| \le \alpha_2, \| \boldsymbol{A}_i \| \le \gamma \right\} , \tag{48a}$$

$$F_i(\boldsymbol{Q}) := \left\{ \max\{ |\mathsf{R}_i(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})| , |\mathsf{S}_i(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})| \} \le \beta_1, \left| \mu_{i+1} - \overline{\mu}_{i+1} \right| \le \beta_2, \| \boldsymbol{B}_i \| \le \gamma \right\} , \tag{48b}$$

where $\overline{\mu}_{i+1} = \mu \cdot (i+1)/n + \mu_\star(\lambda, \mu) \cdot (1 - (i+1)/n)$ is nonrandom. In particular we set $E_0(\boldsymbol{Q}) = \Omega$ so that $E_i(\boldsymbol{Q})$ and $F_i(\boldsymbol{Q})$ are well-defined for $0 \le i \le n$. It follows then $E_i(\boldsymbol{Q}), F_i(\boldsymbol{Q}) \in \mathcal{F}_i$. Next we can proceed to define two stopping times via

$$\{ T_E(\boldsymbol{Q}) \ge k+1 \} := \left(\bigcap_{i=0}^{k} E_i(\boldsymbol{Q}) \right) \cap \left(\bigcap_{i=0}^{k-1} F_i(\boldsymbol{Q}) \right) , \tag{49a}$$

$$\{T_F(\boldsymbol{Q}) \geq k+1\} := \left(\bigcap_{i=0}^{k} E_i(\boldsymbol{Q})\right) \cap \left(\bigcap_{i=0}^{k} F_i(\boldsymbol{Q})\right), \tag{49b}$$

for $k = 0, 1, \cdots, n$, with $T_E(\boldsymbol{Q}), T_F(\boldsymbol{Q}) \in \{0, 1, \cdots, n+1\}$. One can easily check that $T_E(\boldsymbol{Q})$ and $T_F(\boldsymbol{Q})$ are indeed stopping times since the sets in the above displays are in $\mathcal{F}_k$, and another immediate consequence is that $T_E(\boldsymbol{Q}) \geq T_F(\boldsymbol{Q})$. These stopping times are helpful since the event $\{T_F(\boldsymbol{Q}) = n+1\}$ implies

$$\max_{0 \leq i \leq n} \left\{ \left| \mathsf{R}_i(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q}) \right|, \left| \mathsf{S}_i(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q}) \right| \right\} \leq \beta_1 ,$$

and thus if β_1 is much smaller than $\mathsf{R}_0(\boldsymbol{Q})$, we can show $\mathsf{R}_n(\boldsymbol{Q}) \approx \mathsf{R}_0(\boldsymbol{Q})$ as desired. Therefore, we want to lower bound the probability for the event $\{T_F(\boldsymbol{Q}) = n+1\}$. We use the shorthand $p_{i,j}(T_1, T_2, \boldsymbol{Q}) := \mathbb{P}(T_1(\boldsymbol{Q}) \geq i, T_2(\boldsymbol{I}) \geq j)$ for $T_1, T_2 \in \{T_E, T_F\}$. By telescoping sum, we have

$$\begin{aligned}
&\mathbb{P}(T_F(\boldsymbol{Q}) \geq 0, T_F(\boldsymbol{I}) \geq 0) - \mathbb{P}(T_F(\boldsymbol{Q}) = n+1, T_F(\boldsymbol{I}) = n+1) \\
&= p_{0,0}(T_F, T_F, \boldsymbol{Q}) - p_{n+1,n+1}(T_F, T_F, \boldsymbol{Q}) \\
&= \sum_{k=0}^{n} (p_{k,k}(T_F, T_F, \boldsymbol{Q}) - p_{k+1,k+1}(T_E, T_E, \boldsymbol{Q})) + \sum_{k=1}^{n+1} (p_{k,k}(T_E, T_E, \boldsymbol{Q}) - p_{k,k}(T_F, T_F, \boldsymbol{Q})) \\
&\leq \sum_{k=0}^{n} (p_{k,k}(T_F, T_F, \boldsymbol{Q}) - p_{k+1,k}(T_E, T_F, \boldsymbol{Q}) + p_{k,k}(T_F, T_F, \boldsymbol{Q}) - p_{k,k+1}(T_F, T_E, \boldsymbol{Q})) \\
&\qquad + \sum_{k=1}^{n+1} (p_{k,k}(T_E, T_E, \boldsymbol{Q}) - p_{k,k}(T_F, T_E, \boldsymbol{Q}) + p_{k,k}(T_E, T_E, \boldsymbol{Q}) - p_{k,k}(T_E, T_F, \boldsymbol{Q})) \\
&\leq \sum_{k=0}^{n} (p_{k,k}(T_F, T_F, \boldsymbol{Q}) - p_{k+1,k}(T_E, T_F, \boldsymbol{Q}) + p_{k,k}(T_F, T_F, \boldsymbol{I}) - p_{k+1,k}(T_E, T_F, \boldsymbol{I})) \\
&\qquad + \sum_{k=1}^{n+1} (p_{k,k}(T_E, T_E, \boldsymbol{Q}) - p_{k,k}(T_F, T_E, \boldsymbol{Q}) + p_{k,k}(T_E, T_E, \boldsymbol{I}) - p_{k,k}(T_F, T_E, \boldsymbol{I})) ,
\end{aligned} \tag{50}$$

where in the last inequality we use $\mathbb{P}(A \cap B) - \mathbb{P}(A \cap B') \leq \mathbb{P}(B) - \mathbb{P}(B')$ for $B' \subset B$.

We are left with the task of bounding the two terms $p_{k,k}(T_F, T_F, \boldsymbol{Q}) - p_{k+1,k}(T_E, T_F, \boldsymbol{Q})$ and $p_{k,k}(T_E, T_E, \boldsymbol{Q}) - p_{k,k}(T_F, T_E, \boldsymbol{Q})$ for any p.s.d. $\boldsymbol{Q}$, and showing that they are small. Before doing this, we show that, by appropriately choosing γ, we have $\|\boldsymbol{A}_i\| \leq \gamma$ and $\|\boldsymbol{B}_i\| \leq \gamma$ with high probability. The proof of the next lemma is in Appendix C.2.

Lemma 7.2. *Under Assumption 1, for any positive integer D, there exists a fixed $\eta = \eta(\mathsf{C}_{\boldsymbol{x}}) \in (0, 1/2)$, such that for all $n = \Omega_D(1)$, it holds with probability $1 - \mathcal{O}(n^{-D})$ that*

$$\left\| \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\lambda \boldsymbol{I} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X} \right)^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}} \right\| \leq \frac{2}{n} \left(1 + \frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\mathsf{d}_{\boldsymbol{\Sigma}} \sigma_{\lfloor \eta n \rfloor} \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n) \right)}{\lambda} \right), \qquad \text{for all } \lambda > 0 ;$$

additionally, under the same notations of Proposition 2.2, letting $\boldsymbol{\theta}_{\leq k} := \sum_{i \leq k} \langle \boldsymbol{\theta}, \boldsymbol{v}_i \rangle \boldsymbol{v}_i$ and $\boldsymbol{\theta}_{>k} := \boldsymbol{\theta} - \boldsymbol{\theta}_{\leq k}$, we have for all $\lambda > 0$,

$$\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\lambda \boldsymbol{I} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X} \right)^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta} \leq \frac{2}{n} \left(1 + \frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\mathsf{d}_{\boldsymbol{\Sigma}} \sigma_{\lfloor \eta n \rfloor} \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n) \right)}{\lambda} \right) \|\boldsymbol{\theta}_{\leq n}\|^2 + \frac{2 \|\boldsymbol{\beta}_{>n}\|^2}{\lambda} .$$

The next lemma—upper bounding the first term (I)—uses Hanson-Wright inequality to show concentration for events $E_i(\boldsymbol{Q})$ in (48a). A proof is in Appendix C.3.

Lemma 7.3. *Under Assumption 1, choose β_1, β_2 in Eq. (48b) so that $\beta_1 \le \mathsf{R}_0(\boldsymbol{Q})/4$ and $\beta_2 \le \mu/2$. Then for any positive integer D, there exists constants $\eta = \eta(\mathsf{C}_{\boldsymbol{x}}) \in (0, 1/2)$, $\mathsf{C}_\alpha = \mathsf{C}_\alpha(\mathsf{C}_{\boldsymbol{x}}, D)$ and $\mathsf{C}_\gamma = \mathsf{C}_\gamma(\mathsf{C}_{\boldsymbol{x}}, D)$ such that if we take*

$$\begin{aligned}
\gamma &= \min\left\{\frac{2}{n}\left(1 + \frac{\mathsf{C}_\gamma \mathsf{d}_{\boldsymbol{\Sigma}} \sigma_{\lfloor \eta n \rfloor} \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n)}{\lambda}\right) + \frac{2}{\mu_\star(\lambda,\mu)}, \frac{1}{\lambda}\right\}, \\
\alpha_1 &= \mathsf{C}_\alpha \log n \cdot \sqrt{\gamma \mathsf{R}_0(\boldsymbol{I})}, \\
\alpha_2 &= \mathsf{C}_\alpha \log n \cdot \sqrt{\gamma^3 \mathsf{R}_0(\boldsymbol{Q})},
\end{aligned}$$

it holds for all $n = \Omega_D(1)$ that

$$p_{k,k}(T_F, T_F, \boldsymbol{Q}) - p_{k+1,k}(T_E, T_F, \boldsymbol{Q}) = \mathcal{O}(n^{-D}).$$

In addition, on the event $\{T_F(\boldsymbol{Q}) \ge k, T_F(\boldsymbol{I}) \ge k\} \in \mathcal{F}_{k-1}$ we have (using the shorthand $\mathbb{E}_{k-1}\{\cdot\} := \mathbb{E}\{\cdot|\mathcal{F}_{k-1}\}$)

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[\left|\boldsymbol{z}_k^\mathsf{T} \boldsymbol{B}_{k-1} \boldsymbol{z}_k - \mathsf{S}_{k-1}(\boldsymbol{I})\right| \mathbb{I}\left\{\left|\boldsymbol{z}_k^\mathsf{T} \boldsymbol{B}_{k-1} \boldsymbol{z}_k - \mathsf{S}_{k-1}(\boldsymbol{I})\right| \ge \alpha_1\right\}\right] \\
&\quad = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}}\left(n^{-D} \cdot \sqrt{\gamma \mathsf{R}_0(\boldsymbol{I})}\right) = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D}\left(n^{-D} \cdot \alpha_1\right), \\
&\mathbb{E}_{k-1}\left[\left|\boldsymbol{z}_k^\mathsf{T} \boldsymbol{B}_{k-1} \boldsymbol{Q} \boldsymbol{B}_{k-1} \boldsymbol{z}_k - \mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{k-1}^2\right)\right| \mathbb{I}\left\{\left|\boldsymbol{z}_k^\mathsf{T} \boldsymbol{B}_{k-1} \boldsymbol{Q} \boldsymbol{B}_{k-1} \boldsymbol{z}_k - \mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{k-1}^2\right)\right| \ge \alpha_2\right\}\right] \\
&\quad = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}}\left(n^{-D} \cdot \sqrt{\gamma^3 \mathsf{R}_0(\boldsymbol{Q})}\right) = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D}\left(n^{-D} \cdot \alpha_2\right).
\end{aligned}$$

We then proceed to bound the term $p_{k,k}(T_E, T_E, \boldsymbol{Q}) - p_{k,k}(T_F, T_E, \boldsymbol{Q})$. The proof of the next lemma is in Appendix C.4.

Lemma 7.4. *Under Assumption 1, for any positive integer D, there exists a constant $\mathsf{C}_\beta = \mathsf{C}_\beta(\mathsf{C}_{\boldsymbol{x}}, D) > 0$ such that the following holds. Consider $\alpha_1, \alpha_2, \gamma$ as defined in Lemma 7.3, and set β_1, β_2 by*

$$\begin{aligned}
\beta_1 &= \mathsf{C}_\beta\left(\sqrt{n \log n} \cdot \frac{\alpha_1 \gamma \mathsf{R}_0(\boldsymbol{Q}) + \alpha_2(1 + \mathsf{R}_0(\boldsymbol{I}))}{1 + \mathsf{R}_0(\boldsymbol{I})^2} + n \cdot \left\{\frac{\gamma^2 \mathsf{R}_0(\boldsymbol{Q}) + \alpha_1 \alpha_2}{1 + \mathsf{R}_0(\boldsymbol{I})^2} + \frac{\alpha_1^2 \gamma \mathsf{R}_0(\boldsymbol{Q})}{1 + \mathsf{R}_0(\boldsymbol{I})^3}\right\} + \frac{\gamma \mathsf{R}_0(\boldsymbol{Q})}{1 + \mathsf{R}_0(\boldsymbol{I})}\right), \\
\beta_2 &= \frac{\mathsf{C}_\beta n \beta_1}{1 + \mathsf{R}_0(\boldsymbol{I})^2}.
\end{aligned}$$

If $\alpha_1 \le \mathsf{R}_0(\boldsymbol{I})/4$, $\beta_1 \le \mathsf{R}_0(\boldsymbol{Q})/4$, $\beta_2 \le \mu/2$ and $n^{-D} = \mathcal{O}\left(\alpha_1/\left(1 + \mathsf{R}_0(\boldsymbol{I})\right)\right)$, then for all $1 \le k \le n+1$ and $n = \Omega_D(1)$,

$$p_{k,k}(T_E, T_E, \boldsymbol{Q}) - p_{k,k}(T_F, T_E, \boldsymbol{Q}) = \mathcal{O}(n^{-D}).$$

Applying Lemmas 7.3 and 7.4 to Eq. (50) (note that we can take $\boldsymbol{Q} = \boldsymbol{I}$), we have shown that

$$\begin{aligned}
1 - \mathbb{P}(T_F(\boldsymbol{Q}) = n+1, T_F(\boldsymbol{I}) = n+1) &= \mathbb{P}(T_F(\boldsymbol{Q}) \ge 0, T_F(\boldsymbol{I}) \ge 0) - \mathbb{P}(T_F(\boldsymbol{Q}) = n+1, T_F(\boldsymbol{I}) = n+1) \\
&= \mathcal{O}(n^{-D+1}),
\end{aligned}$$

which implies by choosing the parameter Δ given by the above lemmas, with probability $1 - \mathcal{O}(n^{-D+1})$

$$|\mathsf{R}_n(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})| \le \beta_1, \qquad |\mu_n - \overline{\mu}_n| \le \beta_2.$$

Therefore, since $\overline{\mu}_n = \mu$, $\mu_0 = \mu_\star(\lambda, \mu)$, and recalling the definition of $\mathsf{R}_k(\boldsymbol{Q})$, cf. Eq. (44), we have

$$
\begin{aligned}
|\mathscr{R}_n(\lambda,\mu;\boldsymbol{Q}) - \mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{Q})| &\le |\mathscr{R}_n(\lambda,\mu;\boldsymbol{Q}) - \mathscr{R}_n(\lambda,\mu_n;\boldsymbol{Q})| + |\mathscr{R}_n(\lambda,\mu_n;\boldsymbol{Q}) - \mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{Q})| \\
&= \left|(\mu_n - \mu)\cdot \mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I} + \mu\boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}_n\right)\right| + |\mathsf{R}_n(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})| \\
&\overset{\text{(i)}}{\le} \beta_2 \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I} + \mu\boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| \mathsf{R}_n(\boldsymbol{Q}) + |\mathsf{R}_n(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})| \\
&\overset{\text{(ii)}}{\le} \gamma\beta_2 \mathsf{R}_n(\boldsymbol{Q}) + |\mathsf{R}_n(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})| \le \gamma\beta_2\mathsf{R}_0(\boldsymbol{Q}) + (1+\gamma\beta_2)\,|\mathsf{R}_n(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})| \\
&\le \gamma\beta_2\mathsf{R}_0(\boldsymbol{Q}) + \beta_1(1+\gamma\beta_2) \overset{\text{(ii)}}{\le} \frac{5}{4}\gamma\beta_2\mathsf{R}_0(\boldsymbol{Q}) + \beta_1\,,
\end{aligned}
$$

where in (ii) we used Lemma 7.2; in (iii) we used the fact that $\beta_1 \le \mathsf{R}_0(\boldsymbol{Q})/4$ by assumption. We explain the inequality in (i) more carefully as it is less evident. Denoting by $\boldsymbol{B} = \boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I} + \mu\boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}$, we first show $\boldsymbol{B}$ and $\boldsymbol{A}_n$ commute. Clearly commutativity holds if $\mu = \mu_n$, otherwise we have

$$
\begin{aligned}
\boldsymbol{B}\boldsymbol{A}_n &= \boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I} + \mu\boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}} \cdot \boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I} + \mu_n\boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}} \\
&= (\mu_n - \mu)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\left\{(\lambda \boldsymbol{I} + \mu\boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} - (\lambda \boldsymbol{I} + \mu_n\boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right\}\boldsymbol{\Sigma}^{\frac{1}{2}} \\
&= (\mu_n - \mu)^{-1}\,(\boldsymbol{B} - \boldsymbol{A}_n) = \boldsymbol{A}_n\boldsymbol{B}\,.
\end{aligned}
$$

Noting that $\boldsymbol{B}$ and $\boldsymbol{A}_n$ are both p.s.d. compact self-adjoint operators in Hilbert space, commutativity implies they can be simultaneously orthogonally diagonalized and that $\boldsymbol{B}^{\frac{1}{2}}$ and $\boldsymbol{A}_n^{\frac{1}{2}}$ also commute. Consequently, combined with the fact that $\mathrm{Tr}(\boldsymbol{A}_n\boldsymbol{C}) \le \|\boldsymbol{A}_n\|\mathrm{Tr}(\boldsymbol{C})$ for any p.s.d. matrix $\boldsymbol{C}$, we have (i) from

$$
\begin{aligned}
\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I} + \mu\boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{A}_n\right) &= \mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}\boldsymbol{A}_n\right) = \mathrm{Tr}\left(\boldsymbol{B}\cdot\boldsymbol{A}_n^{\frac{1}{2}}\boldsymbol{Q}\boldsymbol{A}_n^{\frac{1}{2}}\right) \le \|\boldsymbol{B}\|\,\mathrm{Tr}(\boldsymbol{Q}\boldsymbol{A}_n) \\
&= \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I} + \mu\boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| \mathsf{R}_n(\boldsymbol{Q})\,.
\end{aligned}
$$

We therefore proved the following. If $\beta_1 \le \mathsf{R}_0(\boldsymbol{Q})/4$ and $n^{-D} = \mathcal{O}\left(\alpha_1/\left(1+\mathsf{R}_0(\boldsymbol{I})\right)\right)$, then

$$
\beta_2 \le \mu/2 \quad \Rightarrow \quad |\mathscr{R}_n(\lambda,\mu;\boldsymbol{Q}) - \mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{Q})| = \mathcal{O}\left(\gamma\beta_2\mathsf{R}_0(\boldsymbol{Q}) + \beta_1\right)\,. \tag{51}
$$

To remove the condition $\beta_2 \le \mu/2$, we use the following estimate, proven in Appendix C.5.

Lemma 7.5. *Under Assumption 1, consider the parameter tuple $\Delta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma)$ defined in Lemmas 7.3 and 7.4. If $\alpha_1 \le \mathsf{R}_0(\boldsymbol{I})/8$, $\beta_1 \le \mathsf{R}_0(\boldsymbol{Q})/64$, $\gamma\beta_2(1+\mathsf{R}_0(\boldsymbol{I})) \le 1/64$, $n^{-D} = \mathcal{O}(\alpha_1/(1+\mathsf{R}_0(\boldsymbol{I})))$ and $\beta_2 > \mu/2$, then we have*

$$
|\mathscr{R}_n(\lambda,\mu;\boldsymbol{Q}) - \mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{Q})| = \mathcal{O}\left(\gamma\beta_2\left(1+\mathsf{R}_0(\boldsymbol{I})\right)\mathsf{R}_0(\boldsymbol{Q}) + \beta_1\right) \tag{52}
$$

Combining Eqs. (51) and (52), the proof is complete.

Acknowledgements

This work was supported by the NSF through award DMS-2031883, the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning, the NSF grant

CCF-2006489, the ONR grant N00014-18-1-2729, and a grant from Eric and Wendy Schmidt at the Institute for Advanced Studies. C. Cheng is supported by the William R. Hewlett Stanford graduate fellowship.

Part of this work was carried out while A. Montanari was on partial leave from Stanford and a Chief Scientist at Ndata Inc dba Project N. The present research is unrelated to A. Montanari's activity while on leave.

References

[Ada15] Radoslaw Adamczak, *A note on the Hanson-Wright inequality for random vectors with dependencies*, Electronic Communications in Probability **20** (2015), 1–13.

[AEK+14] Bloemendal Alex, László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin, *Isotropic local laws for sample covariance and generalized wigner matrices*, Electronic Journal of Probability **19** (2014), 1–53.

[Apo00] Tom M Apostol, *Calculating higher derivatives of inverses*, The American Mathematical Monthly **107** (2000), no. 8, 738–741.

[ASS20] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky, *High-dimensional dynamics of generalization error in neural networks*, Neural Networks **132** (2020), 428–446.

[BBEKY13] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu, *Optimal m-estimation in high-dimensional regression*, Proceedings of the National Academy of Sciences **110** (2013), no. 36, 14563–14568.

[Bel21] Mikhail Belkin, *Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation*, Acta Numerica **30** (2021), 203–248.

[BGL+14] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al., *Analysis and geometry of markov diffusion operators*, vol. 103, Springer, 2014.

[BKM+19] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová, *Optimal errors and phase transitions in high-dimensional generalized linear models*, Proceedings of the National Academy of Sciences **116** (2019), no. 12, 5451–5460.

[BLLT20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler, *Benign overfitting in linear regression*, Proceedings of the National Academy of Sciences **117** (2020), no. 48, 30063–30070.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford University Press, 2013.

[BM11] Mohsen Bayati and Andrea Montanari, *The lasso risk for gaussian matrices*, IEEE Transactions on Information Theory **58** (2011), no. 4, 1997–2017.

[BMR21] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin, *Deep learning: a statistical viewpoint*, Acta numerica **30** (2021), 87–201.

[Bré11] Haim Brézis, *Functional analysis, sobolev spaces and partial differential equations*, vol. 2, Springer, 2011.

[BY08] Zhi-Dong Bai and Yong-Qua Yin, *Limit of the smallest eigenvalue of a large dimensional sample covariance matrix*, Advances In Statistics, World Scientific, 2008, pp. 108–127.

[CM22] Michael Celentano and Andrea Montanari, *Fundamental barriers to high-dimensional regression with convex penalties*, The Annals of Statistics **50** (2022), no. 1, 170–196.

[CMW20] Michael Celentano, Andrea Montanari, and Yuting Wei, *The lasso with general gaussian designs with applications to hypothesis testing*, arXiv:2007.13716 (2020).

[CT05] Emmanuel J Candes and Terence Tao, *Decoding by linear programming*, IEEE transactions on information theory **51** (2005), no. 12, 4203–4215.

[DET05] David L Donoho, Michael Elad, and Vladimir N Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Transactions on information theory **52** (2005), no. 1, 6–18.

[Dic16] Lee H. Dicker, *Ridge regression and asymptotic minimax estimation over spheres of growing dimension*, Bernoulli **22** (2016), no. 1, 1–37.

[DJM13] David L Donoho, Iain Johnstone, and Andrea Montanari, *Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising*, IEEE transactions on information theory **59** (2013), no. 6, 3396–3433.

[DM16] David Donoho and Andrea Montanari, *High dimensional robust m-estimation: Asymptotic variance via approximate message passing*, Probability Theory and Related Fields **166** (2016), no. 3, 935–969.

[DW18] Edgar Dobriban and Stefan Wager, *High-dimensional asymptotics of prediction: ridge regression and classification*, Annals of Statistics **46** (2018), no. 1, 247–279.

[EK18] Noureddine El Karoui, *On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators*, Probability Theory and Related Fields **170** (2018), no. 1, 95–175.

[EKBB+13] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu, *On robust regression with high-dimensional predictors*, Proceedings of the National Academy of Sciences **110** (2013), no. 36, 14557–14562.

[GS73] Janos Galambos and Eugene Seneta, *Regularly varying sequences*, Proceedings of the American Mathematical Society **41** (1973), no. 1, 110–116.

[HLCH19] Lu-Jing Huang, Yin-Ting Liao, Lo-Bin Chang, and Chii-Ruey Hwang, *The smallest eigenvalues of random kernel matrices: Asymptotic results on the min kernel*, Statistics & Probability Letters **148** (2019), 23–29.

[HMRT22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, The Annals of Statistics **50** (2022), no. 2, 949–986.

[Kar33] Jovan Karamata, *Sur un mode de croissance régulière. Théorèmes fondamentaux*, Bulletin de la Société Mathématique de France **61** (1933), 55–62.

[KY17] Antti Knowles and Jun Yin, *Anisotropic local laws for random matrices*, Probability Theory and Related Fields **169** (2017), no. 1, 257–352.

[KZSS21] Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro, *Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting*, Advances in Neural Information Processing Systems **34** (2021), 20657–20668.

[LP09] Anna Lytova and Leonid Pastur, *Central limit theorem for linear eigenvalue statistics of random matrices with independent entries*, The Annals of Probability **37** (2009), no. 5, 1778–1840.

[MM21] Léo Miolane and Andrea Montanari, *The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning*, The Annals of Statistics **49** (2021), no. 4, 2313–2335.

[RMR21] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco, *Asymptotics of ridge (less) regression under general source condition*, International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3889–3897.

[RV09] Mark Rudelson and Roman Vershynin, *Smallest singular value of a random rectangular matrix*, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences **62** (2009), no. 12, 1707–1739.

[RV13] ———, *Hanson-Wright inequality and sub-Gaussian concentration*, Electronic Communications in Probability **18** (2013), 1–9.

[T+15] Joel A Tropp et al., *An introduction to matrix concentration inequalities*, Foundations and Trends® in Machine Learning **8** (2015), no. 1-2, 1–230.

[TAH18] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi, *Precise error analysis of regularized m-estimators in high dimensions*, IEEE Transactions on Information Theory **64** (2018), no. 8, 5592–5628.

[TB20] Alexander Tsigler and Peter L Bartlett, *Benign overfitting in ridge regression*, arXiv:2009.14286 (2020).

[Tib96] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Methodological) **58** (1996), no. 1, 267–288.

[TPT21] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis, *Fundamental limits of ridge-regularized empirical risk minimization in high dimensions*, International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 2773–2781.

[Tsy09] Alexandre B Tsybakov, *Introduction to nonparametric estimation*, Springer, 2009.

[VdV00] Aad W Van der Vaart, *Asymptotic statistics*, vol. 3, Cambridge university press, 2000.

[Ver12] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, Compressed Sensing: Theory and Applications (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012, pp. 210–268.

[WX20] Denny Wu and Ji Xu, *On the optimal weighted ℓ_2 regularization in overparameterized linear regression*, Advances in Neural Information Processing Systems **33** (2020), 10112–10123.

[Yas14] Pavel Yaskov, *Lower bounds on the smallest eigenvalue of a sample covariance matrix*, Electronic Communications in Probability **19** (2014), 1–10.

A Proof of Proposition 2.2

Since $\sigma_{k_\star} \geq \lambda_\star \geq \sigma_{k_\star+1}$, we have

$$\begin{aligned} k_\star + \frac{r_1(k_\star)}{b_{k_\star}} &= \sum_{l=1}^{k_\star} \frac{\sigma_l}{\sigma_l} + \sum_{l=k_\star+1}^{d} \frac{\sigma_l}{\sigma_{k_\star}} \leq \sum_{l=1}^{k_\star} \frac{\sigma_l+\lambda_\star}{\sigma_l+\lambda_\star} + \sum_{l=k_\star+1}^{d} \frac{\sigma_l}{\lambda_\star} \\ &\leq \sum_{l=1}^{k_\star} \frac{2\sigma_l}{\sigma_l+\lambda_\star} + \sum_{l=k_\star+1}^{d} \frac{2\sigma_l}{\sigma_l+\lambda_\star} = 2\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right) \leq 2n\,. \end{aligned}$$

Next we bound $\mathsf{V}_n(\lambda)$. Recalling that $\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right) \leq n(1-c_\star^{-1})$, it then follows

$$\begin{aligned} \mathsf{V}_n(\lambda) &= \frac{\tau^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right)}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right)} \leq \frac{c_\star \tau^2}{n} \cdot \left(\sum_{l=1}^{k_\star} \frac{\sigma_l^2}{(\sigma_l+\lambda_\star)^2} + \sum_{l=k_\star+1}^{d} \frac{\sigma_l^2}{(\sigma_l+\lambda_\star)^2} \right) \\ &\leq \frac{c_\star \tau^2}{n} \cdot \left(k_\star + \sum_{l=k_\star+1}^{d} \frac{\sigma_l^2}{\lambda_\star^2} \right) \leq c_\star \tau^2 \Big(\frac{k_\star}{n} + \frac{r_2(k_\star)}{n} \Big) \overset{\text{(i)}}{\leq} c_\star \tau^2 \Big(\frac{k_\star}{n} + \frac{4 b_{k_\star}^2 n}{\overline{r}(k_\star)} \Big)\,, \end{aligned}$$

where in (i) we use the previous bound $r_1(k_\star) \leq 2 b_{k_\star} n$. Finally, for the bias term, we have

$$\begin{aligned} \mathsf{B}_n(\lambda) &= \frac{\lambda_\star^2 \langle \boldsymbol{\beta}, (\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2} \boldsymbol{\Sigma} \boldsymbol{\beta} \rangle}{1 - n^{-1}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right)} \leq c_\star \sum_{l=1}^{d} \frac{\lambda_\star^2 \sigma_l}{(\sigma_l+\lambda_\star)^2} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2 \\ &\leq c_\star \left(\sum_{l=1}^{k_\star} \lambda_\star^2 \sigma_l^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2 + \sum_{l=k_\star+1}^{d} \sigma_l \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2 \right) \leq c_\star \Big(\sigma_{k_\star}^2 \|\boldsymbol{\beta}_{\leq k_\star}\|_{\boldsymbol{\Sigma}^{-1}}^2 + \|\boldsymbol{\beta}_{>k_\star}\|_{\boldsymbol{\Sigma}}^2 \Big)\,. \end{aligned}$$

B Auxiliary lemmas

B.1 Proof of Lemma 6.1

The lemma follows by pure calculations.

Identities for $\mathscr{V}_{\boldsymbol{X}}(\lambda)$ and $\mathscr{B}_{\boldsymbol{X}}(\lambda)$. Substitute in Eq. (35), we have

$$\begin{aligned} \tau^2 \cdot \frac{\partial}{\partial \lambda} \mathscr{F}_n(\lambda, 0; \boldsymbol{I}) &= \tau^2 \cdot \frac{\partial}{\partial \lambda} \left(\lambda \mathrm{Tr}\left(\boldsymbol{\Sigma} (\lambda \boldsymbol{I} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X})^{-1} \right) \right) \\ &= \tau^2 \left\{ \mathrm{Tr}\left(\boldsymbol{\Sigma} (\lambda \boldsymbol{I} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X})^{-1} \right) - \lambda \mathrm{Tr}\left(\boldsymbol{\Sigma} (\lambda \boldsymbol{I} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X})^{-2} \right) \right\} \\ &= \tau^2 \mathrm{Tr}\left(\boldsymbol{\Sigma} \boldsymbol{X}^\mathsf{T} \boldsymbol{X} (\lambda \boldsymbol{I} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X})^{-2} \right) = \mathscr{V}_{\boldsymbol{X}}(\lambda)\,, \end{aligned}$$

and similarly for the bias term

$$\begin{aligned} -\lambda \cdot \frac{\partial}{\partial \mu} \mathscr{F}_n(\lambda, 0; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}) &= -\lambda \cdot \frac{\partial}{\partial \mu} \left(\lambda \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\Sigma}^{\frac{1}{2}} (\lambda \boldsymbol{I} + \mu \boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X})^{-1} \right) \right) \Big|_{\mu=0} \\ &= \lambda^2 \, \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\Sigma}^{\frac{1}{2}} (\lambda \boldsymbol{I} + \mu \boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X})^{-1} \boldsymbol{\Sigma} (\lambda \boldsymbol{I} + \mu \boldsymbol{\Sigma} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X})^{-1} \right) \Big|_{\mu=0} \\ &= \lambda^2 \mathrm{Tr}\left(\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T} (\lambda \boldsymbol{I} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X})^{-1} \boldsymbol{\Sigma} (\lambda \boldsymbol{I} + \boldsymbol{X}^\mathsf{T} \boldsymbol{X})^{-1} \right) = \mathscr{B}_{\boldsymbol{X}}(\lambda)\,. \end{aligned}$$

Identities for $\mathsf{V}_n(\lambda)$ and $\mathsf{B}_n(\lambda)$. First we verify that $\mu_\star(\lambda, 0) = \lambda/\lambda_\star$. Set $\mu = 0$ in Eq. (36), we obtain

$$\mu_\star = \frac{n}{1 + \mathscr{R}_0(\lambda, \mu_\star; \boldsymbol{I})} = \frac{n}{1 + \mathrm{Tr}\left(\boldsymbol{\Sigma}(\mu_\star \boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1}\right)} = \frac{n}{1 + \mu_\star^{-1}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \frac{\lambda}{\mu_\star}\boldsymbol{I})^{-1}\right)},$$

and thus

$$n - \mu_\star = \mathrm{Tr}\left(\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \frac{\lambda}{\mu_\star}\boldsymbol{I}\right)^{-1}\right),$$

which proves the claim comparing to Eq. (5). Further by (36), we can compute the derivatives

$$\frac{\partial}{\partial \lambda}\mu_\star(\lambda, 0) = \frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)}, \tag{53a}$$

$$\frac{\partial}{\partial \mu}\mu_\star(\lambda, 0) = \frac{n}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)}. \tag{53b}$$

We can then proceed to write

$$\begin{aligned}
&\tau^2 \cdot \frac{\partial}{\partial \lambda}\mathscr{F}_0(\lambda, \mu_\star(\lambda, 0); \boldsymbol{I}) \\
&= \tau^2 \cdot \frac{\partial}{\partial \lambda}\left(\lambda \mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-1}\right)\right) \\
&= \tau^2 \left\{\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-1}\right) - \lambda \mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-2}\right) - \lambda \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-2}\right) \cdot \frac{\partial}{\partial \lambda}\mu_\star(\lambda, 0)\right\} \\
&= \frac{\tau^2}{\mu_\star}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right) \cdot \left(1 - \frac{\lambda_\star \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)}\right) \\
&= \frac{\tau^2}{\mu_\star}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right) \cdot \frac{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-1}\right)}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)} \\
&\overset{\text{(i)}}{=} \frac{\tau^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)} = \mathsf{V}_n(\lambda),
\end{aligned}$$

where in (i) we use Eq. (5) which implies $\mu_\star = n - \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-1}\right)$. For the bias we can compute

$$\begin{aligned}
&-\lambda \cdot \frac{\partial}{\partial \mu}\mathscr{F}_0(\lambda, \mu_\star(\lambda, 0); \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}) \\
&= \lambda^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-1}\right) \cdot \frac{n}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)} \\
&= \frac{\lambda_\star^2 \boldsymbol{\beta}^\mathsf{T}\left(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}}{1 - n^{-1}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right)} = \mathsf{B}_n(\lambda).
\end{aligned}$$

The proof is complete.

B.2 Proof of Lemma 6.2

The lemma is an analogue of [HMRT22, Lemma. 5], which requires two-sided differentiability around 0 and makes use of higher order central difference operators from numerical analysis. Here we apply

a more straightforward argument. For any $0 \le j \le k$, by Taylor expansion with Lagrange remainder, we can write

$$f(j\delta) = \sum_{l=0}^{k} j^l \cdot \frac{\delta^l}{j!} f^{(l)}(0) + j^{k+1} \cdot \frac{\delta^{k+1}}{(k+1)!} f^{(k+1)}(t_j)\,,$$

for some $t_j \in [0, j\delta]$. We can write the $k+1$ equations in matrix form,

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 4 & \cdots & 2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & k & k^2 & \cdots & k^k \end{bmatrix}}_{:=\boldsymbol{V}_k} \begin{bmatrix} f(0) \\ f'(0)\delta \\ f''(0)\delta^2/2 \\ \vdots \\ f^{(k)}(0)\delta^k/k! \end{bmatrix} + \frac{\delta^{k+1}}{(k+1)!} \begin{bmatrix} 0 \\ f^{(k+1)}(t_1) \\ 2^{k+1} f^{(k+1)}(t_2) \\ \vdots \\ k^{k+1} f^{(k+1)}(t_k) \end{bmatrix} = \begin{bmatrix} f(0) \\ f(\delta) \\ f(2\delta) \\ \vdots \\ f(k\delta) \end{bmatrix}.$$

The Vandermonde matrix $\boldsymbol{V}_k$ is invertible, and therefore we can write

$$\begin{bmatrix} f(0) \\ f'(0)\delta \\ f''(0)\delta^2/2 \\ \vdots \\ f^{(k)}(0)\delta^k/k! \end{bmatrix} = \boldsymbol{V}_k^{-1} \begin{bmatrix} f(0) \\ f(\delta) \\ f(2\delta) \\ \vdots \\ f(k\delta) \end{bmatrix} - \frac{\delta^{k+1}}{(k+1)!} \boldsymbol{V}_k^{-1} \begin{bmatrix} 0 \\ f^{(k+1)}(t_1) \\ 2^{k+1} f^{(k+1)}(t_2) \\ \vdots \\ k^{k+1} f^{(k+1)}(t_k) \end{bmatrix}.$$

Denote by $\|\boldsymbol{M}\|_\infty$ the ℓ_∞-induced operator norm, we thus have

$$\begin{aligned} |f'(0)\delta| &\le \left\| \begin{bmatrix} f(0) \\ f'(0)\delta \\ f''(0)\delta^2/2 \\ \vdots \\ f^{(k)}(0)\delta^k/k! \end{bmatrix} \right\|_\infty \le \left\| \boldsymbol{V}_k^{-1} \right\|_\infty \cdot \left(\left\| \begin{bmatrix} f(0) \\ f(\delta) \\ f(2\delta) \\ \vdots \\ f(k\delta) \end{bmatrix} \right\|_\infty + \frac{\delta^{k+1}}{(k+1)!} \left\| \begin{bmatrix} 0 \\ f^{(k+1)}(t_1) \\ 2^{k+1} f^{(k+1)}(t_2) \\ \vdots \\ k^{k+1} f^{(k+1)}(t_k) \end{bmatrix} \right\|_\infty \right) \\ &= \mathcal{O}_k \left(\max_{0 \le j \le k} |f(j\delta)| + \sup_{t \in [0, k\delta]} |f^{(k+1)}(t)| \cdot \delta^{k+1} \right). \end{aligned}$$

Dividing δ from both sides completes the proof.

B.3 Proof of Lemma 6.3

Part I: Derivative w.r.t. λ. By Lemma 6.1,

$$\frac{\partial}{\partial \lambda} \mathscr{F}_n(\lambda, 0; \boldsymbol{I}) = \mathscr{V}_{\boldsymbol{X}}(\lambda)/\tau^2 = \operatorname{Tr}\left(\boldsymbol{\Sigma} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-2} \right),$$

we can easily write out derivatives with respect to λ up to any order $k \ge 1$ as

$$\frac{\partial^k}{\partial \lambda^k} \mathscr{F}_n(\lambda, 0; \boldsymbol{I}) = \mathcal{O}_k \left(\operatorname{Tr}\left(\boldsymbol{\Sigma} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1-k} \right) \right),$$

and therefore

$$\begin{aligned} \left| \frac{\partial^k}{\partial \lambda^k} \mathscr{F}_n(\lambda, 0; \boldsymbol{I}) \right| &\le \mathcal{O}_k \left(\frac{1}{\lambda^{k-1}} \operatorname{Tr}\left(\boldsymbol{\Sigma} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-2} \right) \right) \le \mathcal{O}_k \left(\frac{1}{\lambda^{k-1}} \operatorname{Tr}\left(\boldsymbol{\Sigma} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \right) \right) \\ &= \mathcal{O}_k \left(\frac{\mathscr{F}_n(\lambda, 0; \boldsymbol{I})}{\lambda^k} \right). \end{aligned}$$

Part II: Derivative w.r.t. μ. We can directly compute that

$$\begin{aligned}
\left|\frac{\partial^k}{\partial \mu^k}\mathscr{F}_n(\lambda,\mu;\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\right| &= \left|\frac{\partial^k}{\partial \mu^k}\cdot \lambda \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I}+\mu\boldsymbol{\Sigma}+\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right)\right| \\
&= \mathcal{O}_k\left(\lambda \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}\left((\lambda \boldsymbol{I}+\mu\boldsymbol{\Sigma}+\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}\right)^k(\lambda \boldsymbol{I}+\mu\boldsymbol{\Sigma}+\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right)\right) \\
&= \mathcal{O}_k\left(\lambda \mathrm{Tr}\left(\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I}+\mu\boldsymbol{\Sigma}+\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{k+1}\right)\right) \\
&\overset{\text{(i)}}{=} \mathcal{O}_k\left(\lambda^{1-k}\mathrm{Tr}\left(\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda \boldsymbol{I}+\mu\boldsymbol{\Sigma}+\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right)\right) = \mathcal{O}_k\left(\frac{\mathscr{F}_n(\lambda,\mu;\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\lambda^k}\right),
\end{aligned}$$

where in (i) we use $\|\boldsymbol{\Sigma}\| = 1$.

B.4 Proof of Lemma 6.4

Part I: Derivative w.r.t. λ. Note that

$$\mathscr{F}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{I}) = \lambda \mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda \boldsymbol{I}+\mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right) = \lambda_\star \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right).$$

Combining with the fixed-point equation (5) that determines $\lambda_\star$, we further get

$$\mathscr{F}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{I}) = n\lambda_\star - \lambda\,.$$

Therefore, for all $k \geq 1$.

$$\frac{\partial}{\partial \lambda^k}\mathscr{F}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{I}) = n\cdot\frac{\partial^k \lambda_\star}{\partial \lambda^k} - \mathbb{I}\{k=1\}\,, \tag{54}$$

and it boils down to controlling higher order derivatives of $\lambda_\star$ w.r.t. λ. Of course, we need to first show that we can actually write $\lambda_\star = \lambda_\star(\lambda)$ locally by implicit function theorem. Since

$$\lambda = \lambda_\star \cdot \left(n - \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right)\right)$$

which is clearly a increasing function of $\lambda_\star$ on the right hand side, and thus $\partial\lambda/\partial\lambda_\star > 0$ and the implicit function theorem applies. To calculate the higher order derivative of the inverse function, we apply the formula for higher order derivatives of inverse function [Apo00]

$$\frac{\partial^k \lambda_\star}{\partial \lambda^k} = \left|\frac{\partial \lambda}{\partial \lambda_\star}\right|^{1-2k}\cdot \sum_{\substack{m_1+m_2+\cdots+m_p=k-1\\ m_1+2m_2+\cdots+pm_p=2k-2}} \mathcal{O}_k\left(\prod_{l=1}^p\left(\frac{\partial^l \lambda}{\partial \lambda_\star^l}\right)^{m_l}\right). \tag{55}$$

To further upper bound the above display, we need a lower bound for the derivative $\partial\lambda/\partial\lambda_\star$ and upper bounds for higher order derivatives $\partial^l\lambda/\partial\lambda_\star^l$. Using the Leibniz rule, we can compute that

$$\begin{aligned}
\frac{\partial^l \lambda}{\partial \lambda_\star^l} &= \sum_{r=0}^l \binom{l}{r}\frac{\partial^r \lambda_\star}{\partial \lambda_\star^r}\cdot\frac{\partial^{l-r}\left(n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right)\right)}{\partial \lambda_\star^{l-r}} \\
&= \lambda_\star\cdot\frac{\partial^l\left(n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right)\right)}{\partial \lambda_\star^l} + l\cdot\frac{\partial^{l-1}\left(n-\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right)\right)}{\partial \lambda_\star^{l-1}}\,.
\end{aligned}$$

For $l = 1$, since

$$\frac{\partial \lambda}{\partial \lambda_\star} = \lambda_\star \cdot \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right) + n - \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right) = n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right)$$
$$\geq n - \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right) = \frac{\lambda}{\lambda_\star} \geq n\kappa\,,$$

we have $n\kappa \leq \partial\lambda/\partial\lambda_\star \leq n$. When $l \geq 2$, we get

$$\frac{\partial^l \lambda}{\partial \lambda_\star^l} = (-1)^{l-1} l! \cdot \lambda_\star \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-l-1}\right) + (-1)^{l-2} l! \cdot \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-l}\right) = (-1)^{l-2} l! \cdot \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-l-1}\right)$$
$$= \mathcal{O}_l\left(\left\|(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-l+1}\right\| \cdot \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right)\right) = \mathcal{O}_l\left(\frac{n}{\lambda_\star^{l-1}}\right).$$

Substituting the above displays into Eq. (55) yields

$$\frac{\partial^k \lambda_\star}{\partial \lambda^k} = \left(\frac{1}{(n\kappa)^{2k-1}}\right) \cdot \sum_{\substack{m_1+m_2+\cdots+m_p=k-1 \\ m_1+2m_2+\cdots+pm_p=2k-2}} \mathcal{O}_k\left(\prod_{l=1}^{p} \mathcal{O}_l\left(\frac{n^{m_l}}{\lambda_\star^{lm_l-m_l}}\right)\right) = \mathcal{O}_k\left(\frac{1}{n^k \lambda_\star^{k-1} \cdot \kappa^{2k-1}}\right)$$

Taken collectively with Eq. (54) and $\mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{I}) = \lambda_\star \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right) \geq \kappa n \lambda_\star$, we obtain for all $k \geq 2$,

$$\left|\frac{\partial^k}{\partial \lambda^k}\mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{I})\right| = n\left|\frac{\partial^k \lambda_\star}{\partial \lambda^k}\right| = \mathcal{O}_k\left(\frac{\mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{I})}{n^k \lambda_\star^k \kappa^{2k}}\right) = \mathcal{O}_k\left(\frac{\mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{I})}{\lambda^k \kappa^{2k}}\right),$$

where we use Assumption (21) again for the final bound. This is also valid for $k = 1$ as

$$\left|\frac{\partial}{\partial \lambda}\mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{I})\right| = n\left|\frac{\partial \lambda_\star}{\partial \lambda}\right| + 1 = \mathcal{O}\left(\frac{\mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{I})}{\lambda\kappa^2}\right) + 1 = \mathcal{O}\left(\frac{n\lambda_\star - \lambda}{\lambda\kappa^2}\right) + 1$$
$$= \mathcal{O}\left(\frac{n\lambda_\star - \lambda}{\lambda\kappa^2}\right) = \mathcal{O}\left(\frac{\mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{I})}{\lambda\kappa^2}\right),$$

where we use $\mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{I})/\lambda = n\lambda_\star/\lambda - 1$ and

$$\frac{n\lambda_\star - \lambda}{\lambda\kappa^2} \geq \left((1-\kappa)^{-1} - 1\right)\kappa^{-2} \geq \kappa^{-1} \geq 1\,.$$

Part II: Derivative w.r.t. μ. Now we fix λ and allow μ be take nonzero values. We will also use the shorthand $\mu_\star = \mu_\star(\lambda, \mu)$. Similar to the previous part, we apply Faà di Bruno's formula to $\mathscr{F}_0$ and bound

$$\left|\frac{\partial^k}{\partial \mu}\mathscr{F}_0(\lambda, \mu_\star; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\right| = \sum_{m_1+2m_2+\cdots+pm_p=k} \mathcal{O}_k\left(\frac{\partial^{m_1+\cdots+m_p}}{\partial \mu_\star^{m_1+\cdots+m_p}}\mathscr{F}_0(\lambda, \mu_\star; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}) \cdot \prod_{l=1}^{p}\left(\frac{\partial^l \mu_\star}{\partial \mu^l}\right)^{m_l}\right). \quad (56)$$

For any $1 \leq l \leq k-1$, we have

$$\left|\frac{\partial^l}{\partial \mu_\star^l}\mathscr{F}_0(\lambda, \mu_\star; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\right| = \mathcal{O}_l\left(\lambda \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}} \cdot \boldsymbol{\Sigma}^l(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-1-l}\right)\right) = \mathcal{O}_l\left(\frac{\mathscr{F}_0(\lambda, \mu_\star; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\mu_\star^l}\right).$$

To bound higher order derivatives $\partial^l \mu_\star/\partial \mu^l$, we apply again the formula for higher order derivatives of inverse function. Of course, this would first require showing the existence of inverse function

by implicit function theorem, which will be evident as we will provide a lower bound for $|\partial\mu/\partial\mu_\star|$ below. By [Apo00], we have for all $1 \le l \le k-1$,

$$\left|\frac{\partial^l \mu_\star}{\partial \mu^l}\right| = \left|\frac{\partial \mu}{\partial \mu_\star}\right|^{1-2l} \cdot \sum_{\substack{m_1+m_2+\cdots+m_p=l-1 \\ m_1+2m_2+\cdots+pm_p=2l-2}} \mathcal{O}_l\left(\prod_{r=1}^{p}\left(\frac{\partial^r \mu}{\partial \mu_\star^r}\right)^{m_r}\right). \tag{57}$$

This is a more manageable formula as we can explicitly write μ as a function of $\mu_\star$

$$\mu = \mu_\star - \frac{n}{1+\mathscr{R}_0(\lambda,\mu_\star;\boldsymbol{I})} = \mu_\star - \frac{n}{1+\mathrm{Tr}(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-1})}.$$

We can compute the first order derivative as

$$\frac{\partial \mu}{\partial \mu_\star} = 1 - \frac{n\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-2}\right)}{\left(1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-1}\right)\right)^2} = 1 - \frac{(\mu_\star-\mu)\cdot\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-2}\right)}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-1}\right)},$$

which, together with $0 \le \mu \le \mu_\star/2$, implies a lower bound

$$\begin{aligned}\frac{\partial \mu}{\partial \mu_\star} &\ge 1 - \frac{\mu_\star\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-2}\right)}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-1}\right)} = \frac{1+\lambda\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-2}\right)}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-1}\right)} \\ &\ge \frac{1}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-1}\right)} = \frac{\mu_\star-\mu}{n} \ge \frac{\mu_\star}{2n}.\end{aligned}$$

To further bound higher order derivatives, we again appeal to Faà di Bruno's formula. Use the shorthand $\mathscr{R}_0 = \mathscr{R}_0(\lambda,\mu_\star;\boldsymbol{I})$, we have for all $r \ge 1$,

$$\left|\frac{\partial^r \mu}{\partial \mu_\star^r}\right| = \sum_{m_1+2m_2+\cdots+pm_p=r} \mathcal{O}_r\left(\frac{\partial^{m_1+\cdots+m_p}}{\partial \mathscr{R}_0^{m_1+\cdots+m_p}}\frac{n}{1+\mathscr{R}_0}\cdot\prod_{s=1}^{p}\left(\frac{\partial^s \mathscr{R}_0}{\partial \mu_\star^s}\right)^{m_s}\right).$$

Making use of the following two bounds,

$$\begin{aligned}\frac{\partial^s}{\partial \mathscr{R}_0^s}\frac{n}{1+\mathscr{R}_0} &= \mathcal{O}_s\left(\frac{n}{(1+\mathscr{R}_0)^{s+1}}\right) = \mathcal{O}_s\left(\frac{\mu_\star}{(1+\mathscr{R}_0)^s}\right), \\ \frac{\partial^s \mathscr{R}_0}{\partial \mu_\star^s} &= \mathcal{O}_s\left(\mathrm{Tr}\left(\boldsymbol{\Sigma}^{s+1}(\lambda\boldsymbol{I}+\mu_\star\boldsymbol{\Sigma})^{-s-1}\right)\right) = \mathcal{O}_s\left(\frac{\mathscr{R}_0}{\mu_\star^s}\right),\end{aligned}$$

we can further obtain

$$\left|\frac{\partial^r \mu}{\partial \mu_\star^r}\right| = \sum_{m_1+2m_2+\cdots+pm_p=r} \mathcal{O}_r\left(\frac{\mu_\star}{(1+\mathscr{R}_0)^{m_1+\cdots+m_p}}\cdot\prod_{s=1}^{p}\frac{\mathscr{R}_0^{m_s}}{\mu_\star^{sm_s}}\right) = \mathcal{O}_r\left(\frac{1}{\mu_\star^{r-1}}\right).$$

Taking the above displays into Eq. (57) and use the condition $\mu_\star/n \ge \kappa$, we have

$$\left|\frac{\partial^l \mu_\star}{\partial \mu^l}\right| = \mathcal{O}_l\left(\frac{1}{\kappa^{2l-1}}\right)\cdot\sum_{\substack{m_1+m_2+\cdots+m_p=l-1 \\ m_1+2m_2+\cdots+pm_p=2l-2}} \mathcal{O}_l\left(\prod_{r=1}^{p}\mathcal{O}_r\left(\frac{1}{\mu_\star^{rm_r-m_r}}\right)\right) = \mathcal{O}_l\left(\frac{1}{\mu_\star^{l-1}\kappa^{2l-1}}\right).$$

Finally, taking the above display back into Eq. (56) yields

$$\begin{aligned}\left|\frac{\partial^k}{\partial \mu^k}\mathscr{F}_0(\lambda,\mu_\star;\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})\right| &= \sum_{m_1+2m_2+\cdots+pm_p=k} \mathcal{O}_k\left(\frac{\mathscr{F}_0(\lambda,\mu_\star;\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\mu_\star^{m_1+\cdots+m_p}}\cdot\prod_{l=1}^{p}\mathcal{O}_l\left(\frac{1}{\mu_\star^{lm_l-m_l}\kappa^{2lm_l-m_l}}\right)\right) \\ &= \mathcal{O}_k\left(\frac{\mathscr{F}_0(\lambda,\mu_\star;\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\mu_\star^k\kappa^{2k}}\right).\end{aligned}$$

B.5 Proof of Lemma 6.6

First we show $\mu_\star(\lambda,\mu)$ is increasing in μ when $\mu \geq 0$. To this end, we consider the function

$$f(t) = t - \frac{n}{1+\mathscr{R}_0(\lambda,t;\boldsymbol{I})}\,.$$

By Eq. (36), we have $f(\mu_\star(\lambda,\mu)) = \mu$ for all $\mu \geq 0$. Further, we prove $f(t)$ is increasing in $[\mu_\star(\lambda,0),\infty)$. We write

$$f'(t) = 1 - \frac{n\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda\boldsymbol{I}+t\boldsymbol{\Sigma})^{-2}\right)}{\left(1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+t\boldsymbol{\Sigma})^{-1}\right)\right)^2} \overset{\text{(i)}}{=} 1 - \frac{(t-f(t))\cdot\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda\boldsymbol{I}+t\boldsymbol{\Sigma})^{-2}\right)}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+t\boldsymbol{\Sigma})^{-1}\right)}\,,$$

where in (i) we use that $t - f(t) = n/(1+\mathscr{R}_0(\lambda,t;\boldsymbol{I}))$. Define

$$g(t) := \frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda\boldsymbol{I}+t\boldsymbol{\Sigma})^{-2}\right)}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+t\boldsymbol{\Sigma})^{-1}\right)}\,,$$

we have

$$f'(t) - f(t)g(t) = 1 - \frac{t\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda\boldsymbol{I}+t\boldsymbol{\Sigma})^{-2}\right)}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+t\boldsymbol{\Sigma})^{-1}\right)} = \frac{1+\lambda\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+t\boldsymbol{\Sigma})^{-2}\right)}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+t\boldsymbol{\Sigma})^{-1}\right)} > 0\,,$$

and therefore $e^{-g(t)}f(t)$ is increasing. As $f(\mu_\star(\lambda,0)) = 0$ (cf. Eq. (36)), we must have $f(t) \geq 0$ for all $t \geq \mu_\star(\lambda,0)$. Substituting back into the above display with $g(t) \geq 0$ yields

$$f'(t) \geq f'(t) - f(t)g(t) > 0\,.$$

We then proceed to show a sufficient condition for $\mu_\star(\lambda,\mu) \leq (1-\kappa/2)^{-1}\mu_\star(\lambda,0)$ is $0 \leq \mu \leq n\kappa^3/2$ under Assumption (21). Provided with monotonicity of $f(t)$, the desired condition $\mu_\star(\lambda,\mu) \leq (1-\kappa/2)^{-1}\mu_\star(\lambda,0)$ is essentially equivalent to $\mu = f(\mu_\star(\lambda,\mu)) \leq f((1-\kappa/2)^{-1}\mu_\star(\lambda,0))$. Together with $\mu_\star(\lambda,0) = n/(1+\mathscr{R}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{I}))$, we obtain a lower bound for the right hand side

$$\begin{aligned}
&f((1-\kappa/2)^{-1}\mu_\star(\lambda,0)) \\
&= (1-\kappa/2)^{-1}\mu_\star(\lambda,0) - \frac{n}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+(1-\kappa/2)^{-1}\mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right)} \\
&= \frac{n(1-\kappa/2)^{-1}}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right)} - \frac{n(1-\kappa/2)^{-1}}{(1-\kappa/2)^{-1}+\mathrm{Tr}\left(\boldsymbol{\Sigma}((1-\kappa/2)\lambda\boldsymbol{I}+\mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right)} \\
&\overset{\text{(i)}}{\geq} \frac{n(1-\kappa/2)^{-1}}{1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right)} - \frac{n(1-\kappa/2)^{-1}}{(1-\kappa/2)^{-1}+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right)} \\
&= \frac{n\left((1-\kappa/2)^{-1}-1\right)}{\left(1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right)\right)\cdot\left(1+(1-\kappa/2)\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right)\right)} \\
&\overset{\text{(ii)}}{\geq} \frac{n\left((1-\kappa/2)^{-1}-1\right)}{\left(1+\mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right)\right)^2} \\
&= \frac{\left((1-\kappa/2)^{-1}-1\right)\mu_\star(\lambda,0)^2}{n}\,,
\end{aligned}$$

where in (i) and (ii) we use two times the trivial bound $1-\kappa/2 \leq 1$. By Assumption (21),

$$\frac{\mu_\star(\lambda,0)}{n} = \frac{\lambda}{n\lambda_\star} \geq \kappa,$$

we know

$$f((1-\kappa/2)^{-1}\mu_\star(\lambda,0)) \geq n\cdot\kappa^2\left((1-\kappa/2)^{-1}-1\right) \geq n\cdot\kappa^3/2\,,$$

and thus a sufficient condition for $\mu_\star(\lambda,\mu) \leq (1-\kappa/2)^{-1}\mu_\star(\lambda,0)$ is $\mu \leq n\kappa^3/2$.

C Proofs for Theorem 5

C.1 Proof of Lemma 7.1

We define $\overline{\mu}_i := \inf\left\{\mu \mid \boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda \boldsymbol{I} + \mu\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}} \succ 0\right\}$. Note that, by construction $\overline{\mu}_{i+1} \le \overline{\mu}_i$. Let $\boldsymbol{\varphi} \in \mathbb{R}^d$ be the leading normalized eigenvector of $\boldsymbol{\Sigma}$. If $\mu \le -(\lambda + \|\boldsymbol{X}_i\boldsymbol{\varphi}\|^2)/\|\boldsymbol{\Sigma}\|$, it follows that

$$\boldsymbol{\varphi}^\mathsf{T}\left(\lambda\boldsymbol{I} + \mu\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)\boldsymbol{\varphi} = \lambda + \mu\|\boldsymbol{\Sigma}\| + \|\boldsymbol{X}_i\boldsymbol{\varphi}\|^2 \le 0\,,$$

which implies $\overline{\mu}_i \ge -(\lambda + \|\boldsymbol{X}_i\boldsymbol{\varphi}\|^2)/\|\boldsymbol{\Sigma}\| > -\infty$. The update rule is equivalent to solving the equation

$$\mu_{i+1} + \frac{1}{1 + \operatorname{Tr}\left(\boldsymbol{\Sigma}\left(\lambda\boldsymbol{I} + \mu_{i+1}\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)^{-1}\right)} = \mu_i\,, \qquad \mu_{i+1} \in (\overline{\mu}_i, \infty)\,.$$

For all $t \in (\overline{\mu}_i, \infty)$, let

$$f(t) = t + \frac{1}{1 + \operatorname{Tr}\left(\boldsymbol{\Sigma}\left(\lambda\boldsymbol{I} + t\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)^{-1}\right)}\,.$$

In this given domain, $\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I} + t\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}} \succ 0$ and thus $\operatorname{Tr}\left(\boldsymbol{\Sigma}\left(\lambda\boldsymbol{I} + t\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)^{-1}\right)$ is decreasing in t (this can be seen by computing its derivative with respect to t), which further implies $f(t)$ is strictly increasing in t. Since

$$\lim_{t \downarrow \overline{\mu}_i} \operatorname{Tr}\left(\boldsymbol{\Sigma}\left(\lambda\boldsymbol{I} + t\boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}_i\right)^{-1}\right) = \infty\,,$$

and we have

$$\lim_{t \downarrow \overline{\mu}_i} f(t) = \overline{\mu}_i < \mu_i\,, \qquad f(\mu_i) > \mu_i\,.$$

(The first inequality follows since $\mu_i \in (\overline{\mu}_{i-1}, \infty)$ and $\overline{\mu}_i \le \overline{\mu}_{i-1}$.) Thus, there must be a unique $\mu_{i+1} \in (\overline{\mu}_i, \mu_i)$ that solves $f(\mu_{i+1}) = \mu_i$, proving the lemma.

C.2 Proof of Lemma 7.2

Without loss of generality, we can always assume $d \ge n$ or simply $d = \infty$ by embedding $\mathbb{R}^d$ into the Hilbert space ℓ_2 since we always have

$$\sum_{l=k}^{d} \sigma_l \le \mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\,,$$

when $\sigma_k = 0$. We write the spectral decomposition of $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \sum_{i=1}^{d} \sigma_i \boldsymbol{v}_i\boldsymbol{v}_i^\mathsf{T}\,,$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq \cdots$ where $\{\boldsymbol{v}_i\}$ form an orthogonal basis of eigenvectors. For any $k \leq n$, define the projection operators

$$\boldsymbol{P}_k := \sum_{i=1}^{k} \boldsymbol{v}_i \boldsymbol{v}_i^\mathsf{T}\,, \qquad \boldsymbol{P}_k^\perp := \boldsymbol{I} - \boldsymbol{P}_k = \sum_{i=k+1}^{d} \boldsymbol{v}_i \boldsymbol{v}_i^\mathsf{T}\,,$$

and we write

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{P}_k + \boldsymbol{X}\boldsymbol{P}_k^\perp := \boldsymbol{U}_k + \boldsymbol{W}_k\,.$$

Part I: Decomposing into the top and lower eigenspaces. By writing $\boldsymbol{X} = \boldsymbol{U}_k + \boldsymbol{W}_k$, we can have the following inequality:

Lemma C.1. *For any $1 \leq k \leq n-1$,*

$$\lambda \boldsymbol{I} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X} \succeq \frac{\lambda}{2}\boldsymbol{I} + \left(1 + \frac{2\left\|\boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k\right\|}{\lambda}\right)^{-1} \boldsymbol{U}_k^\mathsf{T}\boldsymbol{U}_k\,.$$

Proof. Note that

$$\begin{aligned}
\lambda \boldsymbol{I} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X} &= \lambda \boldsymbol{I} + \boldsymbol{U}_k^\mathsf{T}\boldsymbol{U}_k + \boldsymbol{U}_k^\mathsf{T}\boldsymbol{W}_k + \boldsymbol{W}_k^\mathsf{T}\boldsymbol{U}_k + \boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k \\
&\succeq \frac{\lambda}{2}\boldsymbol{I} + \boldsymbol{U}_k^\mathsf{T}\boldsymbol{U}_k + \boldsymbol{U}_k^\mathsf{T}\boldsymbol{W}_k + \boldsymbol{W}_k^\mathsf{T}\boldsymbol{U}_k + \left(1 + \frac{\lambda}{2\left\|\boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k\right\|}\right)\boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k \\
&= \frac{\lambda}{2}\boldsymbol{I} + \left(1 + \frac{2\left\|\boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k\right\|}{\lambda}\right)^{-1} \boldsymbol{U}_k^\mathsf{T}\boldsymbol{U}_k + \boldsymbol{C}_k\,,
\end{aligned}$$

where

$$\boldsymbol{C}_k = \left(1 + \frac{\lambda}{2\left\|\boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k\right\|}\right)^{-1} \boldsymbol{U}_k^\mathsf{T}\boldsymbol{U}_k + \boldsymbol{U}_k^\mathsf{T}\boldsymbol{W}_k + \boldsymbol{W}_k^\mathsf{T}\boldsymbol{U}_k + \left(1 + \frac{\lambda}{2\left\|\boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k\right\|}\right)\boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k = \boldsymbol{D}_k^\mathsf{T}\boldsymbol{D}_k \succeq 0\,,$$

with

$$\boldsymbol{D}_k = \left(1 + \frac{\lambda}{2\left\|\boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k\right\|}\right)^{-\frac{1}{2}} \boldsymbol{U}_k + \left(1 + \frac{\lambda}{2\left\|\boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k\right\|}\right)^{\frac{1}{2}} \boldsymbol{W}_k\,.$$

□

To apply the above lemma, we need to further provide an upper bound on $\|\boldsymbol{W}_k^\mathsf{T}\boldsymbol{W}_k\|$, which we summarize as the following result.

Lemma C.2. *Let Assumption 1 holds, we have for any $1 \leq k \leq n-1$ with probability $1 - \mathcal{O}(n^{-D})$ that,*

$$\|\boldsymbol{W}_k\boldsymbol{W}_k^\mathsf{T}\| = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n)\right)\,.$$

Proof. Let $\boldsymbol{\zeta}_i = \boldsymbol{P}_k^\perp \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{P}_k^\perp \in \mathbb{R}^{d\times d}$, we can write

$$\boldsymbol{S}_k := \boldsymbol{W}_k^\mathsf{T} \boldsymbol{W}_k = \sum_{i=1}^n \boldsymbol{P}_k^\perp \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{P}_k^\perp = \sum_{i=1}^n \boldsymbol{\zeta}_i \,.$$

Since $\|\boldsymbol{\zeta}_i\| = \left\|\boldsymbol{P}_k^\perp \boldsymbol{x}_i\right\|^2$, we can apply Hanson-Wright inequality (cf. Lemma 2.1) and conclude that

$$\begin{aligned}\mathbb{P}\left(\left|\|\boldsymbol{\zeta}_i\| - \operatorname{Tr}\left(\boldsymbol{P}_k^\perp \boldsymbol{\Sigma}\right)\right| \geq t\right) &= \mathbb{P}\left(\left|\boldsymbol{x}_i^\mathsf{T} \boldsymbol{P}_k^\perp \boldsymbol{x}_i - \operatorname{Tr}\left(\boldsymbol{P}_k^\perp \boldsymbol{\Sigma}\right)\right| \geq t\right) \\ &\leq 2\exp\left\{-\Omega\left(\min\left\{\frac{t^2}{\mathsf{C}_{\boldsymbol{x}}^4 \left\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{P}_k^\perp \boldsymbol{\Sigma}^{\frac{1}{2}}\right\|_F^2}, \frac{t}{\mathsf{C}_{\boldsymbol{x}}^2 \left\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{P}_k^\perp \boldsymbol{\Sigma}^{\frac{1}{2}}\right\|}\right\}\right)\right\} .\end{aligned}$$

For $t = \Theta_{\mathsf{C}_{\boldsymbol{x}},D}(\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{P}_k^\perp \boldsymbol{\Sigma}^{\frac{1}{2}}\|_F \log n)$ we have with probability $1 - \mathcal{O}(n^{-D})$ that for all $i = 1, 2, \cdots, n$

$$\|\boldsymbol{\zeta}_i\| \leq \operatorname{Tr}\left(\boldsymbol{P}_k^\perp \boldsymbol{\Sigma}\right) + \Theta_{\mathsf{C}_{\boldsymbol{x}},D}\left(\left\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{P}_k^\perp \boldsymbol{\Sigma}^{\frac{1}{2}}\right\|_F \log n\right) = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\operatorname{Tr}\left(\boldsymbol{P}_k^\perp \Sigma\right) \log n\right) ,$$

where the last inequality follows from $\|\boldsymbol{\Sigma}\| = 1$ and

$$\left\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{P}_k^\perp \boldsymbol{\Sigma}^{\frac{1}{2}}\right\|_F = \sqrt{\operatorname{Tr}\left(\boldsymbol{P}_k^\perp \boldsymbol{\Sigma} \boldsymbol{P}_k^\perp \boldsymbol{\Sigma}\right)} \leq \operatorname{Tr}\left(\boldsymbol{P}_k^\perp \boldsymbol{\Sigma}\right) .$$

In the next step, we will adopt a standard truncation argument and apply a matrix concentration inequality. By setting $L_k := \Theta_{\mathsf{C}_{\boldsymbol{x}},D}(\operatorname{Tr}\left(\boldsymbol{P}_k^\perp \boldsymbol{\Sigma}\right) \log n)$, $\tilde{\boldsymbol{\zeta}}_i := \boldsymbol{\zeta}_i \mathbb{I}\{\|\boldsymbol{\zeta}_i\| \leq L_k\}$ and considering

$$\tilde{\boldsymbol{S}}_k := \sum_{i=1}^k \tilde{\boldsymbol{\zeta}}_i = \sum_{i=1}^k \boldsymbol{\zeta}_i \mathbb{I}\left\{\|\boldsymbol{\zeta}_i\| \leq L_k\right\} ,$$

we have $\tilde{\boldsymbol{S}}_k = \boldsymbol{S}_k$ with probability $1 - \mathcal{O}(n^{-D})$. It order to bound $\|\tilde{\boldsymbol{S}}_k\|$, we will use matrix Bernstein inequality. Since we know $\|\tilde{\boldsymbol{\zeta}}_i\| \leq L_k$ by construction, we only need to upper bound the matrix variance. The $\tilde{\boldsymbol{\zeta}}_i$'s' are independent symmetric random matrices and therefore we have

$$\operatorname{Var}(\tilde{\boldsymbol{S}}_k) \preceq \sum_{i=1}^n \mathbb{E}[\tilde{\boldsymbol{\zeta}}_i^2] \overset{\text{(i)}}{\preceq} \sum_{i=1}^n L_k \mathbb{E}[\tilde{\boldsymbol{\zeta}}_i] \overset{\text{(ii)}}{\preceq} \sum_{i=1}^n L_k \mathbb{E}\left[\boldsymbol{\zeta}_i\right] \overset{\text{(iii)}}{\preceq} nL_k \cdot \boldsymbol{P}_k^\perp \boldsymbol{\Sigma} \boldsymbol{P}_k^\perp =: \boldsymbol{V}_k \,,$$

where in (i) we use $\|\tilde{\boldsymbol{\zeta}}_i\| \leq L_k$, in (ii) we apply $\tilde{\boldsymbol{\zeta}}_i \preceq \boldsymbol{\zeta}_i$ and lastly in (iii) we use $\mathbb{E}[\boldsymbol{\zeta}_i] = \mathbb{E}[\boldsymbol{P}_k^\perp \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{P}_k^\perp] = \boldsymbol{P}_k^\perp \boldsymbol{\Sigma} \boldsymbol{P}_k^\perp$. It then follows that $\|\boldsymbol{V}_k\| \leq nL_k \|\boldsymbol{P}_k^\perp \boldsymbol{\Sigma} \boldsymbol{P}_k^\perp\| = n\sigma_{k+1} L_k \leq n\sigma_k L_k =: v_k$. Combine with the bound on the intrinsic dimension under Assumption 1,

$$\mathsf{intdim}\left(\boldsymbol{V}_k\right) = \frac{\operatorname{Tr}\left(\boldsymbol{V}_k\right)}{\|\boldsymbol{V}_k\|} = \frac{\sum_{l=k+1}^\infty \sigma_l}{\sigma_{k+1}} \leq \mathsf{d}_{\boldsymbol{\Sigma}} \,,$$

we can thus deduce from the Bernstein inequality with intrinsic dimension [T+15, Theorem 7.3.1] that for $t \geq \sqrt{v_k} + L_k/3$

$$\mathbb{P}(\|\tilde{\boldsymbol{S}}_k - \mathbb{E}[\tilde{\boldsymbol{S}}_k]\| \geq t) \leq 4\mathsf{d}_{\boldsymbol{\Sigma}} \cdot \exp\left(\frac{-t^2/2}{v_k + L_k t/3}\right) .$$

Finally, by further bounding the mean

$$\|\mathbb{E}[\tilde{\boldsymbol{S}}_k]\| \leq \|n\mathbb{E}[\tilde{\boldsymbol{\zeta}}_i]\| \leq n \cdot \|\mathbb{E}\left[\boldsymbol{\zeta}_i\right]\| = n \cdot \left\|\boldsymbol{P}_k^\perp \boldsymbol{\Sigma} \boldsymbol{P}_k^\perp\right\| \leq n\sigma_k \,,$$

we can obtain with probability $1-\mathcal{O}(n^{-D})$,

$$\begin{aligned}\|\tilde{\boldsymbol{S}}_k\| &= \mathcal{O}_D\left((\sqrt{v_k}+L_k)\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)\right)+\|\mathbb{E}[\tilde{\boldsymbol{S}}_k]\| \\ &\overset{\text{(i)}}{=} \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\left\{\sqrt{n\sigma_k\cdot\mathrm{Tr}\left(\boldsymbol{P}_k^{\perp}\boldsymbol{\Sigma}\right)\log n}+\mathrm{Tr}\left(\boldsymbol{P}_k^{\perp}\boldsymbol{\Sigma}\right)\log n\right\}\cdot\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)+n\sigma_k\right) \\ &\overset{\text{(ii)}}{=} \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\left\{\sqrt{n\sigma_k\cdot\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\log n}+\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\log n\right\}\cdot\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)+n\sigma_k\right)\end{aligned}$$

where in (i) we make use of $v_k=n\sigma_k L_k$, and apply Assumption 1 for the spectrum in (ii). Next by the fact that $\mathsf{d}_{\boldsymbol{\Sigma}}\geq n$, we can further write

$$\|\tilde{\boldsymbol{S}}_k\| = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)\right).$$

The proof is complete as $\boldsymbol{W}_k\boldsymbol{W}_k^{\mathsf{T}}=\boldsymbol{S}_k=\tilde{\boldsymbol{S}}_k=$ holds with probability $1-\mathcal{O}(n^{-D})$. □

To bound the norm of $\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I}+\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}$, we apply Lemmas C.1 and C.2 and obtain

$$\begin{aligned}\lambda\boldsymbol{I}+\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} &\succeq \frac{\lambda}{2}\boldsymbol{I}+\left(1+\frac{2\left\|\boldsymbol{W}_k^{\mathsf{T}}\boldsymbol{W}_k\right\|}{\lambda}\right)^{-1}\boldsymbol{U}_k^{\mathsf{T}}\boldsymbol{U}_k \\ &\succeq \frac{\lambda}{2}\boldsymbol{I}+\left(1+\frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)\right)}{\lambda}\right)^{-1}\boldsymbol{U}_k^{\mathsf{T}}\boldsymbol{U}_k.\end{aligned}$$

Therefore by block matrix inverse, we can further get

$$\begin{aligned}&\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I}+\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}} \\ &\preceq \boldsymbol{\Sigma}^{\frac{1}{2}}\left(\frac{\lambda}{2}\boldsymbol{I}+\left(1+\frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)\right)}{\lambda}\right)^{-1}\boldsymbol{U}_k^{\mathsf{T}}\boldsymbol{U}_k\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}} \\ &= \boldsymbol{\Sigma}^{\frac{1}{2}}\left(\frac{\lambda}{2}\boldsymbol{I}+\left(1+\frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)\right)}{\lambda}\right)^{-1}\boldsymbol{P}_k\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{P}_k\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}} \\ &\preceq \left(1+\frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)\right)}{\lambda}\right)\left(\boldsymbol{P}_k\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{P}_k\right)^{\dagger}+\frac{2\boldsymbol{P}_k^{\perp}\boldsymbol{\Sigma}\boldsymbol{P}_k^{\perp}}{\lambda},\end{aligned}\tag{58}$$

where $\boldsymbol{X}=\boldsymbol{Z}\boldsymbol{\Sigma}^{\frac{1}{2}}$. Define the matrix $\boldsymbol{V}_k=\begin{bmatrix}\boldsymbol{v}_1 & \cdots & \boldsymbol{v}_k\end{bmatrix}\in\mathbb{R}^{d\times k}$, we can then write $\boldsymbol{P}_k=\boldsymbol{V}_k\boldsymbol{V}_k^{\mathsf{T}}$. Thus by exploiting the block matrix structure, it follows that

$$\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I}+\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\|\leq\left(1+\frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)\right)}{\lambda}\right)\lambda_{\min}\left(\boldsymbol{V}_k^{\mathsf{T}}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{V}_k\right)^{-1}+\frac{2\sigma_k}{\lambda}.\tag{59}$$

Substituting $\boldsymbol{\theta}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}$ into Eq. (58), we also obtain

$$\begin{aligned}&\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I}+\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta} \\ &\leq\left(1+\frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)\right)}{\lambda}\right)\lambda_{\min}\left(\boldsymbol{V}_k^{\mathsf{T}}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{V}_k\right)^{-1}\left\|\boldsymbol{\theta}_{\leq k}\right\|^2+\frac{2\left\|\boldsymbol{\beta}_{>k}\right\|^2}{\lambda}.\end{aligned}\tag{60}$$

Part II: Lower bounding the smallest eigenvalue $\lambda_{\min}\left(\boldsymbol{V}_k^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{V}_k\right)$. The last step is then to provide a lower bound for the smallest eigenvalue of $\boldsymbol{V}_k^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{V}_k$. Consider $\tilde{\boldsymbol{Z}} = \begin{bmatrix}\tilde{\boldsymbol{z}}_1 & \cdots & \tilde{\boldsymbol{z}}_n\end{bmatrix}^\mathsf{T} \in \mathbb{R}^{n\times k}$ with

$$\tilde{\boldsymbol{z}}_i = \boldsymbol{V}_k^\mathsf{T}\boldsymbol{z}_i = \begin{bmatrix}\langle \boldsymbol{z}, \boldsymbol{v}_1\rangle \\ \vdots \\ \langle \boldsymbol{z}, \boldsymbol{v}_{k-1}\rangle\end{bmatrix}.$$

We therefore need to lower bound $\lambda_{\min}(\tilde{\boldsymbol{Z}}^\mathsf{T}\tilde{\boldsymbol{Z}})$ where $\tilde{\boldsymbol{Z}}$ has i.i.d. rows $\tilde{\boldsymbol{z}}_i$ in $\mathbb{R}^k$. An immediate consequence is that $\mathbb{E}[\tilde{\boldsymbol{z}}_i] = \boldsymbol{0}$ and $\mathrm{Var}(\tilde{\boldsymbol{z}}_i) = \boldsymbol{I}_k$. Moreover, for any unit vector $\boldsymbol{\varphi} \in \mathbb{R}^k$, we can apply Hanson-Wright (cf. Lemma 2.1) and deduce that for any $t \geq 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\langle\tilde{\boldsymbol{z}}_i, \boldsymbol{\varphi}\rangle^2 - 1\right| \geq t\right) &= \mathbb{P}\left(\left|\boldsymbol{z}_i^\mathsf{T}\boldsymbol{V}_k\boldsymbol{\varphi}\boldsymbol{\varphi}^\mathsf{T}\boldsymbol{V}_k^\mathsf{T}\boldsymbol{z}_i - \mathrm{Tr}\left(\boldsymbol{V}_k\boldsymbol{\varphi}\boldsymbol{\varphi}^\mathsf{T}\boldsymbol{V}_k^\mathsf{T}\right)\right| \geq t\right) \\ &\leq 2\exp\left\{-\Omega\left(\min\left\{\frac{t^2}{\mathsf{C}_{\boldsymbol{x}}^4\left\|\boldsymbol{V}_k\boldsymbol{\varphi}\boldsymbol{\varphi}^\mathsf{T}\boldsymbol{V}_k^\mathsf{T}\right\|_F^2}, \frac{t}{\mathsf{C}_{\boldsymbol{x}}^2\left\|\boldsymbol{V}_k\boldsymbol{\varphi}\boldsymbol{\varphi}^\mathsf{T}\boldsymbol{V}_k^\mathsf{T}\right\|}\right\}\right)\right\} \\ &= 2\exp\left(-\Omega_{\mathsf{C}_{\boldsymbol{x}}}\left(\min\left\{t^2, t\right\}\right)\right), \end{aligned} \tag{61}$$

where we use the fact that $\left\|\boldsymbol{V}_k\boldsymbol{\varphi}\boldsymbol{\varphi}^\mathsf{T}\boldsymbol{V}_k^\mathsf{T}\right\|_F = \left\|\boldsymbol{V}_k\boldsymbol{\varphi}\boldsymbol{\varphi}^\mathsf{T}\boldsymbol{V}_k^\mathsf{T}\right\| \leq 1$. Thus we can bound the fourth moment of $\langle\tilde{\boldsymbol{z}}_i, \boldsymbol{\varphi}\rangle$ by

$$\begin{aligned}\mathbb{E}\left[\langle\tilde{\boldsymbol{z}}_i, \boldsymbol{\varphi}\rangle^4\right] &= \int_0^\infty 2t\mathbb{P}\left(\langle\tilde{\boldsymbol{z}}_i, \boldsymbol{\varphi}\rangle^2 \geq t\right)\mathrm{d}t \\ &\leq 1 + \int_0^\infty 2(t+1)\mathbb{P}\left(\langle\tilde{\boldsymbol{z}}_i, \boldsymbol{\varphi}\rangle^2 \geq t+1\right)\mathrm{d}t \\ &\leq 1 + 4\int_0^\infty (t+1)\exp\left(-\Omega_{\mathsf{C}_{\boldsymbol{x}}}\left(\min\left\{t^2, t\right\}\right)\right)\mathrm{d}t = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}}(1).\end{aligned}$$

Clearly the above bound holds uniformly for all $\boldsymbol{\varphi} \in \mathbb{S}^{k-1}$ from the unit sphere in $\mathbb{R}^k$. Since the upper bound $\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}}(1)$ does not depend on k, we can appeal to [Yas14, Theorem 2.2] and obtain that with probability $1 - \mathcal{O}(n^{-D})$

$$\lambda_{\min}\left(n^{-1}\tilde{\boldsymbol{Z}}^\mathsf{T}\tilde{\boldsymbol{Z}}\right) \geq 1 - \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}}\left(\sqrt{\frac{k}{n}}\right) - \mathcal{O}_D\left(\sqrt{\frac{\log n}{n}}\right).$$

Therefore, if we choose $k = \lfloor \eta n \rfloor$ for some fixed η such that $\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}}(\sqrt{\eta}) \leq 1/4$, it holds for $n = \Omega_D(1)$ that

$$\lambda_{\min}\left(n^{-1}\tilde{\boldsymbol{Z}}^\mathsf{T}\tilde{\boldsymbol{Z}}\right) \geq 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2},$$

and we therefore conclude the proof by taking $k = \lfloor \eta n \rfloor$ as above and substituting into Eq. (59)

$$\begin{aligned}\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I} + \boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\| &\leq \frac{2}{n}\left(1 + \frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n)\right)}{\lambda}\right) + \frac{2\sigma_k}{\lambda} \\ &= \frac{2}{n}\left(1 + \frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n)\right)}{\lambda}\right),\end{aligned}$$

where in the last line we use the fact that $\mathsf{d}_{\boldsymbol{\Sigma}} \geq n$ and therefore $\sigma_k = \mathcal{O}(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n)/n)$.

Similarly for Eq. (60), we have

$$\begin{aligned}\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\Sigma}^{\frac{1}{2}}\left(\lambda\boldsymbol{I}+\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta} &\le \frac{2}{n}\left(1+\frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)\right)}{\lambda}\right)\|\boldsymbol{\theta}_{\le k}\|^2+\frac{2\|\boldsymbol{\beta}_{>k}\|^2}{\lambda}\\ &\le \frac{2}{n}\left(1+\frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)\right)}{\lambda}\right)\|\boldsymbol{\theta}_{\le n}\|^2+\frac{2\|\boldsymbol{\beta}_{>n}\|^2}{\lambda}\,,\end{aligned}$$

where in the last line we use the fact that for all $k+1\le i\le n$,

$$\langle\boldsymbol{\beta},\boldsymbol{v}_i\rangle^2=\sigma_i\langle\boldsymbol{\theta},\boldsymbol{v}_i\rangle^2\le\sigma_k\langle\boldsymbol{\theta},\boldsymbol{v}_i\rangle^2=\mathcal{O}(\mathsf{d}_{\boldsymbol{\Sigma}}\sigma_k\cdot\log n\log(\mathsf{d}_{\boldsymbol{\Sigma}}n)/n)\langle\boldsymbol{\theta},\boldsymbol{v}_i\rangle^2\,.$$

The proof is complete.

C.3 Proof of Lemma 7.3

We apply Hanson-Wright inequality in Lemma 2.1 and get

$$\mathbb{P}\left(\left|\boldsymbol{z}_k^{\mathsf{T}}\boldsymbol{B}_{k-1}\boldsymbol{z}_k-\mathsf{S}_{k-1}(\boldsymbol{I})\right|\ge t\mid\boldsymbol{B}_{k-1}\right)\le 2\exp\left\{-\Omega\left(\min\left\{\frac{t^2}{\mathsf{C}_{\boldsymbol{x}}^4\|\boldsymbol{B}_{k-1}\|_F^2},\frac{t}{\mathsf{C}_{\boldsymbol{x}}^2\|\boldsymbol{B}_{k-1}\|}\right\}\right)\right\}$$

and

$$\begin{aligned}&\mathbb{P}\left(\left|\boldsymbol{z}_k^{\mathsf{T}}\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}\boldsymbol{z}_k-\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{k-1}^2\right)\right|\ge t\mid\boldsymbol{B}_{k-1}\right)\\ &\le 2\exp\left\{-\Omega\left(\min\left\{\frac{t^2}{\mathsf{C}_{\boldsymbol{x}}^4\|\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}\|_F^2},\frac{t}{\mathsf{C}_{\boldsymbol{x}}^2\|\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}\|}\right\}\right)\right\}.\end{aligned}$$

In particular, on the event $\{T_F(\boldsymbol{Q})\ge k,T_F(\boldsymbol{I})\ge k\}$, we have

$$|\mathsf{S}_{k-1}(\boldsymbol{I})-\mathsf{R}_0(\boldsymbol{I})|\le\beta_1\le\frac{1}{4}\mathsf{R}_0(\boldsymbol{I})\,,\qquad|\mathsf{S}_{k-1}(\boldsymbol{Q})-\mathsf{R}_0(\boldsymbol{Q})|\le\beta_1\le\frac{1}{4}R_0(\boldsymbol{Q})\,,\qquad\|\boldsymbol{B}_{k-1}\|\le\gamma\,,$$

which further implies that

$$\|\boldsymbol{B}_{k-1}\|\le\|\boldsymbol{B}_{k-1}\|_F=\sqrt{\mathrm{Tr}\left(\boldsymbol{B}_{k-1}^2\right)}\le\sqrt{\|\boldsymbol{B}_{k-1}\|\cdot\mathsf{S}_{k-1}(\boldsymbol{I})}=\mathcal{O}\left(\sqrt{\gamma\mathsf{R}_0(\boldsymbol{I})}\right),$$

and

$$\begin{aligned}\|\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}\|\le\|\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}\|_F&=\sqrt{\mathrm{Tr}\left(\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}^2\boldsymbol{Q}\boldsymbol{B}_{k-1}\right)}\\ &\le\sqrt{\left\|\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{B}_{k-1}^2\boldsymbol{Q}^{\frac{1}{2}}\right\|\cdot\mathrm{Tr}\left(\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}\right)}\\ &\le\sqrt{\|\boldsymbol{B}_{k-1}\|^2\cdot\mathrm{Tr}\left(\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{B}_{k-1}^2\boldsymbol{Q}^{\frac{1}{2}}\right)}\\ &\overset{\text{(i)}}{\le}\sqrt{\|\boldsymbol{B}_{k-1}\|^3\cdot\mathsf{S}_{k-1}(\boldsymbol{Q})}=\mathcal{O}\left(\sqrt{\gamma^3\mathsf{R}_0(\boldsymbol{Q})}\right),\end{aligned}$$

Substituting the above bounds into the Hanson-Wright inequalities, we have conditioning on $H_k:=\{T_F(\boldsymbol{Q})\ge k,T_F(\boldsymbol{I})\ge k\}$ for some constant $\mathsf{C}=\mathsf{C}(\mathsf{C}_{\boldsymbol{x}},D)$ that

$$\exp\left\{-\Omega\left(\frac{\mathsf{C}\log n\cdot\sqrt{\gamma\mathsf{R}_0(\boldsymbol{I})}}{\mathsf{C}_{\boldsymbol{x}}^2\|\boldsymbol{B}_{k-1}\|}\right)\right\}=\mathcal{O}(n^{-D})\,,$$

$$\exp\left\{-\Omega\left(\frac{\mathsf{C}\log n\cdot\sqrt{\gamma^3\mathsf{R}_0(\boldsymbol{Q})}}{\mathsf{C}_{\boldsymbol{x}}^2\left\|\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}\right\|}\right)\right\}=\mathcal{O}(n^{-D})\,,$$

and therefore it holds with probability $1-\mathcal{O}(n^{-D})$ that

$$\begin{aligned}\left|\boldsymbol{z}_k^\mathsf{T}\boldsymbol{B}_{k-1}\boldsymbol{z}_k-\mathsf{S}_{k-1}(\boldsymbol{I})\right|&\le\mathsf{C}\log n\cdot\sqrt{\gamma\mathsf{R}_0(\boldsymbol{I})}=:\alpha_1\,,\\ \left|\boldsymbol{z}_k^\mathsf{T}\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}\boldsymbol{z}_k-\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{k-1}^2\right)\right|&\le\mathsf{C}\log n\cdot\sqrt{\gamma^3\mathsf{R}_0(\boldsymbol{Q})}=:\alpha_2\,.\end{aligned}$$

The Hanson-Wright inequalities also give the following upper bounds on the expectations conditioning on the tail event when $\|\boldsymbol{B}_{k-1}\|\le\gamma$. In particular, we would have

$$\begin{aligned}&\mathbb{E}_{k-1}\left[\left|\boldsymbol{z}_k^\mathsf{T}\boldsymbol{B}_{k-1}\boldsymbol{z}_k-\mathsf{S}_{k-1}(\boldsymbol{I})\right|\mathbb{I}\left\{\left|\boldsymbol{z}_k^\mathsf{T}\boldsymbol{B}_{k-1}\boldsymbol{z}_k-\mathsf{S}_{k-1}(\boldsymbol{I})\right|\ge\alpha_1\right\}\right]\mathbb{I}(H_k)\\
&=\int_{\mathsf{C}\log n\cdot\sqrt{\gamma\mathsf{R}_0(\boldsymbol{I})}}^{\infty}\mathbb{P}\left(\left|\boldsymbol{z}_k^\mathsf{T}\boldsymbol{B}_{k-1}\boldsymbol{z}_k-\mathsf{S}_{k-1}(\boldsymbol{I})\right|\ge t\mid\boldsymbol{B}_{k-1}\right)\mathbb{I}(H_k)\mathrm{d}t\\
&\le\int_{\mathsf{C}\log n\cdot\sqrt{\gamma\mathsf{R}_0(\boldsymbol{I})}}^{\infty}2\exp\left\{-\Omega\left(\min\left\{\frac{t^2}{\mathsf{C}_{\boldsymbol{x}}^4\left\|\boldsymbol{B}_{k-1}\right\|_F^2},\frac{t}{\mathsf{C}_{\boldsymbol{x}}^2\left\|\boldsymbol{B}_{k-1}\right\|}\right\}\right)\right\}\mathbb{I}(H_k)\mathrm{d}t\\
&\le\int_{\mathsf{C}\log n\cdot\sqrt{\gamma R_0(I)}}^{\infty}2\exp\left\{-\Omega\left(\frac{t}{\mathsf{C}_{\boldsymbol{x}}^2\left\|\boldsymbol{B}_{k-1}\right\|_F}\right)\right\}\mathbb{I}(H_k)\mathrm{d}t\\
&\le\mathcal{O}\left(\mathsf{C}_{\boldsymbol{x}}^2\left\|\boldsymbol{B}_{k-1}\right\|_F\right)\cdot\mathcal{O}(n^{-D})=\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}}\left(n^{-D}\cdot\sqrt{\gamma\mathsf{R}_0(\boldsymbol{I})}\right)\,.\end{aligned}$$

Similarly it also holds that

$$\begin{aligned}&\mathbb{E}_{k-1}\left[\left|\boldsymbol{z}_k^\mathsf{T}\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}\boldsymbol{z}_k-\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{k-1}^2\right)\right|\mathbb{I}\left\{\left|\boldsymbol{z}_k^\mathsf{T}\boldsymbol{B}_{k-1}\boldsymbol{Q}\boldsymbol{B}_{k-1}\boldsymbol{z}_k-\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{k-1}^2\right)\right|\ge\alpha_2\right\}\right]\mathbb{I}(H_k)\\
&=\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}}\left(n^{-D}\cdot\sqrt{\gamma^3\mathsf{R}_0(\boldsymbol{Q})}\right)\,.\end{aligned}$$

To finish the proof, we now only need to show $\|\boldsymbol{A}_k\|\le\gamma$ holds with probability $1-\mathcal{O}(n^{-D})$. We provide upper bounds for small $k\le n/2$ and large $k>n/2$ separately. Under the assumption $\beta_2\le\mu/2$, we make use of the fact that $H_k\subset F_{k-1}(\boldsymbol{Q})$ which enables us to derive

$$\left|\mu_k-\mu_\star(\lambda,\mu)+\frac{k}{1+\mathsf{R}_0(\boldsymbol{I})}\right|\le\beta_2\le\frac{\mu}{2}\,,\tag{62}$$

and thus for all $k\le n/2$,

$$\mu_k\ge\mu_\star(\lambda,\mu)-\frac{k}{1+\mathsf{R}_0(\boldsymbol{I})}-\frac{\mu}{2}=\frac{\mu}{2}+\frac{n-k}{1+\mathsf{R}_0(\boldsymbol{I})}\ge\frac{\mu_\star(\lambda,\mu)}{2}\,,$$

which in particular implies for $k\le n/2$ that

$$\|\boldsymbol{A}_k\|\le\left\|\boldsymbol{\Sigma}^{\frac12}\left(\lambda\boldsymbol{I}+\mu_k\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^{\frac12}\right\|\le\frac{2}{\mu_\star(\lambda,\mu)}\,.$$

On the other hand, if $k>n/2$, we can still deduce from Eq. (62) that $\mu_k\ge\mu/2>0$ and thus

$$\boldsymbol{A}_k=\boldsymbol{\Sigma}^{\frac12}\left(\lambda\boldsymbol{I}+\mu_k\boldsymbol{\Sigma}+\boldsymbol{X}_k^\mathsf{T}\boldsymbol{X}_k\right)^{-1}\boldsymbol{\Sigma}^{\frac12}\preceq\boldsymbol{\Sigma}^{\frac12}\left(\lambda\boldsymbol{I}+\boldsymbol{X}_k^\mathsf{T}\boldsymbol{X}_k\right)^{-1}\boldsymbol{\Sigma}^{\frac12}\,.$$

Applying Lemma 7.2, we obtain with probability $1-\mathcal{O}(n^{-D})$ for all $k > n/2$,

$$\|\boldsymbol{A}_k\| \le \left\| \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\lambda \boldsymbol{I} + \boldsymbol{X}_{\lceil n/2 \rceil}^{\mathsf{T}} \boldsymbol{X}_{\lceil n/2 \rceil} \right)^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}} \right\| = \frac{2}{n} \left(1 + \frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}} \sigma_{\lfloor \eta n \rfloor} \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n) \right)}{\lambda} \right),$$

for some $\eta = \eta(\mathsf{C}_{\boldsymbol{x}})$. Combine with the trivial bound $\|\boldsymbol{A}_k\| \le 1/\lambda$, we conclude that $\|\boldsymbol{A}_k\| \le \gamma$ with probability $1-\mathcal{O}(n^{-D})$, provided we take

$$\gamma = \min\left\{ \frac{2}{n} \left(1 + \frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\mathsf{d}_{\boldsymbol{\Sigma}} \sigma_{\lfloor \eta n \rfloor} \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n) \right)}{\lambda} \right) + \frac{2}{\mu_\star(\lambda,\mu)}, \frac{1}{\lambda} \right\}.$$

The proof is completed by noting that

$$p_{k,k}(T_F, T_F, \boldsymbol{Q}) - p_{k+1,k}(T_E, T_F, \boldsymbol{Q}) = \mathbb{P}(E_k(\boldsymbol{Q})^c; H_k) = \mathcal{O}(n^{-D}).$$

C.4 Proof of Lemma 7.4

We begin by noticing that

$$p_{k,k}(T_E, T_E, \boldsymbol{Q}) - p_{k,k}(T_F, T_E, \boldsymbol{Q}) = \mathbb{P}\big(T_E(\boldsymbol{Q}) \ge k, T_E(\boldsymbol{I}) \ge k; F_{k-1}^c(\boldsymbol{Q})\big). \tag{63}$$

We therefore need to control $|\mathsf{R}_{k-1}(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})|$ and $|\mathsf{S}_{k-1}(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})|$, as well as $|\mu_k - \overline{\mu}_k|$ and $\|\boldsymbol{B}_{k-1}\|$.

Part I: Decomposing into martingale part and bias part. Recall the calculations for Eq. (47), we have

$$\mathsf{R}_i(\boldsymbol{Q}) - \mathsf{R}_{i-1}(\boldsymbol{Q}) = -\frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\right)}{1+\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i} + \frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{A}_{i-1}\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}$$

Define the stopping time

$$\overline{T} = T_E(\boldsymbol{Q}) \wedge T_E(\boldsymbol{I}),$$

and on the event $\{T_E(\boldsymbol{Q}) \ge k, T_E(\boldsymbol{I}) \ge k\} = \{\overline{T} \ge k\}$, it holds

$$\mathsf{R}_{k-1}(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q}) = (\mathsf{R}_{k-1}(\boldsymbol{Q}) - R_0(\boldsymbol{Q}))\,\mathbb{I}\left\{\overline{T} \ge k\right\} = \sum_{i=1}^{k-1} (\mathsf{R}_i(\boldsymbol{Q}) - \mathsf{R}_{i-1}(\boldsymbol{Q}))\,\mathbb{I}\left\{\overline{T} \ge i+1\right\}.$$

For each of the summand, we can decompose it into two parts—the martingale difference part $D_i(\boldsymbol{Q}, \overline{T})$ and a bias part $B_i(\boldsymbol{Q}, \overline{T})$—to be specific, we can write

$$(\mathsf{R}_i(\boldsymbol{Q}) - \mathsf{R}_{i-1}(\boldsymbol{Q}))\,\mathbb{I}\left\{\overline{T} \ge i+1\right\} = D_i(\boldsymbol{Q}, \overline{T}) + B_i(\boldsymbol{Q}, \overline{T}),$$

where by setting $G_i := F_i(\boldsymbol{Q}) \cap F_i(\boldsymbol{I}) \in \mathcal{F}_i$ and $S_i := \left\{\overline{T} = i\right\} \cap G_{i-1} \in \mathcal{F}_i$, the explicit forms of D_i and B_i are (recall that $\mathbb{E}_i(\,\cdot\,) := \mathbb{E}(\,\cdot\,|\mathcal{F}_i)$):

$$\begin{aligned} D_i(\boldsymbol{Q}, \overline{T}) := &-\frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\right)}{1+\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i}\mathbb{I}\left\{\overline{T} \ge i+1\right\} - \frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}(S_i) \\ &+ \mathbb{E}_{i-1}\left[\frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\right)}{1+\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i}\mathbb{I}\left\{\overline{T} \ge i+1\right\} + \frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}(S_i)\right], \end{aligned} \tag{64}$$

and

$$B_i(\boldsymbol{Q},\overline{T}) := \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{A}_{i-1}\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}\left\{\overline{T}\geq i+1\right\} + \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}\left(S_i\right) \\ - \mathbb{E}_{i-1}\left[\frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\right)}{1+\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i}\mathbb{I}\left\{\overline{T}\geq i+1\right\} + \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}\left(S_i\right)\right]. \tag{65}$$

Since $\overline{T}$ is a stopping time, one can easily have that $D_i(\boldsymbol{Q},\overline{T})$ is a martingale difference sequence for $i=0,1,\cdots,n$. (We note in passing that the above decomposition is similar but does not coincide with the standard Doob decomposition. In particular $B_i(\boldsymbol{Q},\overline{T})$ is not measurable on $\mathcal{F}_{i-1}$. We find the present decomposition more convenient.)

Part II: Controlling the martingale part. We will show $D_i(\boldsymbol{Q},\overline{T})$ is bounded and thus by concentration inequality for bounded martingale differences, we can obtain an upper bound for the sum of the $D_i(\boldsymbol{Q},\overline{T})$'s. To this end, we use the fact that if for some $m_{i-1}\in\mathcal{F}_{i-1}$

$$\left|D_i(\boldsymbol{Q},\overline{T}) - m_{i-1}\right| \leq M\,,$$

then $\left|D_i(\boldsymbol{Q},\overline{T})\right| = \left|D_i(\boldsymbol{Q},\overline{T}) - \mathbb{E}\left[D_i(\boldsymbol{Q},\overline{T})\mid\mathcal{F}_{i-1}\right]\right| \leq 2M$. Substitute the following $m_{i-1}\in\mathcal{F}_{i-1}$

$$m_{i-1} = -\frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}\left\{\overline{T}\geq i\right\}\mathbb{I}\left(G_{i-1}\right) + \mathbb{E}_{i-1}\left[\frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\right)}{1+\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i}\mathbb{I}\left\{\overline{T}\geq i+1\right\} + \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}\left(S_i\right)\right]$$

into the previous display, we have

$$\begin{aligned}
&\left|D_i(\boldsymbol{Q},\overline{T}) - m_{i-1}\right| \\
&\overset{\text{(i)}}{=} \left|\frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\right)}{1+\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i} - \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\right|\mathbb{I}\left\{\overline{T}\geq i+1\right\} \\
&\leq \left|\frac{\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i}\right|\mathbb{I}\left\{\overline{T}\geq i+1\right\} + \left|\frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)\cdot\left(\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right)}{\left(1+\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\right)\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)}\right|\mathbb{I}\left\{\overline{T}\geq i+1\right\},
\end{aligned}$$

where in (i) we use $\{\overline{T}\geq i+1\}\subset F_{i-1}(\boldsymbol{Q})\cap F_{i-1}(\boldsymbol{I}) = G_{i-1}$, and therefore

$$\mathbb{I}\left\{\overline{T}\geq i\right\}\mathbb{I}\left(G_{i-1}\right) = \mathbb{I}\left\{\overline{T}=i\right\}\mathbb{I}\left(G_{i-1}\right) + \mathbb{I}\left\{\overline{T}\geq i+1\right\}\mathbb{I}\left(G_{i-1}\right) = \mathbb{I}\left(S_i\right) + \mathbb{I}\left\{\overline{T}\geq i+1\right\}.$$

Recalling our assumptions for α_1, we observe that on the event $\{\overline{T}\geq i+1\}\subset E_i(\boldsymbol{Q})$,

$$\left|\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right| \leq \alpha_1 \leq \frac{1}{4}\mathsf{R}_0(\boldsymbol{I})\,, \qquad \left|\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)\right| \leq \alpha_2\,, \tag{66}$$

and on the event $\{\overline{T}\geq i+1\}\subset F_{i-1}(\boldsymbol{I})$ and $\{\overline{T}\geq i+1\}\subset F_{i-1}(\boldsymbol{Q})$ by assumptions on β_1,

$$\left|\mathsf{S}_{i-1}(\boldsymbol{I}) - \mathsf{R}_0(\boldsymbol{I})\right| \leq \beta_1 \leq \frac{1}{4}\mathsf{R}_0(\boldsymbol{I})\,, \qquad \left|\mathsf{S}_{i-1}(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})\right| \leq \beta_1 \leq \frac{1}{4}\mathsf{R}_0(\boldsymbol{Q})\,, \tag{67}$$

and finally on the event $\{\overline{T}\geq i+1\}\subset G_{i-1}$ it holds

$$\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right) \leq \|\boldsymbol{B}_{i-1}\|\cdot\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\right) \leq \gamma\mathsf{S}_{i-1}(\boldsymbol{Q}) = \mathcal{O}\left(\gamma\mathsf{R}_0(\boldsymbol{Q})\right). \tag{68}$$

Putting together bounds in Eqs. (66), (67), (68) and making use of the fact that $\left\{\overline{T} \geq i+1\right\} \subset E_i(\boldsymbol{Q}) \cap F_{i-1}(\boldsymbol{I}) \cap F_{i-1}(\boldsymbol{Q})$ yield

$$\left|D_i(\boldsymbol{Q}, \overline{T}) - m_{i-1}\right| \leq \left|\frac{\alpha_2}{1+\frac{1}{2}\mathsf{R}_0(\boldsymbol{I})}\right| + \left|\frac{\mathcal{O}(\gamma \mathsf{R}_0(\boldsymbol{Q})) \cdot \alpha_1}{\left(1+\frac{1}{2}\mathsf{R}_0(\boldsymbol{I})\right)\left(1+\frac{3}{4}\mathsf{R}_0(\boldsymbol{I})\right)}\right| = \mathcal{O}\left(\frac{\alpha_1 \gamma \mathsf{R}_0(\boldsymbol{Q}) + \alpha_2(1+\mathsf{R}_0(\boldsymbol{I}))}{1+\mathsf{R}_0(\boldsymbol{I})^2}\right).$$

Then we can apply Azuma-Hoeffding inequality and obtain

$$\max_{k \leq n}\left|\sum_{i=1}^{k} D_i(\boldsymbol{Q}, \overline{T})\right| = \mathcal{O}_D\left(\sqrt{n \log n} \cdot \frac{\alpha_1 \gamma \mathsf{R}_0(\boldsymbol{Q}) + \alpha_2(1+\mathsf{R}_0(\boldsymbol{I}))}{1+\mathsf{R}_0(\boldsymbol{I})^2}\right),$$

with probability $1 - \mathcal{O}(n^{-D})$.

Part III: Controlling the bias part. Now we proceed to bound the bias part $|B_i(\boldsymbol{Q}, \overline{T})|$ in Eq. (65). We can write an upper bound

$$\begin{aligned}\left|B_i(\boldsymbol{Q}, \overline{T})\right| \leq & \underbrace{\left|\frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{A}_{i-1}\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})} - \frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\right| \mathbb{I}\left\{\overline{T} \geq i+1\right\}}_{\text{(I)}} \\ & + \underbrace{\left|\frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})} \mathbb{I}\left\{\overline{T} \geq i\right\} \mathbb{I}\left(G_{i-1}\right) - \mathbb{E}_{i-1}\left[\frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\right)}{1+\boldsymbol{z}_i^{\mathsf{T}}\boldsymbol{B}_{i-1}\boldsymbol{z}_i} \mathbb{I}\left\{\overline{T} \geq i+1\right\} + \frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})} \mathbb{I}\left(S_i\right)\right]\right|}_{\text{(II)}}.\end{aligned}$$

Using the fact that

$$\boldsymbol{B}_{i-1} - \boldsymbol{A}_{i-1} = (\mu_{i-1} - \mu_i)\boldsymbol{B}_{i-1}\boldsymbol{A}_{i-1} = \frac{\boldsymbol{B}_{i-1}\boldsymbol{A}_{i-1}}{1+\mathsf{S}_i(\boldsymbol{I})},$$

we have

$$\text{(I)} \leq \left|\frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\left(\boldsymbol{A}_{i-1} - \boldsymbol{B}_{i-1}\right)\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\right| \mathbb{I}\left\{\overline{T} \geq i+1\right\} = \left|\frac{\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\boldsymbol{A}_{i-1}\right)}{\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)^2}\right| \mathbb{I}\left\{\overline{T} \geq i+1\right\}.$$

Upper bounding the term $\operatorname{Tr}(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\boldsymbol{A}_{i-1})$ requires more careful treatment. Note that $\boldsymbol{B}_{i-1}$ and $\boldsymbol{A}_{i-1}$ commute, as follows from the observation that

$$\begin{aligned}&\boldsymbol{B}_{i-1}\boldsymbol{A}_{i-1} \\ &= (1+\mathsf{S}_i(\boldsymbol{I})) \cdot \boldsymbol{\Sigma}^{\frac{1}{2}}\left\{\left(\lambda\boldsymbol{I} + \mu_i\boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^{\mathsf{T}}\boldsymbol{X}_{i-1}\right)^{-1} \cdot (\mu_{i-1} - \mu_i)\,\boldsymbol{\Sigma} \cdot \left(\lambda\boldsymbol{I} + \mu_{i-1}\boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^{\mathsf{T}}\boldsymbol{X}_{i-1}\right)^{-1}\right\}\boldsymbol{\Sigma}^{\frac{1}{2}} \\ &= (1+\mathsf{S}_i(\boldsymbol{I}))\left(\boldsymbol{B}_{i-1} - \boldsymbol{A}_{i-1}\right) \\ &= (1+\mathsf{S}_i(\boldsymbol{I})) \cdot \boldsymbol{\Sigma}^{\frac{1}{2}}\left\{\left(\lambda\boldsymbol{I} + \mu_{i-1}\boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^{\mathsf{T}}\boldsymbol{X}_{i-1}\right)^{-1} \cdot (\mu_{i-1} - \mu_i)\,\boldsymbol{\Sigma} \cdot \left(\lambda\boldsymbol{I} + \mu_i\boldsymbol{\Sigma} + \boldsymbol{X}_{i-1}^{\mathsf{T}}\boldsymbol{X}_{i-1}\right)^{-1}\right\}\boldsymbol{\Sigma}^{\frac{1}{2}} \\ &= \boldsymbol{A}_{i-1}\boldsymbol{B}_{i-1}.\end{aligned}$$

Since $\boldsymbol{A}_{i-1}$ and $\boldsymbol{B}_{i-1}$ are both p.s.d. compact self-adjoint operators in Hilbert space, commutativity implies they can be simultaneously orthogonally diagonalized, which further implies that $\boldsymbol{A}_{i-1}^{\frac{1}{2}}$ and $\boldsymbol{B}_{i-1}^{\frac{1}{2}}$ also commute. Therefore

$$\operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\boldsymbol{A}_{i-1}\right) = \operatorname{Tr}\left(\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{A}_{i-1}^{\frac{1}{2}}\boldsymbol{B}_{i-1}^2\boldsymbol{A}_{i-1}^{\frac{1}{2}}\boldsymbol{Q}^{\frac{1}{2}}\right) \leq \|\boldsymbol{B}_{i-1}\|^2 \cdot \operatorname{Tr}\left(\boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{A}_{i-1}\boldsymbol{Q}^{\frac{1}{2}}\right) = \|\boldsymbol{B}_{i-1}\|^2 \cdot \mathsf{R}_{i-1}(\boldsymbol{Q}),$$

and thus

$$\text{(I)} \le \frac{\|\boldsymbol{B}_{i-1}\|^2 \cdot \mathsf{R}_{i-1}(\boldsymbol{Q})}{(1+\mathsf{S}_{i-1}(\boldsymbol{I}))^2}\mathbb{I}\left\{\overline{T} \ge i+1\right\} = \mathcal{O}\left(\frac{\gamma^2 \mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^2}\right), \tag{69}$$

where in the last inequality we use Eq. (67) on the event $\{\overline{T} \ge i+1\} \subset F_{i-1}(\boldsymbol{I})$, while $\{\overline{T} \ge i+1\} \subset F_{i-1}(\boldsymbol{Q})$ also implies

$$|\mathsf{R}_{i-1}(\boldsymbol{Q}) - \mathsf{R}_0(\boldsymbol{Q})| \le \beta_1 \le \frac{1}{4}\mathsf{R}_0(\boldsymbol{Q}).$$

Next to bound (II), we make use of the fact that

$$\frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}\left\{\overline{T} \ge i\right\}\mathbb{I}\left(G_{i-1}\right) = \mathbb{E}_{i-1}\left[\frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}\left\{\overline{T} \ge i\right\}\mathbb{I}\left(G_{i-1}\right)\right],$$

and therefore

$$\begin{aligned}
&\text{(II)} \\
&= \left|\mathbb{E}_{i-1}\left[\frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}\left\{\overline{T} \ge i\right\}\mathbb{I}\left(G_{i-1}\right) - \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\right)}{1+\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i}\mathbb{I}\left\{\overline{T} \ge i+1\right\} - \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}\left(S_i\right)\right]\right| \\
&= \left|\mathbb{E}_{i-1}\left[\frac{\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i \cdot \left(\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right)}{\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)\left(1+\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\right)}\mathbb{I}\left\{\overline{T} \ge i+1\right\} + \frac{\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\mathbb{I}\left(S_i\right)\right]\right| \\
&\le \underbrace{\left|\mathbb{E}_{i-1}\left[\left(\frac{\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i \cdot \left(\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right)}{\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)\left(1+\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\right)} - \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)\cdot\left(\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right)}{\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)^2}\right)\mathbb{I}\left\{\overline{T} \ge i+1\right\}\right]\right|}_{\text{(III)}} \\
&\quad + \underbrace{\left|\mathbb{E}_{i-1}\left[\left(\frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)\cdot\left(\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right)}{\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)^2} + \frac{\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{1+\mathsf{S}_{i-1}(\boldsymbol{I})}\right)\mathbb{I}\left(S_i\right)\right]\right|}_{\text{(IV)}},
\end{aligned} \tag{70}$$

where in the last inequality we use that

$$\begin{aligned}
&\mathbb{E}_{i-1}\left[\frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)\cdot\left(\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right)}{\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)^2}\mathbb{I}\left\{\overline{T} \ge i\right\} \cap G_{i-1}\right] \\
&= \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)}{\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)^2}\mathbb{I}\big(\left\{\overline{T} \ge i\right\} \cap G_{i-1}\big)\cdot\mathbb{E}_{i-1}\left[\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right] = 0.
\end{aligned}$$

To control (III), we note that

$$\begin{aligned}
&\frac{\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i \cdot \left(\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right)}{\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)\left(1+\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\right)} - \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)\cdot\left(\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right)}{\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)^2} \\
&= \frac{\left\{\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i \cdot \left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right) - \mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)\cdot\left(1+\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\right)\right\}\cdot\left(\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right)}{\left(1+\mathsf{S}_{i-1}(\boldsymbol{I})\right)^2\left(1+\boldsymbol{z}_i^\mathsf{T}\boldsymbol{B}_{i-1}\boldsymbol{z}_i\right)}
\end{aligned}$$

$$= \frac{1}{(1+\mathsf{S}_{i-1}(\boldsymbol{I}))^2\left(1+\boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{z}_i\right)} \cdot \Big\{ \left(\boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)\right) \cdot (1+\mathsf{S}_{i-1}(\boldsymbol{I})) - \operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right) \cdot \left(\boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right) \Big\} \cdot \left(\boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I})\right) .$$

We again make use of the bounds in Eqs. (66), (67) and (68) on the event $\{\overline{T} \geq i+1\}$, which implies

$$\begin{aligned} \text{(III)} &\leq \left| \mathbb{E}_{i-1}\left[\frac{\left(\alpha_2 \cdot (1+\mathsf{S}_{i-1}(\boldsymbol{I})) + \operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right)\cdot\alpha_1\right)\cdot\alpha_1}{(1+\mathsf{S}_{i-1}(\boldsymbol{I}))^2\left(1+\boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{z}_i\right)} \mathbb{I}\left\{\overline{T} \geq i+1\right\}\right] \right| \\ &= \mathcal{O}\left(\frac{\alpha_1\alpha_2\left(1+\mathsf{R}_0(\boldsymbol{I})\right) + \alpha_1^2\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^3} \right) . \end{aligned} \tag{71}$$

Finally for term (IV) in Eq. (70), we can control it by

$$\begin{aligned} \text{(IV)} \leq{}& \frac{\mathcal{O}\left(\gamma\mathsf{R}_0(\boldsymbol{Q})\right)}{1+\mathsf{R}_0(\boldsymbol{I})^2} \cdot \mathbb{E}_{i-1}\left[\left| \boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I}) \right| \mathbb{I}\left(S_i\right) \right] \\ &+ \frac{1}{1+\mathsf{R}_0(\boldsymbol{I})} \cdot \mathbb{E}_{i-1}\left[\left| \boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right) \right| \mathbb{I}\left(S_i\right) \right] . \end{aligned}$$

Recall $S_i = \{\overline{T} = i\} \cap F_{i-1}(\boldsymbol{Q}) \cap F_{i-1}(\boldsymbol{I})$, which implies $\{T_F(\boldsymbol{Q}) \geq i, T_F(\boldsymbol{I}) \geq i\}$ holds but at least one of $E_i(\boldsymbol{Q})$ and $E_i(\boldsymbol{I})$ doesn't hold. This allows us to invoke Lemma 7.3 and conclude that $\mathbb{P}(S_i|\mathcal{F}_{i-1}) = \mathcal{O}(n^{-D})$. Moreover, we can further deduce from Lemma 7.3 that

$$\begin{aligned} &\mathbb{E}_{i-1}\left[\left| \boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I}) \right| \mathbb{I}\left(S_i\right) \right] \\ &\leq \alpha_1 \mathbb{P}(S_i|\mathcal{F}_{i-1}) + \mathbb{E}_{i-1}\left[\left| \boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I}) \right| \mathbb{I}\left\{ \left| \boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{z}_i - \mathsf{S}_{i-1}(\boldsymbol{I}) \right| \geq \alpha_1 \right\} \right] \mathbb{I}\{T_F(\boldsymbol{Q}) \geq i, T_F(\boldsymbol{I}) \geq i\} \\ &= \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-D}\cdot\alpha_1\right) . \end{aligned}$$

Similarly, we also have

$$\mathbb{E}_{i-1}\left[\left| \boldsymbol{z}_i^\top \boldsymbol{B}_{i-1}\boldsymbol{Q}\boldsymbol{B}_{i-1}\boldsymbol{z}_i - \operatorname{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{i-1}^2\right) \right| \mathbb{I}\left(S_i\right) \right] = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-D}\cdot\alpha_2\right) .$$

Combining the above displays, we obtain that

$$\text{(IV)} = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{n^{-D}\alpha_1\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^2} + \frac{n^{-D}\alpha_2}{1+\mathsf{R}_0(\boldsymbol{I})} \right)$$

Now applying the assumption that

$$n^{-D} = \mathcal{O}\left(\frac{\alpha_1}{1+\mathsf{R}_0(\boldsymbol{I})} \right) ,$$

we obtain

$$\text{(IV)} = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\alpha_1^2\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^3} + \frac{\alpha_1\alpha_2}{1+\mathsf{R}_0(\boldsymbol{I})^2} \right) = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\alpha_1\alpha_2\left(1+\mathsf{R}_0(\boldsymbol{I})\right) + \alpha_1^2\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^3} \right) . \tag{72}$$

Substitute Eqs. (71) and (72) into Eq. (70) we have

$$\text{(II)} = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\alpha_1\alpha_2\left(1+\mathsf{R}_0(\boldsymbol{I})\right) + \alpha_1^2\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^3} \right) ,$$

and together with Eq. (69) we obtain

$$\left| B_i(\boldsymbol{Q}, \overline{T}) \right| = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\gamma^2\mathsf{R}_0(\boldsymbol{Q}) + \alpha_1\alpha_2}{1+\mathsf{R}_0(\boldsymbol{I})^2} + \frac{\alpha_1^2\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^3} \right) .$$

Part IV: Combining the results. Hence, by combining results in part III and IV, we have with probability $1-\mathcal{O}(n^{-D})$ that

$$\begin{aligned}
&\left|(\mathsf{R}_{k-1}(\boldsymbol{Q})-\mathsf{R}_0(\boldsymbol{Q}))\,\mathbb{I}\left\{\overline{T}\geq k\right\}\right| \leq \left|\sum_{i=1}^{k-1} D_i(\boldsymbol{Q},\overline{T})\right| + \left|\sum_{i=1}^{k-1} B_i(\boldsymbol{Q},\overline{T})\right| \\
&\leq \mathcal{O}_D\left(\sqrt{n\log n}\cdot\frac{\alpha_1\gamma\mathsf{R}_0(\boldsymbol{Q})+\alpha_2(1+\mathsf{R}_0(\boldsymbol{I}))}{1+\mathsf{R}_0(\boldsymbol{I})^2}\right) + n\cdot\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\gamma^2\mathsf{R}_0(\boldsymbol{Q})+\alpha_1\alpha_2}{1+\mathsf{R}_0(\boldsymbol{I})^2}+\frac{\alpha_1^2\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^3}\right) \\
&= \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\sqrt{n\log n}\cdot\frac{\alpha_1\gamma\mathsf{R}_0(\boldsymbol{Q})+\alpha_2(1+\mathsf{R}_0(\boldsymbol{I}))}{1+\mathsf{R}_0(\boldsymbol{I})^2} + n\cdot\left\{\frac{\gamma^2\mathsf{R}_0(\boldsymbol{Q})+\alpha_1\alpha_2}{1+\mathsf{R}_0(\boldsymbol{I})^2}+\frac{\alpha_1^2\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^3}\right\}\right).
\end{aligned}$$

We can first see $\|\boldsymbol{B}_{k-1}\|\leq\gamma$ holds with probability $1-\mathcal{O}(n^{-D})$, which follows via exactly the same argument as in Appendix C.3 for $\|\boldsymbol{A}_k\|\leq\gamma$ by invoking Lemma 7.2. Moreover, we have on $\{\overline{T}\geq k\}$ that

$$\begin{aligned}
|\mathsf{S}_{k-1}(\boldsymbol{Q})-\mathsf{R}_{k-1}(\boldsymbol{Q})| &= |\mathrm{Tr}\left(\boldsymbol{Q}\left(\boldsymbol{B}_{k-1}-\boldsymbol{A}_{k-1}\right)\right)| = \frac{\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{B}_{k-1}\boldsymbol{A}_{k-1}\right)}{1+\mathsf{S}_{k-1}(\boldsymbol{I})} \\
&\leq \frac{\|\boldsymbol{B}_{k-1}\|\,\mathsf{R}_{k-1}(\boldsymbol{Q})}{1+\mathsf{S}_{k-1}(\boldsymbol{I})} \overset{\text{(i)}}{\leq} \frac{\gamma\mathsf{R}_{k-1}(\boldsymbol{Q})}{1+\mathsf{R}_{k-1}(\boldsymbol{I})} = \mathcal{O}\left(\frac{\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})}\right),
\end{aligned}$$

where in (i) we apply $\mu_{k-1}\geq\mu_k$ which indicates $\mathsf{R}_{k-1}(\boldsymbol{I})\leq\mathsf{S}_{k-1}(\boldsymbol{I})$. Therefore, by setting a constant $\mathsf{C}_\beta:=\mathsf{C}_\beta(\mathsf{C}_{\boldsymbol{x}},D)$ large enough and take

$$\begin{aligned}
\beta_1 &= \mathsf{C}_\beta\left(\sqrt{n\log n}\cdot\frac{\alpha_1\gamma\mathsf{R}_0(\boldsymbol{Q})+\alpha_2(1+\mathsf{R}_0(\boldsymbol{I}))}{1+\mathsf{R}_0(\boldsymbol{I})^2} + n\cdot\left\{\frac{\gamma^2\mathsf{R}_0(\boldsymbol{Q})+\alpha_1\alpha_2}{1+\mathsf{R}_0(\boldsymbol{I})^2}+\frac{\alpha_1^2\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})^3}\right\}+\frac{\gamma\mathsf{R}_0(\boldsymbol{Q})}{1+\mathsf{R}_0(\boldsymbol{I})}\right), \\
\beta_2 &= \frac{\mathsf{C}_\beta n\beta_1}{1+\mathsf{R}_0(\boldsymbol{I})^2},
\end{aligned}$$

if this satisfies the assumption $\beta_1\leq\mathsf{R}_0(\boldsymbol{I})/4$, we can conclude that with probability $1-\mathcal{O}(n^{-D})$,

$$\left|(\mathsf{R}_{k-1}(\boldsymbol{Q})-\mathsf{R}_0(\boldsymbol{Q}))\mathbb{I}\left\{\overline{T}\geq k\right\}\right|\leq\beta_1, \qquad \left|(\mathsf{S}_{k-1}(\boldsymbol{Q})-\mathsf{R}_0(\boldsymbol{Q}))\mathbb{I}\left\{\overline{T}\geq k\right\}\right|\leq\beta_1.$$

Further, since

$$\begin{aligned}
\left|(\mu_k-\overline{\mu}_k)\,\mathbb{I}\left\{\overline{T}\geq k\right\}\right| &= \left|\sum_{i=0}^{k-1}\left(\frac{1}{1+\mathsf{S}_i(\boldsymbol{I})}-\frac{1}{1+\mathsf{R}_0(\boldsymbol{I})}\right)\mathbb{I}\left\{\overline{T}\geq k\right\}\right| \\
&\leq \sum_{i=0}^{k-1}\frac{|\mathsf{S}_i(\boldsymbol{I})-\mathsf{R}_i(\boldsymbol{I})|}{(1+\mathsf{S}_i(\boldsymbol{I}))(1+\mathsf{R}_i(\boldsymbol{I}))}\mathbb{I}\left\{\overline{T}\geq k\right\} = \mathcal{O}\left(\frac{n\beta_1}{1+\mathsf{R}_0(\boldsymbol{I})^2}\right),
\end{aligned}$$

taking C_β large will guarantee $\left|(\mu_k-\overline{\mu}_k)\,\mathbb{I}\left\{\overline{T}\geq k\right\}\right|\leq\beta_2$. Combining the above displays, we see on the event $\{\overline{T}\geq k\}$, it holds with probability $1-\mathcal{O}(n^{-D})$ that

$$\max\{|\mathsf{R}_{k-1}(\boldsymbol{Q})-\mathsf{R}_0(\boldsymbol{Q})|,|\mathsf{S}_{k-1}(\boldsymbol{Q})-\mathsf{R}_0(\boldsymbol{Q})|\}\leq\beta_1, \qquad |\mu_k-\overline{\mu}_k|\leq\beta_2, \qquad \|\boldsymbol{B}_{k-1}\|\leq\gamma,$$

which is exactly the event $F_{k-1}(\boldsymbol{Q})$ (cf. Eq. (48b)). Substituting into Eq. (63) completes the proof.

C.5 Proof of Lemma 7.5

As $\beta_2 > \mu/2$, we cannot directly apply Lemmas 7.3 and 7.4. We will instead use a perturbation argument, reducing ourselves to the case $\beta_2 \le \mu/2$. We will define a second sequence μ_i' following the recursion Eq. (46) but with a different initialization $\mu_0' = \mu_\star(\lambda, \mu')$ with $\mu' := 64\beta_2 > \mu$. We use the notations

$$\boldsymbol{A}_i' := \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\lambda \boldsymbol{I} + \mu_i' \boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T} \boldsymbol{X}_i\right)^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}},$$
$$\boldsymbol{B}_i' := \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\lambda \boldsymbol{I} + \mu_{i+1}' \boldsymbol{\Sigma} + \boldsymbol{X}_i^\mathsf{T} \boldsymbol{X}_i\right)^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}},$$

and also denote by $\mathsf{R}_i'(\boldsymbol{Q}) = \mathscr{R}_i(\lambda, \mu_i'; \boldsymbol{Q}) := \mathrm{Tr}(\boldsymbol{Q}\boldsymbol{A}_i')$. For this second iteration, we define a parameter tuple $\Delta' = (\alpha_1', \alpha_2', \beta_1', \beta_2', \gamma)$ defined in Lemmas 7.3 and 7.4 as

$$\gamma' = \min\left\{\frac{2}{n}\left(1 + \frac{\mathsf{C}_\gamma \mathsf{d}_{\boldsymbol{\Sigma}} \sigma_{\lfloor \eta n \rfloor} \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n)}{\lambda}\right) + \frac{2}{\mu_\star(\lambda, \mu')}, \frac{1}{\lambda}\right\}, \tag{73a}$$

$$\alpha_1' = \mathsf{C}_\alpha \log n \cdot \sqrt{\gamma' \mathsf{R}_0'(\boldsymbol{I})}, \tag{73b}$$

$$\alpha_2' = \mathsf{C}_\alpha \log n \cdot \sqrt{\gamma'^3 \mathsf{R}_0'(\boldsymbol{Q})}, \tag{73c}$$

$$\beta_1' = \mathsf{C}_\beta \left(\sqrt{n \log n} \cdot \frac{\alpha_1' \gamma' \mathsf{R}_0'(\boldsymbol{Q}) + \alpha_2'(1 + \mathsf{R}_0'(\boldsymbol{I}))}{1 + \mathsf{R}_0'(\boldsymbol{I})^2} + n \cdot \left\{\frac{\gamma'^2 \mathsf{R}_0'(\boldsymbol{Q}) + \alpha_1' \alpha_2'}{1 + \mathsf{R}_0'(\boldsymbol{I})^2} + \frac{\alpha_1'^2 \gamma' \mathsf{R}_0'(\boldsymbol{Q})}{1 + \mathsf{R}_0'(\boldsymbol{I})^3}\right\} + \frac{\gamma' \mathsf{R}_0'(\boldsymbol{Q})}{1 + \mathsf{R}_0'(\boldsymbol{I})}\right), \tag{73d}$$

$$\beta_2' = \frac{\mathsf{C}_\beta n \beta_1'}{1 + \mathsf{R}_0'(\boldsymbol{I})^2}. \tag{73e}$$

We want to show $\alpha_1' \le \mathsf{R}_0'(\boldsymbol{I})/4$, $\beta_1' \le \mathsf{R}_0'(\boldsymbol{Q})/4$, $\beta_2' \le \mu'/2$ and $n^{-D} = \mathcal{O}\left(\alpha_1'/\left(1 + \mathsf{R}_0'(\boldsymbol{I})\right)\right)$ so that Lemmas 7.3 and 7.4 are valid for λ, μ' and Δ'. To prove this claim, we need the following result bounding the perturbation of $\mu_\star$.

Lemma C.3. *For any $\lambda > 0$ and $\mu \ge 0$,*

$$0 \le \frac{\partial \mu_\star(\lambda, \mu)}{\partial \mu} \le 1 + \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{I}).$$

Proof. Taking derivatives w.r.t. μ on both sides of

$$\mu = \mu_\star - \frac{n}{1 + \mathrm{Tr}(\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-1})},$$

we have

$$1 = \frac{\partial \mu_\star}{\partial \mu} \cdot \left(1 - \frac{(\mu_\star - \mu)\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-2}\right)}{1 + \mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-1}\right)}\right).$$

Further

$$1 - \frac{(\mu_\star - \mu)\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-2}\right)}{1 + \mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-1}\right)} \ge 1 - \frac{\mu_\star \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-2}\right)}{1 + \mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-1}\right)} = \frac{1 + \lambda \mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-2}\right)}{1 + \mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star \boldsymbol{\Sigma})^{-1}\right)}$$
$$\ge \frac{1}{1 + \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{I})},$$

which gives the desired bounds $0 \le \partial \mu_\star / \partial \mu \le 1 + \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{I})$. □

Recalling that $\mu_\star(\lambda,\mu)$ is increasing in μ and therefore $\mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{I})$ is decreasing in μ, as a direct consequence of Lemma C.3 we have

$$\begin{aligned}0 \le \mu_\star(\lambda,\mu') - \mu_\star(\lambda,\mu) &= \int_\mu^{\mu'} \frac{\partial \mu_\star}{\partial \nu}(\lambda,\nu)\,\mathrm{d}\nu \\ &\le \int_\mu^{\mu'} (1+\mathscr{R}_0(\lambda,\mu_\star(\lambda,\nu);\boldsymbol{I}))\mathrm{d}\nu \le (\mu'-\mu)(1+\mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{I})) \le \mu'(1+\mathsf{R}_0(\boldsymbol{I}))\,, \end{aligned} \tag{74}$$

and further

$$0 \le \mathsf{R}_0(\boldsymbol{Q}) - \mathsf{R}_0'(\boldsymbol{Q}) = (\mu_\star(\lambda,\mu') - \mu_\star(\lambda,\mu))\mathrm{Tr}(\boldsymbol{Q}\boldsymbol{A}_0\boldsymbol{A}_0') \le \gamma\mu'(1+\mathsf{R}_0(\boldsymbol{I}))\mathsf{R}_0'(\boldsymbol{Q})\,,$$

where in the last inequality we used $\|\boldsymbol{A}_0\| \le \min\{1/\mu_\star(\lambda,\mu), 1/\lambda\} \le \gamma$. Substituting $\mu' = 64\beta_2$ and using the condition $\gamma\beta_2(1+\mathsf{R}_0(\boldsymbol{I})) \le 1/64$, it then follows that

$$\frac{1}{2}\mathsf{R}_0(\boldsymbol{Q}) \le \mathsf{R}_0'(\boldsymbol{Q}) \le \mathsf{R}_0(\boldsymbol{Q})\,, \qquad \forall \text{ p.s.d. } \boldsymbol{Q}\,.$$

Using the last inequalities in Eqs. (73a) to (73c), it follows immediately that

$$\gamma' \le \gamma\,, \qquad \alpha_1' \le \alpha_1\,, \qquad \alpha_2' \le \alpha_2\,.$$

We then first see that $\alpha_1 \le \mathsf{R}_0(\boldsymbol{I})/8$ implies $\alpha_1' \le \alpha_1 \le \mathsf{R}_0'(\boldsymbol{I})/4$. For β_1' and β_2', using $1+\mathsf{R}_0'(\boldsymbol{I})^k \ge 2^{-k}(1+\mathsf{R}_0(\boldsymbol{I}))$ with $k=1,2,3$, we can deduce from Eqs. (73d) and (73e) that

$$\beta_1' \le 8\beta_1\,, \qquad \beta_2' \le \frac{4\mathsf{C}_\beta n\beta_1'}{1+\mathsf{R}_0(\boldsymbol{I})^2} \le 32\beta_2\,.$$

The last inequality verifies $\beta_2' \le 32\beta_2 = \mu'/2$. The condition $\beta_1 \le \mathsf{R}_0(\boldsymbol{Q})/64$ implies $\beta_1' \le 8\beta_1 \le \mathsf{R}_0(\boldsymbol{Q})/8 \le \mathsf{R}_0'(\boldsymbol{Q})/4$.

Finally we need to show $n^{-D} = \mathcal{O}\left(\alpha_1'/\left(1+\mathsf{R}_0'(\boldsymbol{I})\right)\right)$. From Eq. (74), we can obtain that

$$\mu_\star(\lambda,\mu') \le \mu_\star(\lambda,\mu) + 64\beta_2(1+\mathsf{R}_0(\boldsymbol{I})) \le \mu_\star(\lambda,\mu) + \frac{1}{\gamma}\,.$$

Recalling that $\gamma \le 2/\mu_\star(\lambda,\mu)$, we then know $\mu_\star(\lambda,\mu') \le 3/\gamma$ and thus

$$\gamma' \ge \min\left\{\frac{2}{\mu_\star(\lambda,\mu')}, \frac{1}{\lambda}\right\} \ge \frac{2}{3}\gamma\,.$$

Together with $\mathsf{R}_0(\boldsymbol{I}) = \Theta(\mathsf{R}_0'(\boldsymbol{I}))$, we then show $\alpha_1 = \mathcal{O}(\alpha_1')$ and further that $n^{-D} = \mathcal{O}\left(\alpha_1/\left(1+\mathsf{R}_0(\boldsymbol{I})\right)\right) = \mathcal{O}\left(\alpha_1'/\left(1+\mathsf{R}_0'(\boldsymbol{I})\right)\right)$. Hence, we can apply Lemmas 7.3 and 7.4, and by Eq. (51)

$$\left|\mathscr{R}_n(\lambda,\mu';\boldsymbol{Q}) - \mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu');\boldsymbol{Q})\right| = \mathcal{O}(\gamma'\beta_2'\mathsf{R}_0'(\boldsymbol{Q}) + \beta_1') = \mathcal{O}(\gamma\beta_2\mathsf{R}_0(\boldsymbol{Q}) + \beta_1)\,. \tag{75}$$

In order to finish the perturbation argument, we bound

$$\begin{aligned}\left|\mathscr{R}_n(\lambda,\mu;\boldsymbol{Q}) - \mathscr{R}_n(\lambda,\mu';\boldsymbol{Q})\right| &\le \left|(\mu'-\mu)\cdot\mathrm{Tr}\left(\boldsymbol{Q}\boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda\boldsymbol{I}+\mu\boldsymbol{\Sigma}+\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}(\lambda\boldsymbol{I}+\mu'\boldsymbol{\Sigma}+\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right)\right| \\ &\le \mu'\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\lambda\boldsymbol{I}+\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}\right\|\mathscr{R}_n(\lambda,\mu';\boldsymbol{Q}) \\ &\overset{\text{(i)}}{=} \mathcal{O}\left(\gamma\beta_2\left(\mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu');\boldsymbol{Q}) + \left|\mathscr{R}_n(\lambda,\mu';\boldsymbol{Q}) - \mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu');\boldsymbol{Q})\right|\right)\right)\end{aligned}$$

$$= \mathcal{O}\left(\gamma\beta_2\left(\mathsf{R}_0(\boldsymbol{Q}) + \gamma\beta_2\mathsf{R}_0(\boldsymbol{Q}) + \beta_1\right)\right)$$
$$= \mathcal{O}\left(\gamma\beta_2\mathsf{R}_0(\boldsymbol{Q})\right), \tag{76}$$

where in (i) we apply Lemma 7.2 and in the last line we use $\beta_1 = \mathcal{O}(\mathsf{R}_0(\boldsymbol{Q}))$ and $\gamma\beta_2 = \mathcal{O}((1+\mathsf{R}_0(\boldsymbol{I}))^{-1}) = \mathcal{O}(1)$. Similarly, invoke Lemma C.3 and we have

$$\begin{aligned}\left|\mathscr{R}_0(\lambda, \mu_\star(\lambda,\mu'); \boldsymbol{Q}) - \mathscr{R}_0(\lambda, \mu_\star(\lambda,\mu); \boldsymbol{Q})\right| &\le (\mu_\star(\lambda,\mu) - \mu_\star(\lambda,\mu'))\gamma\mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{Q}) \\ &\le \mu'(1+\mathsf{R}_0(\boldsymbol{I}))\gamma\mathsf{R}_0(\boldsymbol{Q}) \\ &= \mathcal{O}\left(\gamma\beta_2(1+\mathsf{R}_0(\boldsymbol{I}))\mathsf{R}_0(\boldsymbol{Q})\right). \end{aligned} \tag{77}$$

By triangular inequality, we deduce from Eqs. (75), (76) and (77) that

$$\begin{aligned}&|\mathscr{R}_n(\lambda,\mu;\boldsymbol{Q}) - \mathscr{R}_0(\lambda,\mu_\star(\lambda,\mu);\boldsymbol{Q})| \\ &= \mathcal{O}\left(\gamma\beta_2\mathsf{R}_0(\boldsymbol{Q}) + \beta_1\right) + \mathcal{O}\left(\gamma\beta_2\mathsf{R}_0(\boldsymbol{Q})\right) + \mathcal{O}\left(\gamma\beta_2(1+\mathsf{R}_0(\boldsymbol{I}))\mathsf{R}_0(\boldsymbol{Q})\right) \\ &= \mathcal{O}\left(\gamma\beta_2\left(1+\mathsf{R}_0(\boldsymbol{I})\right)\mathsf{R}_0(\boldsymbol{Q}) + \beta_1\right).\end{aligned}$$

C.6 Proof of Corollary 6.5

We first derive upper bounds for the parameter $\Delta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma)$ in Theorem 5. Since $\mu_\star(\lambda,0) \le \mu_\star(\lambda,\mu) \le (1-\kappa/2)^{-1}\mu_\star(\lambda,0)$, we have

$$\frac{n}{1+\mathsf{R}_0(\boldsymbol{I})} \le \mu_\star(\lambda,\mu) \le \frac{1}{1-\kappa/2}\mu_\star(\lambda,0) = \frac{1}{1-\kappa/2}\cdot\frac{n}{1+\mathscr{R}_0(\lambda,\mu_\star(\lambda,0);\boldsymbol{I})} \le \frac{1}{1-\kappa/2}\cdot\frac{n}{1+\mathsf{R}_0(\boldsymbol{I})}.$$

On the other hand, by Eq. (21) and the fact that $\mu_\star(\lambda,0) = \lambda/\lambda_\star$, we know

$$\frac{1}{1+\kappa^{-1}} \le \kappa \le \frac{\lambda}{n\lambda_\star} \le \frac{1}{1+\mathsf{R}_0(\boldsymbol{I})} \le \frac{1}{1-\kappa/2}\frac{\lambda}{n\lambda_\star} = \frac{1-\kappa}{1-\kappa/2} \le \frac{1}{1+\kappa/2},$$

which implies $\kappa/2 \le \mathsf{R}_0(\boldsymbol{I}) \le \kappa^{-1}$. Generalizing the definition of Eq. (23) to $\mu > 0$ and arbitrary $\boldsymbol{Q}$, we let $\rho := \mathsf{R}_0(\boldsymbol{Q})/\mathsf{R}_0(\boldsymbol{I}) \in (0,1]$.

Upper bound for γ. First we notice that $\mathsf{d}_{\boldsymbol{\Sigma}} \ge n$ by Assumption 1 and therefore

$$\log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n) \le \log(\mathsf{d}_{\boldsymbol{\Sigma}}) \cdot \log(\mathsf{d}_{\boldsymbol{\Sigma}}^2) = \mathcal{O}\left(\log^2(\mathsf{d}_{\boldsymbol{\Sigma}})\right),$$

which yields

$$\gamma = \mathcal{O}\left(\frac{2}{n}\left\{1 + \frac{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\sigma_{\lfloor\eta n\rfloor}\mathsf{d}_{\boldsymbol{\Sigma}}\log^2(\mathsf{d}_{\boldsymbol{\Sigma}})\right)}{\lambda}\right\} + \frac{2}{\mu_\star(\lambda,\mu)}\right).$$

Since $\mu_\star(\lambda,\mu) \ge \mu_\star(\lambda,0) \ge n\kappa$, we can write

$$\gamma = \frac{1}{n\kappa}\cdot\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(1 + \frac{\sigma_{\lfloor\eta n\rfloor}\mathsf{d}_{\boldsymbol{\Sigma}}\log^2(\mathsf{d}_{\boldsymbol{\Sigma}})}{\lambda}\right) = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\chi_n(\lambda)}{n\kappa}\right). \tag{78}$$

Upper bounds for α_1 and α_2. We know that $\mathsf{R}_0(\boldsymbol{I}) \le \kappa^{-1}$ and $\mathsf{R}_0(\boldsymbol{Q}) = \rho\mathsf{R}_0(\boldsymbol{I})$. As a consequence, we have

$$\alpha_1 = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\log n\cdot\sqrt{\frac{\gamma}{\kappa}}\right), \qquad \alpha_2 = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\log n\cdot\gamma\sqrt{\frac{\gamma\rho}{\kappa}}\right). \tag{79}$$

Upper bounds for β_1 and β_2. Using the bounds in the previous displays, we can write

$$\begin{aligned}
\beta_1 &= \mathsf{C}_\beta \left(\sqrt{n \log n} \cdot \frac{\alpha_1 \gamma \mathsf{R}_0(\boldsymbol{Q}) + \alpha_2 (1 + \mathsf{R}_0(\boldsymbol{I}))}{1 + \mathsf{R}_0(\boldsymbol{I})^2} + n \cdot \left\{ \frac{\gamma^2 \mathsf{R}_0(\boldsymbol{Q}) + \alpha_1 \alpha_2}{1 + \mathsf{R}_0(\boldsymbol{I})^2} + \frac{\alpha_1^2 \gamma \mathsf{R}_0(\boldsymbol{Q})}{1 + \mathsf{R}_0(\boldsymbol{I})^3} \right\} + \frac{\gamma \mathsf{R}_0(\boldsymbol{Q})}{1 + \mathsf{R}_0(\boldsymbol{I})} \right) \\
&= \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\sqrt{n \log n} \cdot (\alpha_1 \gamma \rho + \alpha_2) + n \cdot \left\{ (\gamma^2 \rho + \alpha_1 \alpha_2) + \alpha_1^2 \gamma \rho \right\} + \gamma \rho \right) \\
&= \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\sqrt{n \log n} \cdot \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\log n \cdot \gamma \sqrt{\frac{\gamma \rho}{\kappa}} \right) + n \cdot \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\log^2 n \cdot \gamma^2 \sqrt{\frac{\rho}{\kappa}} \right) + \gamma \rho \right) \\
&= \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\sqrt{n (\log n)^3 \gamma^3} + n (\log n)^2 \gamma^2 + \gamma \sqrt{\kappa} \right) \cdot \sqrt{\frac{\rho}{\kappa}} \, .
\end{aligned}$$

Substituting in Eq. (78), we can further bound

$$\beta_1 = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\frac{\sqrt{\rho} \chi_n(\lambda)^2 \log^2 n}{n \kappa^{2.5}} \right) . \tag{80}$$

As for β_2, we simply bound it by

$$\beta_2 = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(n \beta_1 \right) = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\frac{\sqrt{\rho} \chi_n(\lambda)^2 \log^2 n}{\kappa^{2.5}} \right) . \tag{81}$$

Upper bound for resolvent approximation. Recall the approximation bound we have in Theorem 5,

$$\begin{aligned}
|\mathscr{R}_n(\lambda, \mu; \boldsymbol{Q}) - \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q})| &= \mathcal{O} \left(\gamma \beta_2 \left(1 + \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{I}) \right) \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q}) + \beta_1 \right) \\
&= \mathcal{O} \left(\gamma \beta_2 \kappa^{-1} \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q}) + \beta_1 \right) ,
\end{aligned}$$

because $\mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{I}) = \mathsf{R}_0(\boldsymbol{I}) \le \kappa^{-1}$. And thus

$$\begin{aligned}
|\mathscr{R}_n(\lambda, \mu; \boldsymbol{Q}) - \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q})| &= \mathcal{O} \left(\gamma \beta_2 \kappa^{-1} \cdot \rho \kappa^{-1} + \beta_1 \right) \\
&= \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\frac{\sqrt{\rho^3} \chi_n(\lambda)^3 \log^2 n}{n \kappa^{5.5}} \right) + \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\frac{\sqrt{\rho} \chi_n(\lambda)^2 \log^2 n}{n \kappa^{2.5}} \right) \\
&= \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\frac{\sqrt{\rho} \chi_n(\lambda)^3 \log^2 n}{n \kappa^{5.5}} \right) .
\end{aligned}$$

As $\mathsf{R}_0(\boldsymbol{Q}) = \rho \mathsf{R}_0(\boldsymbol{I}) \ge \rho \kappa / 2$, we can also write

$$|\mathscr{R}_n(\lambda, \mu; \boldsymbol{Q}) - \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q})| = \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}, D} \left(\frac{\chi_n(\lambda)^3 \log^2 n}{n \sqrt{\rho} \cdot \kappa^{6.5}} \right) \mathscr{R}_0(\lambda, \mu_\star(\lambda, \mu); \boldsymbol{Q}) \, .$$

Simplifying the conditions. Finally we conclude the proof by simplifying the conditions $\alpha_1 \le \mathsf{R}_0(\boldsymbol{I})/8$, $\beta_1 \le \mathsf{R}_0(\boldsymbol{Q})/64$, $\gamma \beta_2 (1 + \mathsf{R}_0(\boldsymbol{I})) \le 1/64$ and $n^{-D} = \mathcal{O}(\alpha_1 / (1 + \mathsf{R}_0(\boldsymbol{I})))$. As $\mathsf{R}_0(\boldsymbol{I}) \ge \kappa/2$, by Eq. (79) it is sufficient to have the first condition once

$$\frac{\chi_n(\lambda) \log^2 n}{\kappa^4} \le \mathsf{C} n \, ,$$

for some sufficiently small constant $\mathsf{C} = \mathsf{C}(\mathsf{C}_{\boldsymbol{x}}, D)$. Recall that $\mathsf{R}_0(\boldsymbol{Q}) \geq \rho\kappa/2$. Therefore, by Eq. (80), the second requirement can be deduced from

$$\frac{\chi_n(\lambda)^2 \log^2 n}{\kappa^{3.5}} \leq \mathsf{C}' n\sqrt{\rho}\,,$$

for some sufficiently small constant $\mathsf{C}' = \mathsf{C}'(\mathsf{C}_{\boldsymbol{x}}, D)$. By Eqs. (78) and (81), we can derive $\gamma\beta_2(1 + \mathsf{R}_0(\boldsymbol{I})) \leq 1/64$ from

$$\frac{\chi_n(\lambda)^3 \log^2 n}{\kappa^{4.5}} \leq \mathsf{C}'' n/\sqrt{\rho}\,,$$

for some constant $\mathsf{C}'' = \mathsf{C}''(\mathsf{C}_{\boldsymbol{x}}, D)$. For the last condition, we need a lower bound for $\alpha_1 = \mathsf{C}_\alpha \log n \cdot \sqrt{\gamma \mathsf{R}_0(\boldsymbol{I})}$. As $\mu_\star(\lambda, \mu) \leq (1 - \kappa/2)^{-1}\mu_\star(\lambda, 0) \leq (1 - \kappa/2)^{-1} n \leq 2n$, it follows that

$$\begin{aligned}\gamma &= \min\left\{\frac{2}{n}\left(1 + \frac{\mathsf{C}_\gamma \mathsf{d}_{\boldsymbol{\Sigma}} \sigma_{\lfloor \eta n \rfloor} \cdot \log n \log(\mathsf{d}_{\boldsymbol{\Sigma}} n)}{\lambda}\right) + \frac{2}{\mu_\star(\lambda, \mu)}, \frac{1}{\lambda}\right\} \\ &= \Omega\left(\min\left\{\frac{1}{\mu_\star(\lambda, \mu)}, \frac{1}{\lambda}\right\}\right) = \Omega\left(\min\left\{\frac{1}{n}, \frac{1}{\lambda}\right\}\right).\end{aligned}$$

With $\kappa/2 \leq \mathsf{R}_0(\boldsymbol{I}) \leq \kappa^{-1}$, we obtain

$$\frac{\alpha_1}{1 + \mathsf{R}_0(\boldsymbol{I})} = \Omega\left(\sqrt{\frac{\kappa^3 \log^2 n}{\max\{n, \lambda\}}}\right).$$

It is then sufficient to have

$$n^{-D} = \mathcal{O}\left(\sqrt{\frac{\kappa^3 \log^2 n}{\max\{n, \lambda\}}}\right).$$

D Proof of Theorem 2

To apply triangle inequalities

$$\begin{aligned}|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| &\leq |\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| + |\mathscr{V}_{\boldsymbol{X}}(0) - \mathscr{V}_{\boldsymbol{X}}(\lambda)| + |\mathsf{V}_n(0) - \mathsf{V}_n(\lambda)|\,, \\ |\mathscr{B}_{\boldsymbol{X}}(0) - \mathsf{B}_n(0)| &\leq |\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| + |\mathscr{B}_{\boldsymbol{X}}(0) - \mathscr{B}_{\boldsymbol{X}}(\lambda)| + |\mathsf{B}_n(0) - \mathsf{B}_n(\lambda)|\,,\end{aligned} \tag{82}$$

we define $\lambda = \kappa n \lambda_\star$ and bound each term separately. By homogeneity, we will assume $\|\boldsymbol{\theta}\| = 1$ throughout the proof.

Part I: Bounding $|\mathscr{V}_{\boldsymbol{X}}(0) - \mathscr{V}_{\boldsymbol{X}}(\lambda)|$ and $|\mathscr{B}_{\boldsymbol{X}}(0) - \mathscr{B}_{\boldsymbol{X}}(\lambda)|$. Assume $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ has rank r with eigendecomposition $\boldsymbol{X}^\mathsf{T}\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\mathsf{T}$ where $\boldsymbol{U} \in \mathbb{R}^{d \times r}$ has orthonormal columns and $\boldsymbol{D}$ is a diagonal matrix with entries $s_1 \geq \cdots \geq s_r > 0$. Note that $s_r = n s_{\mathsf{min}}$.

For the variance term, by the elementary inequality $|1/x - x/(x + \lambda)^2| \leq 2\lambda/x^2$ for all $x, \lambda > 0$, we have by Eq. (4a)

$$\begin{aligned}|\mathscr{V}_{\boldsymbol{X}}(0) - \mathscr{V}_{\boldsymbol{X}}(\lambda)| &= \left|\tau^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}\boldsymbol{U}\left(\boldsymbol{D}^{-1} - \boldsymbol{D}(\boldsymbol{D} + \lambda\boldsymbol{I})^{-2}\right)\boldsymbol{U}^\mathsf{T}\right)\right| \\ &\leq \tau^2 \mathrm{Tr}\left(\frac{2\lambda}{s_r} \cdot \boldsymbol{\Sigma}\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^\mathsf{T}\right) = \frac{2\kappa\lambda_\star(\lambda)}{s_{\mathsf{min}}} \cdot \mathscr{V}_{\boldsymbol{X}}(0)\,,\end{aligned} \tag{83}$$

where in the last equality we use $\lambda = \kappa n \lambda_\star(\lambda)$ and $s_r = n s_{\mathsf{min}}$. The next lemma bounds the difference between $\lambda_\star(0)$ and $\lambda_\star(\lambda)$.

Lemma D.1. *Under the assumptions of Theorem 2, for λ such that $\lambda = \kappa n \lambda_\star(\lambda)$ it holds that*

$$\lambda_\star(0) \le \lambda_\star(\lambda) \le \left(1 + \frac{2\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\right) \lambda_\star(0) \le 2\lambda_\star(0)\,.$$

Proof. Since

$$\lambda = \lambda_\star \cdot \left(n - \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-1}\right)\right)\,,$$

we can compute that

$$\begin{aligned}\frac{\partial \lambda}{\partial \lambda_\star} &= \lambda_\star \cdot \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right) + n - \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-1}\right) = n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star(\lambda) \boldsymbol{I})^{-2}\right) \\ &\ge n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star(0) \boldsymbol{I})^{-2}\right) \ge \mathsf{C}_{\boldsymbol{\Sigma}} n\,,\end{aligned}$$

and thus

$$\lambda_\star(\lambda) = \lambda_\star(0) + \int_0^\lambda \frac{\partial \lambda_\star(\lambda)}{\partial \lambda} \mathrm{d}\lambda \le \lambda_\star(0) + \frac{\lambda}{\mathsf{C}_{\boldsymbol{\Sigma}} n} = \lambda_\star(0) + \frac{\lambda}{\mathsf{C}_{\boldsymbol{\Sigma}} n \lambda_\star(\lambda)} \cdot \lambda_\star(\lambda) = \lambda_\star(0) + \frac{\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}} \cdot \lambda_\star(\lambda)\,.$$

Rearranging terms, using $\kappa \le \mathsf{C}_{\boldsymbol{\Sigma}}^2/8 \le \mathsf{C}_{\boldsymbol{\Sigma}}/2$ and the fact that $(1-x)^{-1} \le 1 + 2x$ for $0 \le x \le 1/2$ conclude the proof. □

Returning to the bound of the variance term, we can thus further derive the upper bound

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathscr{V}_{\boldsymbol{X}}(\lambda)| \le \frac{4\kappa\lambda_\star(0)}{s_{\mathsf{min}}} \cdot \mathscr{V}_{\boldsymbol{X}}(0)\,.$$

Using the fact that $\kappa \le s_{\mathsf{min}}/(8\lambda_\star(0))$, we further have

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathscr{V}_{\boldsymbol{X}}(\lambda)| \le \left(1 - \frac{4\kappa\lambda_\star(0)}{s_{\mathsf{min}}}\right)^{-1} \cdot \frac{4\kappa\lambda_\star(0)}{s_{\mathsf{min}}} \cdot \mathscr{V}_{\boldsymbol{X}}(\lambda) \le \frac{8\kappa\lambda_\star(0)}{s_{\mathsf{min}}} \cdot \mathscr{V}_{\boldsymbol{X}}(\lambda)\,. \tag{84}$$

Now we look at the bias term. From Eq. (4b), we first have

$$\begin{aligned}\mathscr{B}_{\boldsymbol{X}}(0) &= \lim_{\lambda \downarrow 0} \lambda^2 \mathrm{Tr}\left(\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\right) \\ &= \lim_{\lambda \downarrow 0} \left\|(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}/\lambda + \boldsymbol{I})^{-1}\boldsymbol{\beta}\right\|_{\boldsymbol{\Sigma}}^2 \\ &= \lim_{\lambda \downarrow 0} \left\|\left(\boldsymbol{I} + \boldsymbol{U}\left((\boldsymbol{D}/\lambda + \boldsymbol{I})^{-1} - \boldsymbol{I}\right)\boldsymbol{U}^\mathsf{T}\right)\boldsymbol{\beta}\right\|_{\boldsymbol{\Sigma}}^2 = \left\|\left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\mathsf{T}\right)\boldsymbol{\beta}\right\|_{\boldsymbol{\Sigma}}^2\,.\end{aligned}$$

By triangle inequality, it thus follows

$$\begin{aligned}\left|\mathscr{B}_{\boldsymbol{X}}(0)^{\frac{1}{2}} - \mathscr{B}_{\boldsymbol{X}}(\lambda)^{\frac{1}{2}}\right| &\le \left\|\boldsymbol{U}(\boldsymbol{D}/\lambda + \boldsymbol{I})^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{\beta}\right\|_{\boldsymbol{\Sigma}} \le \frac{\lambda}{\lambda + s_r}\|\boldsymbol{\beta}\| = \frac{\kappa\lambda_\star(\lambda)}{\kappa\lambda_\star(\lambda) + s_{\mathsf{min}}}\|\boldsymbol{\beta}\| \\ &\le \frac{2\kappa\lambda_\star(0)}{2\kappa\lambda_\star(0) + s_{\mathsf{min}}}\|\boldsymbol{\beta}\|\,,\end{aligned}$$

where in the last line we invoke Lemma D.1 and use $\lambda_\star(\lambda) \le 2\lambda_\star(0)$.

Additionally with $\mathscr{B}_{\boldsymbol{X}}(0) \le \|\boldsymbol{\beta}\|^2$ and

$$\mathscr{B}_{\boldsymbol{X}}(\lambda) = \lambda^2 \mathrm{Tr}\left(\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\right) \le \mathrm{Tr}\left(\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X}/\lambda + \boldsymbol{I})^{-2}\right) \le \mathrm{Tr}\left(\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}\right) = \|\boldsymbol{\beta}\|^2\,,$$

we conclude that

$$|\mathscr{B}_{\boldsymbol{X}}(0) - \mathscr{B}_{\boldsymbol{X}}(\lambda)| = \left|\mathscr{B}_{\boldsymbol{X}}(0)^{\frac{1}{2}} + \mathscr{B}_{\boldsymbol{X}}(\lambda)^{\frac{1}{2}}\right| \cdot \left|\mathscr{B}_{\boldsymbol{X}}(0)^{\frac{1}{2}} - \mathscr{B}_{\boldsymbol{X}}(\lambda)^{\frac{1}{2}}\right| \leq \frac{4\kappa\lambda_\star(0)\left\|\boldsymbol{\beta}\right\|^2}{2\kappa\lambda_\star(0) + s_{\min}} \leq \frac{4\kappa\lambda_\star(0)\left\|\boldsymbol{\beta}\right\|^2}{s_{\min}}\,. \tag{85}$$

We obtain an alternative upper bound for $|\mathscr{B}_{\boldsymbol{X}}(0) - \mathscr{B}_{\boldsymbol{X}}(\lambda)|$ in the following way. Note that

$$\begin{aligned}
&\left|\mathscr{B}_{\boldsymbol{X}}(0)^{\frac{1}{2}} - \mathscr{B}_{\boldsymbol{X}}(\lambda)^{\frac{1}{2}}\right| \leq \left\|\boldsymbol{U}(\boldsymbol{D}/\lambda + \boldsymbol{I})^{-1}\boldsymbol{U}^\mathsf{T}\boldsymbol{\beta}\right\|_{\boldsymbol{\Sigma}} = \lambda\left\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-1}\boldsymbol{X}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\right\| \\
&= \lambda\sqrt{\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{1/2}\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-1}\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-1}\boldsymbol{X}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}} \\
&\leq \lambda\sqrt{\left\|(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-\frac{1}{2}}\right\|} \cdot \sqrt{\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}} \\
&= \lambda\sqrt{\left\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\|} \cdot \sqrt{\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}}\,.
\end{aligned}$$

We next apply Lemma 7.2, which implies that with probability $1 - \mathcal{O}(n^{-D})$

$$\begin{aligned}
\left|\mathscr{B}_{\boldsymbol{X}}(0)^{\frac{1}{2}} - \mathscr{B}_{\boldsymbol{X}}(\lambda)^{\frac{1}{2}}\right| &\leq \sqrt{\frac{1}{n}\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\lambda\chi_n'(\kappa))} \cdot \sqrt{\frac{1}{n}\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\lambda\chi_n'(\kappa))\left\|\boldsymbol{\theta}_{\leq n}\right\|^2 + 2\left\|\boldsymbol{\beta}_{>n}\right\|^2} \\
&= \sqrt{\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa^2\lambda_\star(0)^2\chi_n'(\kappa)^2)\left\|\boldsymbol{\theta}_{\leq n}\right\|^2 + \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa\lambda_\star(0)\chi_n'(\kappa))\left\|\boldsymbol{\beta}_{>n}\right\|^2}\,.
\end{aligned}$$

Using the same argument, we can also bound

$$\begin{aligned}
\mathscr{B}_{\boldsymbol{X}}(\lambda) &= \lambda^2\mathrm{Tr}\left(\boldsymbol{\beta}\boldsymbol{\beta}^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\right) \\
&\leq \lambda^2\left\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^{1/2}\right\| \cdot \boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{1/2}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} \\
&= \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa^2\lambda_\star(0)^2\chi_n'(\kappa)^2)\left\|\boldsymbol{\theta}_{\leq n}\right\|^2 + \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa\lambda_\star(0)\chi_n'(\kappa))\left\|\boldsymbol{\beta}_{>n}\right\|^2\,,
\end{aligned}$$

and we can therefore conclude that

$$\begin{aligned}
|\mathscr{B}_{\boldsymbol{X}}(0) - \mathscr{B}_{\boldsymbol{X}}(\lambda)| &= \left|\mathscr{B}_{\boldsymbol{X}}(0)^{\frac{1}{2}} + \mathscr{B}_{\boldsymbol{X}}(\lambda)^{\frac{1}{2}}\right| \cdot \left|\mathscr{B}_{\boldsymbol{X}}(0)^{\frac{1}{2}} - \mathscr{B}_{\boldsymbol{X}}(\lambda)^{\frac{1}{2}}\right| \\
&\leq \left(\left|\mathscr{B}_{\boldsymbol{X}}(0)^{\frac{1}{2}} - \mathscr{B}_{\boldsymbol{X}}(\lambda)^{\frac{1}{2}}\right| + 2\mathscr{B}_{\boldsymbol{X}}(\lambda)^{\frac{1}{2}}\right) \cdot \left|\mathscr{B}_{\boldsymbol{X}}(0)^{\frac{1}{2}} - \mathscr{B}_{\boldsymbol{X}}(\lambda)^{\frac{1}{2}}\right| \\
&= \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa^2\lambda_\star(0)^2\chi_n'(\kappa)^2)\left\|\boldsymbol{\theta}_{\leq n}\right\|^2 + \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa\lambda_\star(0)\chi_n'(\kappa))\left\|\boldsymbol{\beta}_{>n}\right\|^2\,.
\end{aligned} \tag{86}$$

Combining Eqs. (85) and (86), we finally have

$$\begin{aligned}
&|\mathscr{B}_{\boldsymbol{X}}(0) - \mathscr{B}_{\boldsymbol{X}}(\lambda)| \\
&= \min\left\{\mathcal{O}\left(\frac{\kappa\lambda_\star(0)\left\|\boldsymbol{\beta}\right\|^2}{s_{\min}}\right), \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa^2\lambda_\star(0)^2\chi_n'(\kappa)^2)\left\|\boldsymbol{\theta}_{\leq n}\right\|^2 + \mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa\lambda_\star(0)\chi_n'(\kappa))\left\|\boldsymbol{\beta}_{>n}\right\|^2\right\}\,.
\end{aligned} \tag{87}$$

Part II: Bounding $|\mathsf{V}_n(0) - \mathsf{V}_n(\lambda)|$ and $|\mathsf{B}_n(0) - \mathsf{B}_n(\lambda)|$. Note that

$$0 \geq \frac{\partial\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star\boldsymbol{I})^{-2}\right)}{\partial\lambda_\star} = -2\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star\boldsymbol{I})^{-3}\right) \geq -\frac{2}{\lambda_\star(0)}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star(0)\boldsymbol{I})^{-2}\right)\,,$$

we can apply Lemma D.1 and obtain

$$\begin{aligned}\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right) &\geq \operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)\\ &\geq \operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right) - \frac{2\left(\lambda_\star(\lambda)-\lambda_\star(0)\right)}{\lambda_\star(0)}\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)\\ &\geq \left(1-\frac{4\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\right)\cdot \operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)\,.\end{aligned}$$

We then have

$$\begin{aligned}\mathsf{V}_n(0)\geq \mathsf{V}_n(\lambda) &= \frac{n-\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)}{n-\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)}\cdot\frac{\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)}{\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)}\cdot\mathsf{V}_n(0)\\ &\overset{\text{(i)}}{\geq} \frac{\mathsf{C}_{\boldsymbol{\Sigma}}n}{\mathsf{C}_{\boldsymbol{\Sigma}}n+\frac{4\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\cdot\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)}\cdot\left(1-\frac{4\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\right)\cdot\mathsf{V}_n(0)\\ &\overset{\text{(ii)}}{\geq} \frac{\mathsf{C}_{\boldsymbol{\Sigma}}^2}{\mathsf{C}_{\boldsymbol{\Sigma}}^2+4\kappa}\cdot\left(1-\frac{4\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\right)\cdot\mathsf{V}_n(0)\,,\end{aligned}$$

where we use $n-\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)\geq \mathsf{C}_{\boldsymbol{\Sigma}}n$ in (i) and $n\geq \operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)$ in (ii). By the elementary inequality $1-(1-a)(1-b)\leq a+b$ for all $0\leq a,b\leq 1$, we can thus derive that

$$\begin{aligned}|\mathsf{V}_n(0)-\mathsf{V}_n(\lambda)| &\leq \left\{1-\left(1-\frac{4\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2+4\kappa}\right)\cdot\left(1-\frac{4\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\right)\right\}\cdot\mathsf{V}_n(0)\\ &\leq \left(\frac{4\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2+4\kappa}+\frac{4\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\right)\cdot\mathsf{V}_n(0)\leq\frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}\cdot\mathsf{V}_n(0)\,.\end{aligned}\tag{88}$$

For the bias term, we first similarly derive

$$\boldsymbol{\beta}^\mathsf{T}\left(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}\geq\boldsymbol{\beta}^\mathsf{T}\left(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}\geq\left(1-\frac{4\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\right)\cdot\boldsymbol{\beta}^\mathsf{T}\left(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}\,.$$

Note that

$$\begin{aligned}\left|\frac{\mathsf{B}_n(\lambda)}{\mathsf{B}_n(0)}-1\right| &= \left|\frac{n-\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)}{n-\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)}\cdot\frac{\boldsymbol{\beta}^\mathsf{T}\left(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}}{\boldsymbol{\beta}^\mathsf{T}\left(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}}\cdot\frac{\lambda_\star(\lambda)^2}{\lambda_\star(0)^2}-1\right|\\ &\leq \max\left\{1-\frac{n-\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)}{n-\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)}\cdot\frac{\boldsymbol{\beta}^\mathsf{T}\left(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}}{\boldsymbol{\beta}^\mathsf{T}\left(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}};\frac{\lambda_\star(\lambda)^2}{\lambda_\star(0)^2}-1\right\}\,,,\end{aligned}$$

From the previous calculations for the variance term, we know

$$1-\frac{n-\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)}{n-\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)}\cdot\frac{\boldsymbol{\beta}^\mathsf{T}\left(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}}{\boldsymbol{\beta}^\mathsf{T}\left(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I}\right)^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}}\leq\frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}\,,$$

and by Lemma D.1 we have

$$\frac{\lambda_\star(\lambda)^2}{\lambda_\star(0)^2}-1\leq\left(1+\frac{2\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\right)^2-1\leq\frac{4\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}+\frac{2\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\cdot\frac{2\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\leq\frac{6\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}}\,.$$

In the last inequality, recall $\kappa\leq\mathsf{C}_{\boldsymbol{\Sigma}}^2/8\leq\mathsf{C}_{\boldsymbol{\Sigma}}/2$. Putting together, we have error of the bias term bounded by

$$|\mathsf{B}_n(0)-\mathsf{B}_n(\lambda)|\leq\frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}\cdot\mathsf{B}_n(0)\,.\tag{89}$$

Part III: Variance approximation when $\lambda = 0$. Recalling that $\lambda = \kappa n \lambda_\star(\lambda)$, we want to invoke Theorem 1 to bound $|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)|$. Note that by Lemma D.1 it holds $\lambda_\star(\lambda) = \Theta(\lambda_\star(0))$ and thus

$$\begin{aligned}\chi_n(\lambda) &= 1 + \frac{\sigma_{\lfloor \eta n \rfloor} \mathsf{d}_{\boldsymbol{\Sigma}} \log^2(\mathsf{d}_{\boldsymbol{\Sigma}})}{\lambda} = 1 + \frac{\sigma_{\lfloor \eta n \rfloor} \mathsf{d}_{\boldsymbol{\Sigma}} \log^2(\mathsf{d}_{\boldsymbol{\Sigma}})}{\kappa n \lambda_\star(\lambda)} = \Theta\left(1 + \frac{\sigma_{\lfloor \eta n \rfloor} \mathsf{d}_{\boldsymbol{\Sigma}} \log^2(\mathsf{d}_{\boldsymbol{\Sigma}})}{\kappa n \lambda_\star(0)}\right) \\ &= \Theta(\chi_n'(\kappa))\,.\end{aligned}$$

Hence the conditions hold for Theorem 1 by taking $\mathsf{C}_1 = \Theta(\mathsf{C})$, and we have for some constant $\mathsf{C}' := \mathsf{C}'(k, \mathsf{C}_{\boldsymbol{x}}, D) > 0$,

$$|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| \le \mathsf{C}' \cdot \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}} \cdot \mathsf{V}_n(\lambda)\,.$$

Substituting the above display and Eqs. (84), (88) into Eq. (82) yields

$$\begin{aligned}|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| &\le \mathsf{C}' \cdot \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}} \cdot \mathsf{V}_n(\lambda) + \frac{8\kappa\lambda_\star(0)}{s_{\min}} \cdot \mathscr{V}_{\boldsymbol{X}}(\lambda) + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2} \cdot \mathsf{V}_n(0) \\ &\le \mathsf{C}' \cdot \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}} \cdot \mathsf{V}_n(\lambda) + \frac{8\kappa\lambda_\star(0)}{s_{\min}} \cdot \left(1 + \mathsf{C}' \cdot \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}}\right) \mathsf{V}_n(\lambda) + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2} \cdot \mathsf{V}_n(0) \\ &= \left\{\left(1 + \frac{8\kappa\lambda_\star(0)}{s_{\min}}\right)\left(1 + \mathsf{C}' \cdot \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}}\right) - 1\right\} \cdot \mathsf{V}_n(\lambda) + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2} \cdot \mathsf{V}_n(0) \\ &\le \left\{\left(1 + \frac{8\kappa\lambda_\star(0)}{s_{\min}}\right)\left(1 + \mathsf{C}' \cdot \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}}\right) - 1\right\} \cdot \left(1 + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}\right) \mathsf{V}_n(0) + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2} \cdot \mathsf{V}_n(0) \\ &\le \left\{\left(1 + \frac{8\kappa\lambda_\star(0)}{s_{\min}}\right)\left(1 + \mathsf{C}' \cdot \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}}\right)\left(1 + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}\right) - 1\right\} \cdot \mathsf{V}_n(0) \\ &\le \left(\exp\left(\frac{8\kappa\lambda_\star(0)}{s_{\min}} + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2} + \mathsf{C}' \cdot \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}}\right) - 1\right) \cdot \mathsf{V}_n(0)\,.\end{aligned}$$

Since $\kappa \le s_{\min}/(8\lambda_\star(0))$ and $\kappa \le \mathsf{C}_{\boldsymbol{\Sigma}}^2/8$, if we additionally assume

$$\frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}} \le \frac{1}{\mathsf{C}'}\,,$$

we can then conclude that

$$\exp\left(\frac{8\kappa\lambda_\star(0)}{s_{\min}} + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2} + \mathsf{C}' \cdot \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}}\right) - 1 = \mathcal{O}\left(\frac{8\kappa\lambda_\star(0)}{s_{\min}} + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2} + \mathsf{C}' \cdot \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}}\right)\,,$$

and

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| = \mathcal{O}_{k, \mathsf{C}_{\boldsymbol{x}}, D}\left(\kappa \cdot \left(\frac{\lambda_\star(0)}{s_{\min}} + \frac{1}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}\right) + \frac{\chi_n'(\kappa)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}}\right) \cdot \mathsf{V}_n(0)\,.$$

We meet this assumption by setting $\mathsf{C}_2 = 1/\mathsf{C}'$.

Part IV: Bias approximation when $\lambda = 0$. To apply Theorem 1 when $\lambda = \kappa n \lambda_\star(\lambda)$, we first note that the condition $\lambda k n^{-\frac{1}{k}} \le n\kappa/2$ is equivalent to $\lambda_\star(\lambda) k n^{-\frac{1}{k}} \le 1/2$, and by Lemma D.1 it suffices to have $\lambda_\star(0) k n^{-\frac{1}{k}} \le 1/4$, which holds by assumption. Since we know $\chi_n'(\kappa) = \Theta(\chi_n(\lambda))$ from the previous part of the proof, we only need to additionally verify that $\lambda_\star(0) = \Theta(\lambda_\star(\lambda))$ and $\rho(0) = \Theta(\rho(\lambda))$. The first relation is a direct consequence of Lemma D.1, and for the second claim we observe that

$$\rho(\lambda) = \frac{\mathscr{R}_0(\lambda_\star(\lambda), 1; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T})}{\mathscr{R}_0(\lambda_\star(\lambda), 1; \boldsymbol{I})} = \frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\boldsymbol{\Sigma} + \lambda_\star(\lambda)\boldsymbol{I}\right)^{-1}\right)}{\mathrm{Tr}\left(\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma} + \lambda_\star(\lambda)\boldsymbol{I}\right)^{-1}\right)}\,.$$

As for any p.s.d. $\boldsymbol{Q}$,

$$0 \ge \frac{\partial \mathscr{R}_0(\lambda_\star, 1; \boldsymbol{Q})}{\partial \lambda_\star} = -\mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{Q} \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I}\right)^{-2}\right) \ge -\frac{1}{\lambda_\star(0)} \mathrm{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{Q} \boldsymbol{\Sigma}^{\frac{1}{2}} \left(\boldsymbol{\Sigma} + \lambda_\star(0)\boldsymbol{I}\right)^{-1}\right)\,,$$

we have

$$\mathscr{R}_0(\lambda_\star(0), 1; \boldsymbol{Q}) \ge \mathscr{R}_0(\lambda_\star(\lambda), 1; \boldsymbol{Q}) \ge \mathscr{R}_0(\lambda_\star(0), 1; \boldsymbol{Q}) - \frac{\lambda_\star(\lambda) - \lambda_\star(0)}{\lambda_\star(0)} \cdot \mathscr{R}_0(\lambda_\star(0), 1; \boldsymbol{Q})\,.$$

Therefore by Lemma D.1 and $\kappa \le \mathsf{C}_{\boldsymbol{\Sigma}}^2/8$, we can obtain

$$\left|\frac{\mathscr{R}_0(\lambda_\star(\lambda), 1; \boldsymbol{Q})}{\mathscr{R}_0(\lambda_\star(0), 1; \boldsymbol{Q})} - 1\right| \le \left|\frac{\lambda_\star(\lambda)}{\lambda_\star(0)} - 1\right| \le \frac{2\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}} \le \frac{\mathsf{C}_{\boldsymbol{\Sigma}}}{4} \le \frac{1}{4}\,,$$

which implies $\mathscr{R}_0(\lambda_\star(\lambda), 1; \boldsymbol{Q})/\mathscr{R}_0(\lambda_\star(0), 1; \boldsymbol{Q}) = \Theta(1)$ and therefore $\rho(0) = \Theta(\rho(\lambda))$. Now we are able to invoke Theorem 2, yielding for some constant $\mathsf{C}' := \mathsf{C}'(k, \mathsf{C}_{\boldsymbol{x}}, D) > 0$,

$$|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| \le \mathsf{C}' \cdot \left(\frac{\lambda_\star(0)^{k+1}}{n\kappa^3} + \frac{\chi_n'(\kappa)^3 \log^2 n}{\sqrt{\rho(0)} n^{1-\frac{1}{k}} \kappa^{8.5}}\right) \cdot \mathsf{B}_n(\lambda)\,.$$

Now we can substitute the above bound and Eq. (89) into Eq. (82),

$$\begin{aligned}
&|\mathscr{B}_{\boldsymbol{X}}(0) - \mathsf{B}_n(0)| \\
&\le \mathsf{C}' \cdot \left(\frac{\lambda_\star(0)^{k+1}}{n\kappa^3} + \frac{\chi_n'(\kappa)^3 \log^2 n}{\sqrt{\rho(0)} n^{1-\frac{1}{k}} \kappa^{8.5}}\right) \cdot \mathsf{B}_n(\lambda) + |\mathscr{B}_{\boldsymbol{X}}(0) - \mathscr{B}_{\boldsymbol{X}}(\lambda)| + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2} \cdot \mathsf{B}_n(0) \\
&\le \left\{\left(1 + \mathsf{C}' \cdot \left(\frac{\lambda_\star(0)^{k+1}}{n\kappa^3} + \frac{\chi_n'(\kappa)^3 \log^2 n}{\sqrt{\rho(0)} n^{1-\frac{1}{k}} \kappa^{8.5}}\right)\right)\left(1 + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}\right) - 1\right\} \cdot \mathsf{B}_n(0) + |\mathscr{B}_{\boldsymbol{X}}(0) - \mathscr{B}_{\boldsymbol{X}}(\lambda)| \\
&\le \left(\exp\left(\mathsf{C}' \cdot \left(\frac{\lambda_\star(0)^{k+1}}{n\kappa^3} + \frac{\chi_n'(\kappa)^3 \log^2 n}{\sqrt{\rho(0)} n^{1-\frac{1}{k}} \kappa^{8.5}}\right) + \frac{8\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}\right) - 1\right) \cdot \mathsf{B}_n(0) + |\mathscr{B}_{\boldsymbol{X}}(0) - \mathscr{B}_{\boldsymbol{X}}(\lambda)|\,.
\end{aligned}$$

Similar to previous calculations for the variance approximation, setting $\mathsf{C}_3 = 1/\mathsf{C}'$ and thus

$$\frac{\lambda_\star(0)^{k+1}}{n\kappa^3} + \frac{\chi_n'(\kappa)^3 \log^2 n}{\sqrt{\rho(0)} n^{1-\frac{1}{k}} \kappa^{8.5}} \le \frac{1}{\mathsf{C}'}\,.$$

Substituting in Eq. (87), it then holds that

$$|\mathscr{B}_{\boldsymbol{X}}(0) - \mathsf{B}_n(0)|$$

$$= \mathcal{O}_{k,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}+\frac{\lambda_\star(0)^{k+1}}{n\kappa^3}+\frac{\chi_n'(\kappa)^3\log^2 n}{\sqrt{\rho(0)}n^{1-\frac{1}{k}}\kappa^{8.5}}\right)\cdot \mathsf{B}_n(0)+|\mathscr{B}_{\boldsymbol{X}}(0)-\mathscr{B}_{\boldsymbol{X}}(\lambda)|$$

$$= \mathcal{O}_{k,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\kappa}{\mathsf{C}_{\boldsymbol{\Sigma}}^2}+\frac{\lambda_\star(0)^{k+1}}{n\kappa^3}+\frac{\chi_n'(\kappa)^3\log^2 n}{\sqrt{\rho(0)}n^{1-\frac{1}{k}}\kappa^{8.5}}\right)\cdot \mathsf{B}_n(0)$$

$$+\min\left\{\mathcal{O}\left(\frac{\kappa\lambda_\star(0)\,\|\boldsymbol{\beta}\|^2}{s_{\mathsf{min}}}\right),\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa^2\lambda_\star(0)^2\chi_n'(\kappa)^2)\,\|\boldsymbol{\theta}_{\le n}\|^2+\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}(\kappa\lambda_\star(0)\chi_n'(\kappa))\,\|\boldsymbol{\beta}_{>n}\|^2\right\}.$$

E Proof of Theorem 3

We follow the same proof strategy in Appendix D for the overparameterized regime, taking $\lambda=\varepsilon n$.

Part I: Bounding $|\mathscr{V}_{\boldsymbol{X}}(0)-\mathscr{V}_{\boldsymbol{X}}(\lambda)|$. As we assume $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ has rank d, we can write its eigendecomposition $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\mathsf{T}$ with $\boldsymbol{U}\in\mathbb{R}^{d\times d}$ an orthogonal matrix and $\boldsymbol{D}$ is a diagonal matrix with entries $s_1\ge\cdots\ge s_d>0$. In this case, $s_d=ns_{\mathsf{min}}$. Substitute $\lambda=\varepsilon n$ into Eq. (83) instead of $\lambda=\kappa n\lambda_\star(\lambda)$, we have

$$|\mathscr{V}_{\boldsymbol{X}}(0)-\mathscr{V}_{\boldsymbol{X}}(\lambda)|\le\frac{2\varepsilon}{s_{\mathsf{min}}}\cdot\mathscr{V}_{\boldsymbol{X}}(0)\,. \tag{90}$$

Part II: Bounding $|\mathsf{V}_n(0)-\mathsf{V}_n(\lambda)|$. Similar to the overparameterized case, we can control the growth of $\lambda_\star(\lambda)$ by

Lemma E.1. *Under the assumptions of Theorem 3, for λ such that $\lambda=\varepsilon n$ it holds that*

$$0=\lambda_\star(0)\le\lambda_\star(\lambda)\le\frac{\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}}\,.$$

Proof. By the proof of Lemma D.1, we have

$$\frac{\partial\lambda}{\partial\lambda_\star}=n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)\ge n-d\ge\mathsf{C}_{\boldsymbol{\Sigma}}n\,,$$

and thus

$$0\le\lambda_\star(0)\le\lambda_\star(\lambda)=\int_0^\lambda\frac{\partial\lambda_\star(\nu)}{\partial\lambda}\mathrm{d}\nu\le\frac{\lambda}{\mathsf{C}_{\boldsymbol{\Sigma}}n}=\frac{\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}}\,.$$

□

In this case, note that

$$0\ge\frac{\partial\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}{\partial\lambda_\star}=-2\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-3}\right)\ge-\frac{2}{\sigma_d}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)=-\frac{2d}{\sigma_d}\,,$$

we can apply Lemma E.1 and obtain for $\lambda_\star(0)=0$,

$$d=\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)\ge\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)\ge\left(1-\frac{2\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}\sigma_d}\right)\cdot d\,.$$

We then have

$$\mathsf{V}_n(0)\ge\mathsf{V}_n(\lambda)=\frac{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)}\cdot\frac{\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)}{\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-2}\right)}\cdot\mathsf{V}_n(0)$$

$$
\begin{aligned}
&= \frac{n-d}{n - \operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)} \cdot \frac{\operatorname{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-2}\right)}{d} \cdot \mathsf{V}_n(0) \\
&\geq \frac{\mathsf{C}_{\boldsymbol{\Sigma}} n}{\mathsf{C}_{\boldsymbol{\Sigma}} n + \frac{2\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}\sigma_d} \cdot d} \cdot \left(1 - \frac{2\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}\sigma_d}\right) \cdot \mathsf{V}_n(0) \\
&\geq \frac{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d + 2\varepsilon} \cdot \left(1 - \frac{2\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}\sigma_d}\right) \cdot \mathsf{V}_n(0)\,,
\end{aligned}
$$

where in the last line we use $n \geq d$. Again by the elementary inequality $1-(1-a)(1-b) \leq a+b$ for all $0 \leq a, b \leq 1$,

$$
\begin{aligned}
|\mathsf{V}_n(0) - \mathsf{V}_n(\lambda)| &\leq \left\{1 - \left(1 - \frac{2\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d + 2\varepsilon}\right) \cdot \left(1 - \frac{2\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}\sigma_d}\right)\right\} \cdot \mathsf{V}_n(0) \\
&\leq \left(\frac{2\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d + 2\varepsilon} + \frac{2\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}\sigma_d}\right) \cdot \mathsf{V}_n(0) \leq \frac{4\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d} \cdot \mathsf{V}_n(0)\,. \qquad (91)
\end{aligned}
$$

Part III: Variance approximation. Taking $\lambda = \varepsilon n$, we want to invoke Theorem 1 to bound $|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)|$. Using Lemma E.1, we know

$$
\mathsf{C}_{\boldsymbol{\Sigma}} \leq 1 - \frac{d}{n} \leq 1 - \frac{1}{n}\operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-1}\right) \leq 1 - \frac{d}{n} \cdot \frac{\sigma_d}{\sigma_d + \varepsilon/\mathsf{C}_{\boldsymbol{\Sigma}}} \leq 1 - \frac{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d}{\mathsf{C}_{\boldsymbol{\Sigma}}\sigma_d + \varepsilon}\,.
$$

Since by assumption $\varepsilon \leq \mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d/4 \leq \mathsf{C}_{\boldsymbol{\Sigma}}\sigma_d$, Eq. (21) holds with $\kappa = \mathsf{C}_{\boldsymbol{\Sigma}}/2$, because

$$
\mathsf{C}_{\boldsymbol{\Sigma}} \leq \frac{\lambda}{n\lambda_\star(\lambda)} = 1 - \frac{1}{n}\operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-1}\right) \leq 1 - \frac{\mathsf{C}_{\boldsymbol{\Sigma}}}{2}\,.
$$

Thus by Theorem 1, we have for some constant $\mathsf{C}' := \mathsf{C}'(k, \mathsf{C}_{\boldsymbol{x}}, D) > 0$,

$$
|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| \leq \mathsf{C}' \cdot \frac{\chi_n(\varepsilon n)^3 \log^2 n}{n^{1-\frac{1}{k}}\mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}} \cdot \mathsf{V}_n(\lambda)\,.
$$

Combining the above display with Eqs. (90), (91) yields

$$
\begin{aligned}
|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| &\leq \mathsf{C}' \cdot \frac{\chi_n(\varepsilon n)^3 \log^2 n}{n^{1-\frac{1}{k}}\mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}} \cdot \mathsf{V}_n(\lambda) + \frac{2\varepsilon}{s_{\min}} \cdot \mathscr{V}_{\boldsymbol{X}}(\lambda) + \frac{4\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d} \cdot \mathsf{V}_n(0) \\
&\leq \left\{\left(1 + \frac{2\varepsilon}{s_{\min}}\right)\left(1 + \mathsf{C}' \cdot \frac{\chi_n(\varepsilon n)^3 \log^2 n}{n^{1-\frac{1}{k}}\mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}}\right) - 1\right\} \cdot \mathsf{V}_n(\lambda) + \frac{4\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d} \cdot \mathsf{V}_n(0) \\
&\leq \left\{\left(1 + \frac{2\varepsilon}{s_{\min}}\right)\left(1 + \mathsf{C}' \cdot \frac{\chi_n(\varepsilon n)^3 \log^2 n}{n^{1-\frac{1}{k}}\mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}}\right)\left(1 + \frac{4\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d}\right) - 1\right\} \cdot \mathsf{V}_n(0) \\
&\leq \left(\exp\left(\frac{2\varepsilon}{s_{\min}} + \frac{4\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d} + \mathsf{C}' \cdot \frac{\chi_n(\varepsilon n)^3 \log^2 n}{n^{1-\frac{1}{k}}\mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}}\right) - 1\right) \cdot \mathsf{V}_n(0)\,.
\end{aligned}
$$

Since $\varepsilon \leq s_{\min}/2$ and $\varepsilon \leq \mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d/4$, if we additionally assume

$$
\frac{\chi_n(\varepsilon n)^3 \log^2 n}{n^{1-\frac{1}{k}}\mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}} \leq \frac{1}{\mathsf{C}'}\,,
$$

we can then conclude that

$$
\exp\left(\frac{2\varepsilon}{s_{\min}} + \frac{4\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d} + \mathsf{C}' \cdot \frac{\chi_n(\varepsilon n)^3 \log^2 n}{n^{1-\frac{1}{k}}\mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}}\right) - 1 = \mathcal{O}\left(\frac{\varepsilon}{s_{\min}} + \frac{\varepsilon}{\mathsf{C}_{\boldsymbol{\Sigma}}^2\sigma_d} + \mathsf{C}' \cdot \frac{\chi_n(\varepsilon n)^3 \log^2 n}{n^{1-\frac{1}{k}}\mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}}\right),
$$

and the proof is complete with $\mathsf{C}_2 = 1/\mathsf{C}'$.

F Proofs for proportional regime

F.1 Proof of Proposition 4.1

To apply Theorem 1, we first provide upper bounds for $\mathsf{d}_{\mathbf{\Sigma}}(n)$ and κ implying that Assumptions 1 and Eq. (21) hold. Throughout we use the shorthand $\lambda_{\mathsf{p}} = \lambda/n \in [1/M, M]$.

Lemma F.1. *Under Assumption 2 and $\lambda = n\lambda_{\mathsf{p}}$, Assumptions 1 and (21) hold for*

$$\begin{aligned} \mathsf{d}_{\mathbf{\Sigma}}(n) &= \mathcal{O}_M(n)\,, \\ \kappa &= \Omega_M(1)\,. \end{aligned}$$

For such $\mathsf{d}_{\mathbf{\Sigma}}$ and κ, $\chi_n(\lambda) = \mathcal{O}_{\lambda_{\mathsf{p}},M}(\log^2 n)$.

Proof. By Assumption 2 we know $d \le Mn$ and therefore for any $1 \le k \le \min\{n, d\}$,

$$\sum_{l=k}^{d} \sigma_l \le d\sigma_k \le Mn\sigma_k =: \mathsf{d}_{\mathbf{\Sigma}}\sigma_k\,.$$

Using $\lambda = n\lambda_{\mathsf{p}}$ into Eq. (5), we have

$$1 - \frac{\lambda_{\mathsf{p}}}{\lambda_\star} = \frac{1}{n}\mathrm{Tr}\left(\mathbf{\Sigma}(\mathbf{\Sigma} + \lambda_\star \boldsymbol{I})^{-1}\right)\,.$$

which implies

$$1 - \frac{\lambda_{\mathsf{p}}}{\lambda_\star} \le \frac{d}{n}\cdot\frac{1}{1+\lambda_\star} \le \frac{M}{\lambda_\star}\,.$$

This implies $\lambda_\star \le \lambda_{\mathsf{p}} + M$ and therefore

$$1 - \frac{\lambda}{n\lambda_\star} = 1 - \frac{\lambda_{\mathsf{p}}}{\lambda_\star} \le 1 - \frac{\lambda_{\mathsf{p}}}{\lambda_{\mathsf{p}} + M}\,.$$

On the other hand,

$$1 - \frac{\lambda}{n\lambda_\star} = 1 - \frac{\lambda_{\mathsf{p}}}{\lambda_\star} \ge \frac{d}{n}\cdot\frac{\sigma_d}{\sigma_d + \lambda_\star} \ge \frac{1}{M + M^2\lambda_\star} \ge \frac{1}{M + M^2\lambda_{\mathsf{p}} + M^3}\,,$$

and therefore we have

$$\kappa := \min\left\{\frac{\lambda_{\mathsf{p}}}{\lambda_{\mathsf{p}} + M}, \frac{1}{M + M^2\lambda_{\mathsf{p}} + M^3}\right\} = \Omega_M(1)\,.$$

Finally, we can bound $\chi_n(\lambda)$ as $\mathsf{d}_{\mathbf{\Sigma}} = \mathcal{O}_M(n)$, and thus

$$\chi_n(\lambda) = 1 + \frac{\sigma_{\lfloor \eta n\rfloor}\mathsf{d}_{\mathbf{\Sigma}}\log^2(\mathsf{d}_{\mathbf{\Sigma}})}{n\lambda_{\mathsf{p}}} = \mathcal{O}_M\left(1 + \frac{\log^2 n}{\lambda_{\mathsf{p}}}\right) = \mathcal{O}_M\left(\log^2 n\right)\,.$$

□

For any unit vector $\boldsymbol{u} \in \mathbb{R}^d$, since

$$\begin{aligned}
\mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{u}\boldsymbol{u}^\mathsf{T}) &= \lambda \mathrm{Tr}\left(\boldsymbol{u}\boldsymbol{u}^\mathsf{T}\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right) \\
&\geq \frac{\lambda}{\lambda M + \mu_\star(\lambda,0)} \mathrm{Tr}\left(\boldsymbol{u}\boldsymbol{u}^\mathsf{T}\right) = \frac{\lambda}{d(\lambda M + \mu_\star(\lambda,0))} \mathrm{Tr}(\boldsymbol{I}) \\
&\geq \frac{\lambda + \mu_\star(\lambda,0)}{d(\lambda M + \mu_\star(\lambda,0))} \cdot \lambda \mathrm{Tr}\left(\boldsymbol{\Sigma}(\lambda \boldsymbol{I} + \mu_\star(\lambda,0)\boldsymbol{\Sigma})^{-1}\right) \\
&\geq \frac{1}{dM} \mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{I}) \geq \frac{1}{nM^2} \mathscr{F}_0(\lambda, \mu_\star(\lambda,0); \boldsymbol{I}) \,, \qquad (92)
\end{aligned}$$

we have $\rho(\lambda) = \Omega_M(n^{-1})$. Together with Lemma F.1, since $n = \Omega_{M,\mathsf{C}_{\boldsymbol{x}},D}(1)$, the following conditions in Theorem 1 hold

$$\chi_n(\lambda)^3 \log^2 n \leq \mathsf{C} n \kappa^{4.5} \min\left\{1, \sqrt{\rho(\lambda)}\right\} , \qquad n^{-2D+1} = \mathcal{O}\left(\sqrt{\frac{\kappa^3 \log^2 n}{\max\{n, \lambda\}}}\right) .$$

Additionally, $\lambda k n^{-\frac{1}{k}} \leq n\kappa/2$ is equivalent to $\lambda_\mathsf{p} k n^{-\frac{1}{k}} \leq \kappa/2$, which holds for $n = \Omega_{k,M}(1)$. Finally, by using $\lambda_\star(\lambda) \leq \lambda_\mathsf{p} + M = \mathcal{O}_M(1)$, as shown above, we can conclude from Theorem 1 and Lemma F.1 that, for $n = \Omega_{k,M,\mathsf{C}_{\boldsymbol{x}},D}(1)$, with probability $1 - \mathcal{O}_k(n^{-D+1})$,

$$\begin{aligned}
|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| &= \mathcal{O}_{k,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\chi_n(\lambda)^3 \log^2 n}{n^{1-\frac{1}{k}}\kappa^{9.5}}\right) \cdot \mathsf{V}_n(\lambda) = \mathcal{O}_{k,M,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\log^8 n}{n^{1-\frac{1}{k}}}\right) \cdot \mathsf{V}_n(\lambda) \,, \\
|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| &= \mathcal{O}_{k,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\lambda_\star(\lambda)^{k+1}}{n\kappa^3} + \frac{\chi_n(\lambda)^3 \log^2 n}{\sqrt{\rho(\lambda)} n^{1-\frac{1}{k}}\kappa^{8.5}}\right) \cdot \mathsf{B}_n(\lambda) = \mathcal{O}_{k,M,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{\log^8 n}{n^{\frac{1}{2}-\frac{1}{k}}}\right) \cdot \mathsf{B}_n(\lambda) \,.
\end{aligned}$$

The proof is complete.

F.2 Proof of Proposition 4.2

Overparameterized regime. When $d/n \geq 1 + M^{-1}$, by

$$n = \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star(0)\boldsymbol{I})^{-1}\right) \geq \frac{d}{1 + M\lambda_\star(0)} \,,$$

we can deduce that $\lambda_\star(0) \geq M^{-1} \cdot (d/n - 1) \geq M^{-2}$. Hence,

$$n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star(0)\boldsymbol{I})^{-2}\right) \geq n - \frac{1}{1+\lambda_\star(0)} \cdot \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star(0)\boldsymbol{I})^{-1}\right) \geq \frac{\lambda_\star(0)}{1+\lambda_\star(0)} \cdot n \geq \frac{1}{M^2+1} \cdot n \,,$$

and therefore, in Theorem 2 we can take $\mathsf{C}_{\boldsymbol{\Sigma}} \geq 1/(M^2+1) = \Theta_M(1)$. By Eq. (92) we know $\rho(0) = \Omega_M(n^{-1})$. By [BY08, RV09], we know when $n = \Omega_{\mathsf{C}_{\boldsymbol{x}},M,D}(1)$, with probability $1 - \mathcal{O}(n^{-D+1})$ we have $s_{\min} = \Omega_{M,\mathsf{C}_{\boldsymbol{x}},D}(1)$. Substituting $\lambda_\star(0) = \Omega_M(1)$ and $\mathsf{d}_{\boldsymbol{\Sigma}}(n) = \mathcal{O}_M(n)$ (c.f. Lemma F.1) into $\chi_n'(\kappa)$, we get for $\kappa = \mathcal{O}(1)$,

$$\chi_n'(\kappa) = \mathcal{O}_M\left(\frac{\log^2 n}{\kappa}\right) .$$

Thus, by taking $\kappa = n^{-1/14}$, the conditions below hold for $n = \Omega_{k,M,\mathsf{C}_{\boldsymbol{x}},D}(1)$ given $k \geq 15$,

$$\kappa \leq \min\left\{s_{\min}/(8\lambda_\star(0)), \mathsf{C}_{\boldsymbol{\Sigma}}^2/8\right\} , \qquad n^{-2D+1} = \mathcal{O}\left(\sqrt{\frac{\kappa^3 \log^2 n}{\max\{n, \lambda\}}}\right) ,$$

$$\chi_n'(\kappa)^3 \log^2 n \le \mathsf{C}_2 n^{1-\frac{1}{k}} \kappa^{9.5} ,$$

and by taking $\kappa = n^{-1/28}$, the following additional conditions hold when $n = \Omega_{k,M,\mathsf{C}_{\boldsymbol{x}},D}(1)$ given $k \ge 29$,

$$\chi_n'(\kappa)^3 \log^2 n \le \mathsf{C}_1 n\kappa^{4.5} \min\left\{1, \sqrt{\rho(0)}\right\} , \qquad \frac{\lambda_\star(0)^{k+1}}{n\kappa^3} + \frac{\chi_n'(\kappa)^3 \log^2 n}{\sqrt{\rho(0)} n^{1-\frac{1}{k}} \kappa^{8.5}} \le \mathsf{C}_3 .$$

We can then invoke Theorem 2 by taking $\kappa = n^{-1/14}$ for variance approximation and $n^{-1/28}$ for bias approximation. Therefore, we can conclude that for $k \ge 29$,

$$\begin{aligned}
|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| &= \mathcal{O}_{k,M,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/14} + \frac{\log^8 n}{n^{1.5/14-\frac{1}{k}}}\right) \cdot \mathsf{V}_n(0) , \\
|\mathscr{B}_{\boldsymbol{X}}(0) - \mathsf{B}_n(0)| &= \mathcal{O}_{k,M,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/28} + n^{-25/28} + \frac{\log^8 n}{n^{2.5/28-\frac{1}{k}}}\right) \cdot \mathsf{B}_n(0) + \mathcal{O}\left(\|\boldsymbol{\beta}\|^2 n^{-1/28}\right) .
\end{aligned}$$

Use again $\lambda_\star(0) = \Omega_M(1)$ and $\mathsf{C}_{\boldsymbol{\Sigma}} = \Omega_M(1)$, we know

$$\mathsf{B}_n(0) = \frac{\boldsymbol{\beta}^\mathsf{T} \left(\boldsymbol{\Sigma}/\lambda_\star(0) + \boldsymbol{I}\right)^{-2} \boldsymbol{\Sigma}\boldsymbol{\beta}}{1 - n^{-1}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star(0)\boldsymbol{I})^{-2}\right)} = \Omega_M(\|\boldsymbol{\beta}\|^2) .$$

We conclude the proof by fixing $k \ge 57$, and thus

$$\begin{aligned}
|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| &= \mathcal{O}_{M,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/14}\right) \cdot \mathsf{V}_n(0) , \\
|\mathscr{B}_{\boldsymbol{X}}(0) - \mathsf{B}_n(0)| &= \mathcal{O}_{M,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/28}\right) \cdot \mathsf{B}_n(0) .
\end{aligned}$$

Underparameterized regime. Suppose $M^{-1} \le d/n \le 1 - M^{-1}$, we can invoke Theorem 3 with $\mathsf{C}_{\boldsymbol{\Sigma}} = M^{-1}$. By [BY08] we have $s_{\min} = \Omega_{M,\mathsf{C}_{\boldsymbol{x}},D}(1)$. Also as we can take $\mathsf{d}_{\boldsymbol{\Sigma}}(n) = n$ in this case, we have

$$\chi_n(\varepsilon n) \le 1 + \frac{n \log^2 n}{\varepsilon n} = \mathcal{O}\left(\frac{\log^2 n}{\kappa}\right) ,$$

and therefore the conditions below hold for $n = \Omega_{k,M,\mathsf{C}_{\boldsymbol{x}},D}(1)$ by taking $\varepsilon = n^{-1/4}$ when $k \ge 5$,

$$\varepsilon \le \min\left\{s_{\min}/2, \mathsf{C}_{\boldsymbol{\Sigma}}^2 \sigma_d/4\right\} , \qquad \chi_n(\varepsilon n)^3 \log^2 n \le \mathsf{C}_1 n \mathsf{C}_{\boldsymbol{\Sigma}}^{4.5} , \qquad n^{-2D+1} = \mathcal{O}\left(\sqrt{\frac{\mathsf{C}_{\boldsymbol{\Sigma}}^3 \log^2 n}{\max\{n, \lambda\}}}\right) ,$$

$$\chi_n(\varepsilon n)^3 \log^2 n \le \mathsf{C}_2 n^{1-\frac{1}{k}} \mathsf{C}_{\boldsymbol{\Sigma}}^{9.5} .$$

We thus have

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| = \mathcal{O}_{k,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/4} \cdot \left(\frac{1}{s_{\min}} + \frac{1}{\mathsf{C}_{\boldsymbol{\Sigma}}^2 \sigma_d}\right) + \frac{\log^8 n}{n^{\frac{1}{4}-\frac{1}{k}} \mathsf{C}_{\boldsymbol{\Sigma}}^{9.5}}\right) \cdot \mathsf{V}_n(0) .$$

By fixing $k > 20$, we know for all $n = \Omega_{M,\mathsf{C}_{\boldsymbol{x}},D}(1)$,

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| = \mathcal{O}_{M,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{1}{n^{\frac{1}{5}}}\right) \cdot \mathsf{V}_n(0) .$$

G Proofs for bounded varying spectrum regime

G.1 Proof of Proposition 4.3

Throughout this proof, we will use the shorthand $\lambda_{\mathsf{bv}} := \lambda/(n\lambda_\star(0)) \in [1/M, M]$. We begin by lower bounding the constant κ of Eq. (21). Since

$$\begin{aligned} \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-1}\right) &= n\,, \\ \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\sigma_{2n}\boldsymbol{I})^{-1}\right) &\geq \sum_{i=1}^{2n} \frac{\sigma_i}{\sigma_i+\sigma_{2n}} \geq n\,, \end{aligned} \tag{93}$$

we know that $\lambda_\star(0) \geq \sigma_{2n}$ and therefore $\psi(\delta)\lambda_\star(0) \geq \psi(\delta)\sigma_{2n} \geq \sigma_{\lfloor 2\delta n\rfloor}$ for any $\delta \in (0,1]$. We then have

$$\begin{aligned} \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\psi(\delta)\lambda_\star(0)\boldsymbol{I})^{-1}\right) &\leq 2\delta n + \sum_{i=\lfloor 2\delta n\rfloor+1}^{\infty} \frac{\sigma_i}{\sigma_i+\psi(\delta)\lambda_\star(0)} \\ &\leq 2\delta n + \frac{\sigma_{\lfloor 2\delta n\rfloor}+\lambda_\star(0)}{\sigma_{\lfloor 2\delta n\rfloor}+\psi(\delta)\lambda_\star(0)} \cdot \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-1}\right) \\ &\leq \left(\frac{1}{2}+2\delta+\frac{1}{\psi(\delta)}\right)\cdot n\,, \end{aligned}$$

where in the last inequality we use $\psi(\delta)\lambda_\star(0) \geq \sigma_{\lfloor 2\delta n\rfloor}$. Further

$$n - \frac{\lambda}{\psi(\delta)\lambda_\star(0)} = \left(1-\frac{\lambda_{\mathsf{bv}}}{\psi(\delta)}\right)\cdot n\,,$$

and therefore, using the previous inequality, we conclude the following. If $\delta > 0$ is such that

$$1-\frac{\lambda_{\mathsf{bv}}}{\psi(\delta)} \geq \frac{1}{2}+2\delta+\frac{1}{\psi(\delta)}\,,$$

then $\lambda_\star(\lambda) \leq \psi(\delta)\lambda_\star(0)$. Let $\delta_0 = \delta_0(M,\psi)$ be defined follows

$$\delta_0 := \sup\left\{\delta \in (0,1/2):\ 2\delta + \frac{1+M}{\psi(\delta)} \leq \frac{1}{2}\right\}.$$

Them $\lambda_\star(0) \leq \lambda_\star(\lambda) \leq \psi(\delta_0)\lambda_\star(0)$. Hence

$$\frac{\lambda}{n\lambda_\star(\lambda)} = \frac{\lambda_\star(0)\lambda_{\mathsf{bv}}}{\lambda_\star(\lambda)} \geq \frac{\lambda_{\mathsf{bv}}}{\psi(\delta_0)}\,,$$

and

$$1-\frac{\lambda}{n\lambda_\star(\lambda)} = \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star(\lambda)\boldsymbol{I})^{-1}\right) \geq \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\psi(\delta_0)\lambda_\star(0)\boldsymbol{I})^{-1}\right) \geq \frac{1}{\psi(\delta_0)}\,.$$

Putting together the above displays, we conclude Eq (21) holds for $\kappa \geq \min\{\lambda_{\mathsf{bv}},1\}/\psi(\delta_0) = \Omega_{M,\psi}(1)$ when $n = \Omega_{M,\psi}(1)$ (recall that we need $2\delta_0 n \geq 1$).

To verify the conditions of Theorem 1, we first assume $\mathsf{d}_{\boldsymbol{\Sigma}} = \mathcal{O}(n^{1+\gamma})$ for $\gamma \in [0,1/3)$,

$$\chi_n(n\lambda_\star(0)\lambda_{\mathsf{bv}}) = 1+\frac{\sigma_{\lfloor \eta n\rfloor}\mathsf{d}_{\boldsymbol{\Sigma}}\log^2(\mathsf{d}_{\boldsymbol{\Sigma}})}{n\lambda_\star(0)\lambda_{\mathsf{bv}}} \leq 1+\frac{\sigma_{\lfloor \eta n\rfloor}\mathsf{d}_{\boldsymbol{\Sigma}}\log^2(\mathsf{d}_{\boldsymbol{\Sigma}})}{n\sigma_{2n}\lambda_{\mathsf{bv}}} \leq 1+\frac{\psi(\eta/4)\mathsf{d}_{\boldsymbol{\Sigma}}\log^2(\mathsf{d}_{\boldsymbol{\Sigma}})}{n\lambda_{\mathsf{bv}}}$$

$$= \mathcal{O}_{M,\psi,\mathsf{C}_{\boldsymbol{x}}}\left(\frac{\mathsf{d}_{\boldsymbol{\Sigma}}\log^2 n}{n}\right) = \mathcal{O}_{M,\psi,\mathsf{C}_{\boldsymbol{x}}}\left(n^{\gamma}\log^2 n\right)\,,$$

and with $\kappa = \Omega_{M,\psi}(1)$, the conditions

$$\chi_n(n\lambda_\star(0)\lambda_{\mathsf{bv}})^3\log^2 n \le \mathsf{C}n\kappa^{4.5}\,, \qquad n^{-2D+1} = \mathcal{O}\left(\sqrt{\frac{\kappa^3\log^2 n}{\max\{n,\lambda\}}}\right)\,,$$

hold if $n = \Omega_{M,\psi,\gamma,\mathsf{C}_{\boldsymbol{x}},D}(1)$. We then can apply Theorem 1 to approximate the variance. Given any positive integer k, if $n = \Omega_{k,M,\psi,\gamma,\mathsf{C}_{\boldsymbol{x}},D}(1)$, it holds with probability $1 - \mathcal{O}_k(n^{-D+1})$ that

$$|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| = \mathcal{O}_{k,M,\psi,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{(\mathsf{d}_{\boldsymbol{\Sigma}}/n)^3\log^8 n}{n^{1-\frac{1}{k}}}\right)\cdot \mathsf{V}_n(\lambda)\,.$$

If additionally $\mathsf{d}_{\boldsymbol{\Sigma}} = \mathcal{O}_{M,\psi,\mathsf{C}_{\boldsymbol{x}}}(n^{1+\gamma}\lambda^{1/6})$, we have

$$\chi_n(\lambda)^3\log^2 n \le \mathsf{C}n\kappa^{4.5}\sqrt{\rho(\lambda)}\,,$$

when $n = \Omega_{M,\psi,\gamma,\mathsf{C}_{\boldsymbol{x}},D}(1)$. The condition $\lambda k n^{-\frac{1}{k}} \le n\kappa/2$ is equivalent to $\lambda_\star(0)\lambda_{\mathsf{bv}}kn^{-1/k} \le \kappa/2$, which holds when $n = \Omega_{k,M,\psi}(1)$ since we have assumed $\lambda_\star(0) = \mathcal{O}(1)$. Therefore, we can appeal to the bias approximation result in Theorem 1, yielding

$$|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| = \mathcal{O}_{k,M,\psi,\mathsf{C}_{\boldsymbol{x}},D}\left(\frac{(\mathsf{d}_{\boldsymbol{\Sigma}}/n)^3\log^8 n}{\sqrt{\rho(\lambda)}n^{1-\frac{1}{k}}}\right)\cdot \mathsf{B}_n(\lambda)\,.$$

G.2 Proof of Proposition 4.4

We provide the following bounds for the quantities in Theorem 2.

Lemma G.1. *Under the same Assumptions of Theorem 2, we can take*

$$\mathsf{C}_{\boldsymbol{\Sigma}} = \Omega_\psi(1)\,,$$

when $n = \Omega(1)$. For $\kappa = \mathcal{O}(1)$, we have

$$\chi_n'(\kappa) = \mathcal{O}_\psi\left(\frac{\log^{2+\mathcal{O}(1)} n}{\kappa}\right)\,.$$

In addition, $s_{\mathsf{min}} = \Omega_{\psi,\mathsf{C}_{\boldsymbol{x}}}(\sigma_n)$ with probability $1 - \mathcal{O}(n^{-D+1})$.

Proof. Since

$$\begin{aligned} n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star(0)\boldsymbol{I})^{-2}\right) &\ge n - \sum_{i=1}^{\lfloor n/2\rfloor}\frac{\sigma_i}{\sigma_i + \lambda_\star(0)} - \frac{\sigma_{\lfloor n/2\rfloor}}{\sigma_{\lfloor n/2\rfloor} + \lambda_\star(0)}\sum_{i=\lfloor n/2\rfloor+1}^{\infty}\frac{\sigma_i}{\sigma_i + \lambda_\star(0)} \\ &= \frac{\lambda_\star(0)}{\sigma_{\lfloor n/2\rfloor} + \lambda_\star(0)}\sum_{i=\lfloor n/2\rfloor+1}^{\infty}\frac{\sigma_i}{\sigma_i + \lambda_\star(0)}\,, \end{aligned}$$

where in the last line we use $\operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I})^{-1}\right)=n$. Since

$$\sum_{i=\lfloor n/2\rfloor+1}^{\infty}\frac{\sigma_i}{\sigma_i+\lambda_\star(0)}=n-\sum_{i=1}^{\lfloor n/2\rfloor}\frac{\sigma_i}{\sigma_i+\lambda_\star(0)}\geq n-\frac{n}{2}=\frac{n}{2}\,,$$

and $\lambda_\star(0)\geq\sigma_{2n}$ from Eq. (93), we know

$$n-\operatorname{Tr}\left(\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}+\lambda_\star(0)\boldsymbol{I}\right)^{-2}\right)\geq\frac{\sigma_{2n}}{\sigma_{\lfloor n/2\rfloor}+\sigma_{2n}}\cdot\frac{n}{2}\geq\frac{1}{2\psi(1/4)+2}\cdot n\,.$$

We can hence take $\mathsf{C}_{\boldsymbol{\Sigma}}:=(2\psi(1/4)+2)^{-1}=\Omega_\psi(1)$.

Substituting $\lambda_\star(0)\geq\sigma_{2n}$ and $\mathsf{d}_{\boldsymbol{\Sigma}}=\mathcal{O}(n^{1+\gamma})$ for some $1\leq\gamma<1/2$ into Eq. (24), we have for $\kappa=\mathcal{O}(1)$,

$$\chi_n'(\kappa)=\mathcal{O}\left(\frac{\sigma_{\lfloor\eta n\rfloor}\mathsf{d}_{\boldsymbol{\Sigma}}\log^2 n}{\kappa n\sigma_{2n}}\right)=\mathcal{O}_\psi\left(\frac{\log^{2+\mathcal{O}(1)}n}{\kappa}\right)\,.$$

Finally, to get a lower bound for $s_{\min}$, we use Cauchy interlacing theorem which implies

$$\begin{aligned}s_{\min}&\geq\lambda_n\left(\frac{\boldsymbol{X}^\mathsf{T}\boldsymbol{X}}{n}\right)=\lambda_n\left(\frac{\boldsymbol{\Sigma}^{\frac12}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Sigma}^{\frac12}}{n}\right)\geq\lambda_n\left(\frac{\boldsymbol{P}_k\boldsymbol{\Sigma}^{\frac12}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Sigma}^{\frac12}\boldsymbol{P}_k}{n}\right)\\&=\lambda_n\left(\frac{\boldsymbol{Z}\boldsymbol{P}_k\boldsymbol{\Sigma}\boldsymbol{P}_k\boldsymbol{Z}^\mathsf{T}}{n}\right)\,.\end{aligned}$$

where $\boldsymbol{P}_k$ is the projection to the space spanned by the top k eigenvectors. Let $k\geq n$, we further have

$$s_{\min}\geq\sigma_k\cdot\lambda_n\left(\frac{\boldsymbol{Z}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\boldsymbol{Z}^\mathsf{T}}{n}\right)\,,$$

where $\boldsymbol{P}_k=\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}$ and $\boldsymbol{V}_k=\begin{bmatrix}\boldsymbol{v}_1&\cdots&\boldsymbol{v}_k\end{bmatrix}\in\mathbb{R}^{d\times k}$ with $\boldsymbol{v}_i$ being the i-th eigenvector of $\boldsymbol{\Sigma}$. Since $\boldsymbol{Z}\boldsymbol{V}_k$ is a $n\times k$ random matrix with i.i.d. isotropic and sub-Gaussian rows. When $k\geq n$, by [Ver12, Thm. 5.58, generalized version in Sec. 5.7], we have

$$\lambda_n\left(\frac{\boldsymbol{Z}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\boldsymbol{Z}^\mathsf{T}}{n}\right)\geq\left((1-\zeta)\sqrt{\frac{k}{n}}-\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}}}(1)-\mathcal{O}_{\mathsf{C}_{\boldsymbol{x}},D}\left(\sqrt{\frac{\log n}{n}}\right)\right)^2\,,$$

with probability at least $1-\mathcal{O}(n^{-D+1})$, where ζ is the random variable

$$\zeta:=\max_{1\leq i\leq n}\left|\frac{\left\|\boldsymbol{V}_k^\mathsf{T}\boldsymbol{z}_i\right\|^2}{k}-1\right|\,.$$

By Hanson-Wright in Lemma 2.1 and similar to the argument in Eq. (61), we have

$$\mathbb{P}\left(\left|\frac{\left\|\boldsymbol{V}_k^\mathsf{T}\boldsymbol{z}_i\right\|^2}{k}-1\right|\geq t\right)=2\exp\left(-\Omega_{\mathsf{C}_{\boldsymbol{x}}}\left(k\cdot\min\left\{t^2,t\right\}\right)\right)\,.$$

Given the above sharp concentration of ζ, we can therefore conclude by taking $k=\lfloor\mathsf{C}(\mathsf{C}_{\boldsymbol{x}})n\rfloor$ for some $\mathsf{C}>0$, and $n=\Omega_{\mathsf{C}_{\boldsymbol{x}},D}(1)$, we have with probability $1-\mathcal{O}(n^{-D+1})$ that

$$\lambda_n\left(\frac{\boldsymbol{Z}\boldsymbol{V}_k\boldsymbol{V}_k^\mathsf{T}\boldsymbol{Z}^\mathsf{T}}{n}\right)\geq1\,,$$

and therefore $s_{\min}\geq\sigma_k\geq\sigma_n/\psi(\mathsf{C})$ by Assumption 3. □

By Lemma G.1 and the assumption $\lambda_\star(0)/\sigma_n = \mathcal{O}(\log^{\mathcal{O}(1)} n)$, we know by taking $\kappa = n^{-1/14}$, the conditions below hold for $n = \Omega_{k,\psi,\mathsf{C}_{\boldsymbol{x}},D}(1)$ whenever $k \geq 15$,

$$\kappa \leq \min\left\{s_{\mathsf{min}}/(8\lambda_\star(0)), \mathsf{C}_{\boldsymbol{\Sigma}}^2/8\right\}, \qquad \chi_n'(\kappa)^3 \log^2 n \leq \mathsf{C}_1 n\kappa^{4.5}, \qquad n^{-2D+1} = \mathcal{O}\left(\sqrt{\frac{\kappa^3 \log^2 n}{\max\{n,\lambda\}}}\right),$$

$$\chi_n'(\kappa)^3 \log^2 n \leq \mathsf{C}_2 n^{1-\frac{1}{k}} \kappa^{9.5}.$$

Therefore by the variance approximation in Theorem 2, it holds

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| = \mathcal{O}_{k,\psi,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/14}\log^{\mathcal{O}(1)}(n) + \frac{\log^{8+\mathcal{O}(1)} n}{n^{1.5/14-\frac{1}{k}}}\right) \cdot \mathsf{V}_n(0).$$

Fixing $k \geq 29$, we have with probability $1 - \mathcal{O}(n^{-D+1})$, we know as $n = \Omega_{\psi,\mathsf{C}_{\boldsymbol{x}},D}(1)$,

$$|\mathscr{V}_{\boldsymbol{X}}(0) - \mathsf{V}_n(0)| = \mathcal{O}_{\psi,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-1/15}\right) \cdot \mathsf{V}_n(0).$$

For the bias approximation with the assumption $\rho(0) = \Omega(n^{-2+\gamma})$, the following additional conditions hold by taking $\kappa = n^{-\gamma/28}$ when $n = \Omega_{k,\psi,\mathsf{C}_{\boldsymbol{x}},D}(1)$ given $k \geq 29/\gamma$,

$$\chi_n'(\kappa)^3 \log^2 n \leq \mathsf{C}_1 n\kappa^{4.5} \min\left\{1, \sqrt{\rho(0)}\right\}, \qquad \frac{\lambda_\star(0)^{k+1}}{n\kappa^3} + \frac{\chi_n'(\kappa)^3 \log^2 n}{\sqrt{\rho(0)} n^{1-\frac{1}{k}} \kappa^{8.5}} \leq \mathsf{C}_3.$$

In verifying the second condition above, we use $\lambda_\star(0) = \mathcal{O}(\sigma_n \log^{\mathcal{O}(1)} n) = \mathcal{O}(\sigma_n \log^{\mathcal{O}(1)} n)$. By Lemma G.1, we also have

$$\kappa\lambda_\star(0)\chi_n'(\kappa) = \mathcal{O}_\psi(\sigma_n \log^{2+\mathcal{O}(1)}).$$

We can then write out the bias approximation result applying Theorem 2

$$\begin{aligned}|\mathscr{B}_{\boldsymbol{X}}(0) - \mathsf{B}_n(0)| = \mathcal{O}_{k,\psi,\mathsf{C}_{\boldsymbol{x}},D}&\left(n^{-\gamma/28} + n^{-25\gamma/28}\log^{\mathcal{O}(k)} n + \frac{\log^{8+\mathcal{O}(1)} n}{n^{2.5\gamma/28-\frac{1}{k}}}\right) \cdot \mathsf{B}_n(0) \\ &+ \mathcal{O}_{\psi,\mathsf{C}_{\boldsymbol{x}},D}\left(\sigma_n^2 \log^{4+\mathcal{O}(1)} \|\boldsymbol{\theta}_{\leq n}\|^2 + \sigma_n \log^{2+\mathcal{O}(1)} \|\boldsymbol{\beta}_{>n}\|^2\right).\end{aligned}$$

Fixing $k \geq 57/\gamma$, we conclude the proof with

$$|\mathscr{B}_{\boldsymbol{X}}(0) - \mathsf{B}_n(0)| = \mathcal{O}_{\psi,\mathsf{C}_{\boldsymbol{x}},D}\left(n^{-\gamma/29}\right) \cdot \mathsf{B}_n(0) + \log^{\mathcal{O}(1)} \cdot \mathcal{O}_{\psi,\mathsf{C}_{\boldsymbol{x}},D}\left(\sigma_n^2 \|\boldsymbol{\theta}_{\leq n}\|^2 + \sigma_n \|\boldsymbol{\beta}_{>n}\|^2\right).$$

H Proof of Theorem 4

Define the following increasing function in t,

$$f_n(t;\lambda) = 1 - \frac{\lambda}{n \cdot t\sigma_n} - \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + t\sigma_n \boldsymbol{I})^{-1}\right).$$

Case I: regularly varying spectrum when $\alpha > 1$. In the first case, we set $\lambda = \nu n \sigma_n$. For any $t > 0$, we can compute that

$$f_n(t;\lambda) = 1 - \frac{\nu}{t} - \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + t\sigma_n \boldsymbol{I})^{-1}\right) .$$

We will first show $\mathsf{d}_{\boldsymbol{\Sigma}}(n) = \mathcal{O}_{\boldsymbol{\Sigma}}(n)$ and $\lambda_\star = \Theta_\nu(\lambda/n)$, and then we can invoke Proposition 4.3 for variance approximation. For simplicity, we will suppress the dependence on sequences $\{a_i\}$ and $\{b_i\}$ in the big-O and big-Ω notations. For instance, we will just write for all $n = \Omega_\alpha(1)$, $|b_n| \le \alpha$.

We first upper bound $\mathsf{d}_{\boldsymbol{\Sigma}}$. Note that

$$\sum_{l=k}^{d} \sigma_l = \sum_{l=k}^{\infty} l^{-\alpha} a_l \exp\left\{\sum_{j=1}^{l} b_j/j\right\} = \sigma_k \cdot \sum_{l=k}^{\infty} \left(\frac{l}{k}\right)^{-\alpha} \cdot \frac{a_l}{a_k} \cdot \exp\left\{\sum_{j=k+1}^{l} b_j/j\right\} .$$

As a_l converges to a positive limit, we have $a_l/a_k = \mathcal{O}(1)$. For $k = \Omega_\alpha(1)$ such that $|b_l| \le \alpha/2$ for all $l \ge k$, we can further derive that

$$\begin{aligned}
\sum_{l=k}^{d} \sigma_l &\le \sigma_k \cdot \mathcal{O}(1) \cdot \sum_{l=k}^{\infty} \left(\frac{l}{k}\right)^{-\alpha} \cdot \exp\left\{\frac{\alpha}{2} \sum_{j=k+1}^{l} j^{-1}\right\} \\
&\le \sigma_k \cdot \mathcal{O}(1) \cdot \sum_{l=k}^{\infty} \left(\frac{l}{k}\right)^{-\alpha} \cdot \exp\left\{\frac{\alpha}{2} \int_k^l \frac{\mathrm{d}t}{t}\right\} \\
&= \sigma_k \cdot \mathcal{O}(1) \cdot \sum_{l=k}^{\infty} \left(\frac{l}{k}\right)^{\alpha/2-\alpha} = k\sigma_k \cdot \mathcal{O}(1) \cdot \sum_{l=k}^{\infty} \frac{1}{k} \left(\frac{l}{k}\right)^{-\alpha/2} \\
&\le k\sigma_k \cdot \mathcal{O}(1) \cdot \left(\frac{1}{k} + \int_1^{\infty} t^{-\alpha/2} \mathrm{d}t\right) \\
&= \mathcal{O}_\alpha(k\sigma_k) .
\end{aligned}$$

This implies for all $n = \Omega_\alpha(1)$, we can take $\mathsf{d}_{\boldsymbol{\Sigma}}(n) = \mathcal{O}_\alpha(n)$. Next we show $\lambda_\star = \Theta_\nu(\lambda/n)$. Note that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + t\sigma_n \boldsymbol{I})^{-1}\right) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{l=1}^{\infty} \frac{\sigma_l}{\sigma_l + t\sigma_n} = \lim_{M\to\infty} \lim_{n\to\infty} \frac{1}{n} \sum_{M^{-1}n \le l \le Mn} \frac{\sigma_l}{\sigma_l + t\sigma_n} \\
&= \lim_{M\to\infty} \lim_{n\to\infty} \frac{1}{n} \sum_{M^{-1}n \le l \le Mn} \frac{1}{1 + t(l/n)^\alpha} = \int_0^{\infty} \frac{1}{1 + tx^\alpha} \mathrm{d}x = t^{-1/\alpha} \cdot \frac{1}{\alpha} \int_0^{\infty} \frac{u^{1/\alpha - 1}}{1+u} \mathrm{d}u \\
&= t^{-1/\alpha} \cdot \frac{\mathsf{Beta}(1/\alpha, 1 - 1/\alpha)}{\alpha} = t^{-1/\alpha} \cdot \frac{\Gamma(1/\alpha)\Gamma(1 - 1/\alpha)}{\alpha\Gamma(1)} \\
&\overset{\text{(i)}}{=} t^{-1/\alpha} \frac{\pi/\alpha}{\sin(\pi/\alpha)} ,
\end{aligned} \tag{94}$$

where in (i) we use the reflection formula for Γ function. Recall that we define $\mathsf{c}_\star = \mathsf{c}_\star(\nu)$ as the unique solution of

$$1 = \nu \mathsf{c}_\star^{-1} + \frac{\pi/\alpha}{\sin(\pi/\alpha)} \mathsf{c}_\star^{-1/\alpha} ,$$

it then follows from the above displays that

$$\lim_{n\to\infty} f_n(\mathsf{c}_\star;\lambda) = 1-\nu\mathsf{c}_\star^{-1} - \lim_{n\to\infty}\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\mathsf{c}_\star\sigma_n\boldsymbol{I})^{-1}\right)$$
$$= 1-\nu\mathsf{c}_\star^{-1} - \frac{\pi/\alpha}{\sin(\pi/\alpha)}\mathsf{c}_\star^{-1/\alpha} = 0\,.$$

By the definition of $\lambda_\star$ in (5), we can write $f_n(\lambda_\star/\sigma_n;\lambda)=0$. Combining with the above limit, we can then conclude that

$$\lambda_\star = \mathsf{c}_\star\sigma_n(1+o_n(1))\,.$$

Substituting into Eq. (23), we further have

$$\begin{aligned}
\rho(\lambda) &= \frac{\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|^2\,\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\right)} = \frac{\boldsymbol{\beta}^\mathsf{T}(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-1}\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|^2_{\boldsymbol{\Sigma}^{-1}}\,(n-\lambda/\lambda_\star)} \\
&= \frac{\sum_{l=1}^\infty(\sigma_l+\lambda_\star)^{-1}\langle\boldsymbol{\beta},\boldsymbol{v}_l\rangle^2}{n(1-\nu\mathsf{c}_\star^{-1})\sum_{l=1}^\infty\sigma_l^{-1}\langle\boldsymbol{\beta},\boldsymbol{v}_l\rangle^2}\cdot(1+o_n(1)) \\
&= \frac{\sum_{l=1}^\infty\sigma_n(\sigma_l+\lambda_\star)^{-1}\langle\boldsymbol{\beta},\boldsymbol{v}_l\rangle^2}{n(1-\nu\mathsf{c}_\star^{-1})\sum_{l=1}^\infty\sigma_n\sigma_l^{-1}\langle\boldsymbol{\beta},\boldsymbol{v}_l\rangle^2}\cdot(1+o_n(1)) \\
&= \frac{\sum_{l=1}^\infty(l/n)^\alpha(1+\mathsf{c}_\star(l/n)^\alpha)^{-1}\langle\boldsymbol{\beta},\boldsymbol{v}_l\rangle^2}{n(1-\nu\mathsf{c}_\star^{-1})\sum_{l=1}^\infty(l/n)^\alpha\langle\boldsymbol{\beta},\boldsymbol{v}_l\rangle^2}\cdot(1+o_n(1)) \\
&= \frac{\int_0^\infty x^\alpha(1+\mathsf{c}_\star x^\alpha)^{-1}\,\mathrm{d}F_{\boldsymbol{\beta}}(x)}{n(1-\nu\mathsf{c}_\star^{-1})\int_0^\infty x^\alpha\,\mathrm{d}F_{\boldsymbol{\beta}}(x)}\cdot(1+o_n(1))\,.
\end{aligned}$$

Therefore, under the additional condition for some $0<\theta\le 1$ that

$$\int_0^\infty x^\alpha\,\mathrm{d}F_{\boldsymbol{\beta}}(x) = \mathcal{O}\left(n^{1-\theta}\int_0^\infty x^\alpha(1+\mathsf{c}_\star x^\alpha)^{-1}\,\mathrm{d}F_{\boldsymbol{\beta}}(x)\right),$$

we have $\rho(\lambda)=\Omega(n^{-2+\theta})$. By choosing $\gamma=(1-\theta)/3$, we can invoke Proposition 4.3. Choosing a sufficiently large k yields $\mathscr{V}_{\boldsymbol{X}}(\lambda)=\mathsf{V}_n(\lambda)(1+o_n(1))$ and $\mathscr{B}_{\boldsymbol{X}}(\lambda)=\mathsf{B}_n(\lambda)(1+o_n(1))$.

In the next step, we derive explicit asymptotic formulas for V_n and B_n. Similar to the previous calculations in Eq. (94), we can compute that

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+t\sigma_n\boldsymbol{I})^{-2}\right) \\
&= \int_0^\infty\frac{1}{(1+tx^\alpha)^2}\mathrm{d}x = t^{-1/\alpha}\cdot\frac{1}{\alpha}\int_0^\infty\frac{u^{1/\alpha-1}}{(1+u)^2}\mathrm{d}u \\
&= t^{-1/\alpha}\cdot\frac{\mathsf{Beta}(1/\alpha,2-1/\alpha)}{\alpha} = t^{-1/\alpha}\cdot\frac{\Gamma(1/\alpha)\Gamma(2-1/\alpha)}{\alpha\Gamma(2)} = t^{-1/\alpha}\cdot\frac{\Gamma(1/\alpha)\Gamma(1-1/\alpha)}{\alpha}\cdot\left(1-\frac{1}{\alpha}\right) \\
&= t^{-1/\alpha}\frac{\pi/\alpha}{\sin(\pi/\alpha)}\cdot\left(1-\frac{1}{\alpha}\right).
\end{aligned}$$

and further $n^{-1}\mathrm{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2})\to(1-\nu\mathsf{c}_\star^{-1})(1-\alpha^{-1})$. This then gives the variance

$$\mathsf{V}_n(\lambda) = \frac{\tau^2 n^{-1}\mathrm{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2})}{1-n^{-1}\mathrm{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2})} = \frac{\tau^2(1-\nu\mathsf{c}_\star^{-1})(\alpha-1)}{1+\nu\mathsf{c}_\star^{-1}(\alpha-1)}(1+o_n(1))\,.$$

For the bias term, we can similarly write

$$\begin{aligned}\lambda_\star^2\langle\boldsymbol{\beta},(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}\rangle &= (1+o_n(1))\cdot\sigma_n\mathsf{c}_\star^2\sum_{l=1}^{\infty}\frac{\sigma_l\sigma_n}{(\sigma_l+\lambda_\star)^2}\langle\boldsymbol{\beta},\boldsymbol{v}_l\rangle^2\\ &= (1+o_n(1))\cdot\sigma_n\mathsf{c}_\star^2\sum_{l=1}^{\infty}\frac{(l/n)^\alpha}{(1+\mathsf{c}_\star(l/n)^\alpha)^2}\langle\boldsymbol{\beta},\boldsymbol{v}_l\rangle^2\\ &= \sigma_n\mathsf{c}_\star^2\int_0^\infty\frac{x^\alpha}{(1+\mathsf{c}_\star x^\alpha)^2}\,\mathrm{d}F_{\boldsymbol{\beta}}(x)\left(1+o_n(1)\right).\end{aligned}$$

Together with $\mathrm{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}) = n(1-\nu\mathsf{c}_\star^{-1})(1-\alpha^{-1})(1+o_n(1))$, we conclude the proof for this case.

Case II: regularly varying spectrum when $\alpha = 1$. Setting $\lambda = \nu n\sigma_n\log n$. For any $t>0$, we can compute that

$$f_n(t;\lambda) = 1-\frac{\nu\log n}{t}-\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+t\sigma_n\boldsymbol{I})^{-1}\right).$$

We first verify Assumption 1 holds. With $a_i=\mathcal{O}(1)$ bounded and $\alpha'>1$, we indeed have that $\mathrm{Tr}(\boldsymbol{\Sigma})<\infty$ as

$$\mathrm{Tr}(\boldsymbol{\Sigma}) = \sum_{l=1}^{\infty}\frac{a_l}{l(1+\log l)^{\alpha'}} = \mathcal{O}\left(1+\int_1^\infty\frac{M}{t(1+\log t)^{\alpha'}}\mathrm{d}t\right) = \mathcal{O}\left(-\frac{(1+\log t)^{1-\alpha'}}{\alpha'-1}\Big|_{t=1}^{\infty}\right) = \mathcal{O}\left(\frac{1}{\alpha'-1}\right).$$

Since the sequence $\{a_i\}$ converge to a positive limit, we have for $k=\Omega(1)$

$$\begin{aligned}\sum_{l=k}^{d}\sigma_l &= \Theta\left(\sum_{l=k}^{\infty}\frac{1}{l(1+\log l)^{\alpha'}}\right) = \Theta\left(\sigma_k+\int_k^\infty\frac{1}{t(1+\log t)^{\alpha'}}\mathrm{d}t\right) = \Theta\left(\sigma_k+\frac{(1+\log k)^{1-\alpha'}}{\alpha'-1}\right)\\ &= \Theta_{\alpha'}(k\log k\sigma_k),\end{aligned}$$

and therefore, we can take $\mathsf{d}_{\boldsymbol{\Sigma}}(n)=\Theta_{\alpha'}(n\log n)$. We proceed to compute $\lambda_\star$. Taking any $t>0$,

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+t\sigma_n\log n\boldsymbol{I})^{-1}\right)\\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{l=1}^{\infty}\frac{\sigma_l}{\sigma_l+t\sigma_n\log n} = \lim_{M\to\infty}\lim_{n\to\infty}\frac{1}{n}\sum_{l\geq M^{-1}n}\frac{\sigma_l}{\sigma_l+t\sigma_n\log n}\\ &= \lim_{M\to\infty}\lim_{n\to\infty}\frac{1}{n}\sum_{l\geq M^{-1}n}\frac{\sigma_l}{t\sigma_n\log n} = \lim_{M\to\infty}\lim_{n\to\infty}\frac{(1+\log n-\log M)^{1-\alpha'}}{t(\alpha'-1)(\log n)^{1-\alpha'}}\\ &= \frac{1}{t(\alpha'-1)}.\end{aligned}$$

Recalling that $\mathsf{c}_\star$ solves

$$1 = \nu\mathsf{c}_\star^{-1}+(\alpha'-1)^{-1}\mathsf{c}_\star^{-1},$$

we then have

$$\lim_{n\to\infty}f_n(\mathsf{c}_\star\log n;\lambda) = 1-\nu\mathsf{c}_\star^{-1}-\lim_{n\to\infty}\frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\mathsf{c}_\star\sigma_n\log n\boldsymbol{I})^{-1}\right)$$

$$= 1 - \nu \mathsf{c}_\star^{-1} - (\alpha' - 1)^{-1} \mathsf{c}_\star^{-1} = 0\,,$$

and consequently by Eq. (5),

$$\lambda_\star = \mathsf{c}_\star \sigma_n \log n (1 + o_n(1))\,.$$

Taking the above display into Eq. (23), we get

$$\begin{aligned}
\rho(\lambda) &= \frac{\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\Sigma}^{\frac{1}{2}} (\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}}{\|\boldsymbol{\theta}\|^2 \operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-1}\right)} = \frac{\sum_{l=1}^\infty (\sigma_l + \lambda_\star)^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2}{n(1 - \nu \mathsf{c}_\star^{-1}) \sum_{l=1}^\infty \sigma_l^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2} \cdot (1 + o_n(1)) \\
&= \frac{\sum_{l=1}^\infty \sigma_n (\sigma_l + \mathsf{c}_\star \sigma_n \log n)^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2}{n(1 - \nu \mathsf{c}_\star^{-1}) \sum_{l=1}^\infty \sigma_n \sigma_l^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2} \cdot (1 + o_n(1)) \\
&= \frac{\sum_{l=1}^\infty l/n \cdot (1 + l \cdot (\mathsf{c}_\star \log n/n))^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2}{n(1 - \nu \mathsf{c}_\star^{-1}) \sum_{l=1}^\infty l/n \cdot \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2} \cdot (1 + o_n(1)) \\
&= \frac{\sum_{l=1}^\infty l \cdot (\log n/n) \cdot (1 + l \cdot (\mathsf{c}_\star \log n/n))^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2}{n(1 - \nu \mathsf{c}_\star^{-1}) \sum_{l=1}^\infty l \cdot (\log n/n) \cdot \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2} \cdot (1 + o_n(1)) \\
&= \frac{\int_0^\infty x \, (1 + \mathsf{c}_\star x)^{-1} \, \mathrm{d}F_{\boldsymbol{\beta}}(x)}{n(1 - \nu \mathsf{c}_\star^{-1}) \int_0^\infty x \, \mathrm{d}F_{\boldsymbol{\beta}}(x)} \cdot (1 + o_n(1))\,,
\end{aligned}$$

where in the last line we use $F_{\boldsymbol{\beta}}(x) = \sum_{k=1}^{\lfloor (n/\log n) x \rfloor} \langle \boldsymbol{\beta}, \boldsymbol{v}_k \rangle^2$. Thus we can have $\rho(\lambda) = \Omega(n^{-2+\theta})$ provided the condition

$$\int_0^\infty x \, \mathrm{d}F_{\boldsymbol{\beta}}(x) = \mathcal{O}\left(n^{1-\theta} \int_0^\infty x \, (1 + \mathsf{c}_\star x)^{-1} \, \mathrm{d}F_{\boldsymbol{\beta}}(x) \right).$$

Setting $\gamma = (1 - \theta)/3$, we can invoke Proposition 4.3 and obtain $\mathscr{V}_{\boldsymbol{X}}(\lambda) = \mathsf{V}_n(\lambda)(1 + o_n(1))$, $\mathscr{B}_{\boldsymbol{X}}(\lambda) = \mathsf{B}_n(\lambda)(1 + o_n(1))$.

For the variance $\mathsf{V}_n(\lambda)$, we note

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\log n}{n} \operatorname{Tr}\left(\boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + t\sigma_n \log n \boldsymbol{I})^{-2}\right) \\
&= \lim_{n\to\infty} \frac{\log n}{n} \sum_{l=1}^\infty \frac{\sigma_l^2}{(\sigma_l + t\sigma_n \log n)^2} = \lim_{M\to\infty} \lim_{n\to\infty} \frac{\log n}{n} \sum_{M^{-1} n/\log n \le l \le M n/\log n} \frac{\sigma_l^2}{(\sigma_l + t\sigma_n \log n)^2} \\
&= \lim_{M\to\infty} \lim_{n\to\infty} \frac{\log n}{n} \sum_{M^{-1} n/\log n \le l \le M n/\log n} \frac{1}{(1 + tl \log n/n)^2} = \int_0^\infty \frac{1}{(1 + tx)^2} \mathrm{d}x = \frac{1}{t}\,.
\end{aligned}$$

Substituting in $\lambda_\star$, we thus have $n^{-1} \operatorname{Tr}\left(\boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2}\right) = (1 + o_n(1))/(\mathsf{c}_\star \log n)$, which further implies that

$$\mathsf{V}_n(0) = \frac{\tau^2 n^{-1} \operatorname{Tr}(\boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \lambda_\star(0) \boldsymbol{I})^{-2})}{1 - n^{-1} \operatorname{Tr}(\boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \lambda_\star(0) \boldsymbol{I})^{-2})} = \frac{\tau^2}{\mathsf{c}_\star \log n} \left(1 + o_n(1)\right).$$

Finally for the bias, we have

$$\lambda_\star^2 \langle \boldsymbol{\beta}, (\boldsymbol{\Sigma} + \lambda_\star \boldsymbol{I})^{-2} \boldsymbol{\Sigma} \boldsymbol{\beta} \rangle = (1 + o_n(1)) \cdot \mathsf{c}_\star^2 \sigma_n \log n \sum_{l=1}^\infty \frac{\sigma_l \sigma_n \log n}{(\sigma_l + \lambda_\star)^2} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2$$

$$= (1+o_n(1)) \cdot \mathsf{c}_\star^2 \sigma_n \log n \sum_{l=1}^{\infty} \frac{(l \log n/n)}{(1+\mathsf{c}_\star l \log n/n)^2} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2$$
$$= \mathsf{c}_\star^2 \sigma_n \log n \int_0^\infty \frac{x}{(1+\mathsf{c}_\star x)^2} \, \mathrm{d}F_{\boldsymbol{\beta}}(x) \left(1+o_n(1)\right).$$

Combining with $n^{-1}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-2}\right) = (1+o_n(1))/(\mathsf{c}_\star \log n)$, it holds that

$$\mathsf{B}_n(\lambda) = \mathsf{c}_\star^2 \sigma_n \log n \int_0^\infty \frac{x}{(1+\mathsf{c}_\star x)^2} \, \mathrm{d}F_{\boldsymbol{\beta}}(x) \left(1+o_n(1)\right).$$

Case III: a non-regularly varying spectrum. Take $\lambda = \nu n \sigma_n$. For any $t > 0$, we can compute that

$$f_n(t;\lambda) = 1 - \frac{\nu}{t} - \frac{1}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+t\sigma_n \boldsymbol{I})^{-1}\right).$$

If $\sigma_k = p^{-s}$, we can easily have $\sigma_l \le p^{-r-s}$ if $q^r k \le l < q^{r+1}k$. This immediately yields

$$\sum_{l=k}^{d} \sigma_l \le \sum_{r=0}^{\infty} (q^{r+1}k - q^r k) \cdot p^{-r-s} \le k\sigma_k \sum_{r=0}^{\infty} \frac{q^{r+1}}{p^r} = \mathcal{O}_{p,q}(k\sigma_k),$$

as $q < p$ and the geometric sum converges. We can thus take $\mathsf{d}_{\boldsymbol{\Sigma}}(n) = \mathcal{O}_{p,q}(n)$. For $\lambda_\star$, using that

$$\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+t\sigma_n \boldsymbol{I})^{-1}\right) = \sum_{l=0}^{\infty} \frac{(q^{l+1}-q^l)p^{-l}}{p^{-l}+tp^{-s_\star}} = \left(q^{s_\star+1}-q^{s_\star}\right) \cdot \sum_{l=0}^{\infty} \frac{q^{l-s_\star}}{1+tp^{l-s^\star}}.$$

Since $s^\star \to \infty$ as n tends to infinity, we have

$$\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+t\sigma_n \boldsymbol{I})^{-1}\right) = n\rho_\star^{-1} \cdot G_{p,q,1}(t)\big(1+o_n(1)\big).$$

Hence

$$\lim_{n\to\infty} f_n(t;\lambda) = 1 - \nu t^{-1} - \rho_\star^{-1} \cdot G_{p,q,1}(t).$$

While the right hand side is increasing in t ranging in $(-\infty, 1)$. There exists a unique $\mathsf{c}_\star = \mathsf{c}_\star(\nu)$ solving

$$\lim_{n\to\infty} f_n(\mathsf{c}_\star;\lambda) = 0,$$

and substituting into Eq. (5) yields

$$\lambda_\star = \mathsf{c}_\star \sigma_n (1+o_n(1)).$$

Next we compute $\rho(\lambda)$ from Eq. (23),

$$\rho(\lambda) = \frac{\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\Sigma}^{\frac{1}{2}} (\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}}{\|\boldsymbol{\theta}\|^2 \, \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \boldsymbol{I})^{-1}\right)} = \frac{\sum_{l=1}^{\infty} (\sigma_l+\lambda_\star)^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2}{n(1-\nu\mathsf{c}_\star^{-1}) \sum_{l=1}^{\infty} \sigma_l^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2} \cdot (1+o_n(1))$$
$$= \frac{\sum_{l=1}^{\infty} \sigma_n (\sigma_l+\lambda_\star)^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2}{n(1-\nu\mathsf{c}_\star^{-1}) \sum_{l=1}^{\infty} \sigma_n \sigma_l^{-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l \rangle^2} \cdot (1+o_n(1))$$

$$= \frac{\sum_{s=0}^{\infty} p^{s-s_\star}/(1+\mathsf{c}_\star p^{s-s_\star}) \sum_{l=q^s}^{q^{s+1}-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l\rangle^2}{n(1-\nu\mathsf{c}_\star^{-1}) \sum_{s=0}^{\infty} p^{s-s_\star} \sum_{l=q^s}^{q^{s+1}-1} \langle \boldsymbol{\beta}, \boldsymbol{v}_l\rangle^2} \cdot (1+o_n(1))$$
$$= \frac{\int_0^\infty p^{x-s_\star}/(1+\mathsf{c}_\star p^{x-s_\star})\, \mathrm{d}F_{\boldsymbol{\beta}}(x)}{n(1-\nu\mathsf{c}_\star^{-1}) \int_0^\infty p^{x-s_\star}\, \mathrm{d}F_{\boldsymbol{\beta}}(x)} \cdot (1+o_n(1))\,.$$

Given the "rapid-decay" property

$$\int_0^\infty p^{x-s_\star}\, \mathrm{d}F_{\boldsymbol{\beta}}(x) = \mathcal{O}\left(n^{1-\theta} \int_0^\infty p^{x-s_\star}(1+\mathsf{c}_\star p^{x-s_\star})^{-1}\, \mathrm{d}F_{\boldsymbol{\beta}}(x)\right),$$

we have $\rho(\lambda) = \Omega(n^{-2+\theta})$ and Proposition 4.3 holds with $\gamma = (1-\theta)/3$, implying that $\mathscr{V}_{\boldsymbol{X}}(\lambda) = \mathsf{V}_n(\lambda)(1+o_n(1))$ and $\mathscr{B}_{\boldsymbol{X}}(\lambda) = \mathsf{B}_n(\lambda)(1+o_n(1))$.

To compute the effective variance $\mathsf{V}_n(\lambda)$, we first note that

$$\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+t\sigma_n\boldsymbol{I})^{-2}\right) = \left(q^{s_\star+1}-q^{s_\star}\right)\cdot \sum_{l=0}^{\infty} \frac{q^{l-s_\star}}{(1+tp^{l-s^\star})^2} = n\rho_\star^{-1}\cdot G_{p,q,2}(t)\big(1+o_n(1)\big)\,.$$

Thus

$$\mathsf{V}_n(\lambda) = \frac{\tau^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)}{n-\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right)} = \frac{G_{p,q,2}(\mathsf{c}_\star)\tau^2}{\rho_\star - G_{p,q,2}(\mathsf{c}_\star)}\big(1+o_n(1)\big)\,.$$

For the bias term, we have

$$\begin{aligned}
\lambda_\star^2\langle\boldsymbol{\beta}, (\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\boldsymbol{\Sigma}\boldsymbol{\beta}\rangle &= (1+o_n(1))\mathsf{c}_\star^2\sigma_n \sum_{l=0}^{\infty} \frac{\sigma_l\sigma_n}{(\sigma_l+\mathsf{c}_\star\sigma_n)^2}\langle\boldsymbol{\beta},\boldsymbol{v}_l\rangle^2 \\
&= (1+o_n(1))\mathsf{c}_\star^2\sigma_n \sum_{s=0}^{\infty}\left\{ \frac{p^{s-s_\star}}{(1+\mathsf{c}_\star p^{s-s_\star})^2}\cdot \sum_{l=q^s}^{q^{s+1}-1}\langle\boldsymbol{\beta},\boldsymbol{v}_l\rangle^2\right\} \\
&= \mathsf{c}_\star^2\sigma_n \int_0^\infty \frac{p^{x-s_\star}}{(1+\mathsf{c}_\star p^{x-s_\star})^2}\,\mathrm{d}F_{\boldsymbol{\beta}}(x)\,\big(1+o_n(1)\big)\,.
\end{aligned}$$

We conclude the proof for $\mathsf{B}_n(\lambda)$ by substituting in $n^{-1}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\boldsymbol{I})^{-2}\right) = (1+o_n(1))\rho_\star^{-1}G_{p,q,2}(\mathsf{c}_\star)$.