The Interplay Between Implicit Bias and Benign Overfitting in Two-Layer Linear Networks

Niladri S. Chatterji
Computer Science Department
Stanford University
niladri@cs.stanford.edu

Philip M. Long
Google
plong@google.com

Peter L. Bartlett
University of California, Berkeley & Google
peter@berkeley.edu

May 27, 2022

Abstract

The recent success of neural network models has shone light on a rather surprising statistical phenomenon: statistical models that perfectly fit noisy data can generalize well to unseen test data. Understanding this phenomenon of *benign overfitting* has attracted intense theoretical and empirical study. In this paper, we consider interpolating two-layer linear neural networks trained with gradient flow on the squared loss and derive bounds on the excess risk when the covariates satisfy sub-Gaussianity and anti-concentration properties, and the noise is independent and sub-Gaussian. By leveraging recent results that characterize the implicit bias of this estimator, our bounds emphasize the role of both the quality of the initialization as well as the properties of the data covariance matrix in achieving low excess risk.

Contents

1 Introduction

Understanding benign overfitting—the phenomenon where statistical models predict well on test data despite perfectly fitting noisy training data [see, e.g., Zha+17; Bel+19; BMR21; Bel21]—has recently attracted intense attention. One line of work has focused on understanding this phenomenon in relatively simple models such as linear regression [KLS20; Has+19; Bar+20; Mut+20b; NDR20; CL20; WX20; TB20; BSW20; CLG20; Koe+21] including with random features [Has+19; Yan+20; LZG21], linear classification [Mon+19; CL21; LS20; Mut+20a; HMX21; DKT19; WT21], kernel regression [LR20; MM19; LRZ20] and simplicial nearest neighbor methods [BHM18].

A complementary line of work [Sou+18; JT19; Gun+17; NSS19; Gun+18b; Gun+18a; YKM21; Azu+21] has formalized the argument [NTS15] that, even when no explicit regularization is used in training these models, there is nevertheless implicit regularization encoded in the choice of the optimization method, loss function and initialization. They argue that this implicit bias is critical in determining the generalization properties of the learnt model.

Recently, Azulay et al. [Azu+21] characterized the implicit bias of gradient flow applied to two-layer linear neural networks with the squared loss. More concretely, the setting is as follows. Given n data points $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, let $\mathbf{y} := (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ and $X := (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times p}$. They studied two-layer linear networks, with m hidden units, and weights $a \in \mathbb{R}^m$ and $W \in \mathbb{R}^{m \times p}$, that map an input $x \in \mathbb{R}^p$ to the scalar

$$a^\top W x.$$

Let $\theta = a^\top W \in \mathbb{R}^p$ denote the standard parameterization of the resulting linear map. A two-layer linear network with parameters $\{a, W\}$ is said to be *balanced* if

$$aa^\top - WW^\top = 0.$$

Azulay et al. [Azu+21] showed in Proposition 1 that, starting from a balanced initial point $(a(0), W(0))$, if the gradient flow converges to a solution that perfectly fits the data, then the solution can be characterized as follows:

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \|\theta\|^{3/2} - \frac{\theta(0)^\top \theta}{\sqrt{\|\theta(0)\|}}, \qquad \text{s.t., } \mathbf{y} = X\theta. \tag{1}$$

In this paper, we study the generalization properties of this solution in the overparameterized regime, where such interpolation is possible. We prove upper bounds on the excess risk and show that it depends both on the properties of the eigenstructure of population covariance matrix—as in the case of the minimum ℓ_2-norm interpolant (ordinary least squares) [Bar+20; TB20]—and also on the quality of the initialization $\theta(0)$. In particular, we show that to drive the excess risk to zero, it suffices if the number of samples is large relative to the trace of the population covariance matrix and also that the number of "small" eigenvalues is large relative to n. Our bounds also show that the excess risk can be smaller as a rescaling of $\theta(0)$ gets closer to the optimal linear predictor.

An overview of the techniques that drive our analysis is as follows. We begin by showing that the predictor $\hat{\theta}$ can be viewed as a perturbation of the ordinary least squares solution in the subspace orthogonal to the row span of X. To characterize this perturbation we find that it is important to derive upper and lower bounds on $\mathrm{Tr}((XX^\top)^{-1})$. To do this, as done in past work, we instead bound the trace of the "tail" of the matrix—the submatrix formed by the many low variance directions—and show that it not only concentrates but also provides a good approximation for the trace of the inverse of the entire matrix $\mathrm{Tr}((XX^\top)^{-1})$.

Along the way we derive a new multiplicative high-probability lower bound on the least singular value of a non-isotropic rectangular random matrix (Lemma 4.5). We could not find such a result in the literature. The most closely related work that we know of [see RV10, and references therein], characterizing the "hard edge" of a random matrix, has focused on the most difficult case of isotropic square matrices.

The remainder of the paper is organized as follows. In Section 2 we introduce notation and definitions. In Section 3 we present our results. We provide a proof of our main result, Theorem 3.1, in Section 4 and prove our lower bound in Section 5. We conclude with a discussion in Section 6.

2 Preliminaries

This section includes notational conventions and a description of the setting.

2.1 Notation

Given a vector v, let $\|v\|$ denote its Euclidean norm. Given a matrix M, let $\|M\|$ denote its Frobenius norm and $\|M\|_{op}$ denote its operator norm. For any $j \in \mathbb{N}$, we denote the set $\{1, \ldots, j\}$ by $[j]$. Given a symmetric matrix $M \in \mathbb{R}^{p\times p}$ we let $\mu_1(M) \geq \ldots \geq \mu_p(M)$ denote its eigenvalues. We let I_p denote the identity matrix in p dimensions. Given any vector $v \in \mathbb{R}^p$, we let $v_{1:j} \in \mathbb{R}^p$ denote the vector obtained by zeroing out the last $p - j$ coordinates of v and let $v_{j+1:p} \in \mathbb{R}^p$ denote the vector obtained by zeroing out the first j coordinates. Given a symmetric positive semidefinite matrix $M \in \mathbb{R}^{p\times p}$, let $M_{1:j} \in \mathbb{R}^{p\times p}$ be the matrix formed by zeroing out the last $p - j$ rows and columns of M, and let $M_{j+1:p} \in \mathbb{R}^{p\times p}$ be the matrix formed by zeroing out the first j rows and columns. We let $\|v\|_M := \sqrt{v^\top M v}$ denote the matrix norm of v with respect to the matrix M. We use the standard "big Oh notation" [see, e.g., Cor+09]. We will use $c, c', c_1, \ldots$ to denote positive absolute constants, which may take different values in different contexts.

2.2 The setting

Throughout the paper we assume that $p > n$. Although we assume throughout that the input dimension p is finite, it is straightforward to extend our results to infinite p.

For random $(x, y) \in \mathbb{R}^p \times \mathbb{R}$, let

$$\theta^\star \in \arg\min_{\theta \in \mathbb{R}^p} \mathbb{E}\left[(y - x^\top \theta)^2\right]$$

be an arbitrary optimal linear regressor. We assume that x is mean zero and let $\Sigma := \mathbb{E}[xx^\top]$ denote the covariance matrix of the features. Without loss of generality, we will assume that the covariance matrix is diagonal and its eigenvalues are arranged in descending order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$. (Note that such a covariance matrix can always be obtained by a rotation and permutation, and the estimator (1) is correspondingly transformed.) Recall that $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is the vector of responses and $X = (x_1, \ldots, x_n)^\top$ is the data matrix. Define $\boldsymbol{\varepsilon} = (y_1 - x_1^\top \theta^\star, \ldots, y_n - x_n^\top \theta^\star)^\top = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ to be the vector of noise.

We make the following assumptions:

(A.1) the samples $(x_1, y_1), \ldots, (x_n, y_n)$ and (x, y) are drawn i.i.d.;

(A.2) the features x and responses y are mean-zero;

(A.3) the features $x = \Sigma^{1/2} u$, where u has components that are independent σ_x^2-sub-Gaussian random variables with σ_x a positive constant, that is, for all $\phi \in \mathbb{R}^p$

$$\mathbb{E}\left[\exp\left(\phi^\top u\right)\right] \leq \exp\left(\sigma_x^2 \|\phi\|^2 / 2\right);$$

(A.4) there is an absolute constant c such that, for any unit vector $\phi \in \mathbb{S}^{n-1}$ and any $a \leq b \in \mathbb{R}$

$$\mathbb{P}\left[(\Sigma^{-1/2} X^\top \phi)_i \in [a, b]\right] \leq c|b - a|$$

for all $i \in [p]$;

(A.5) the difference $y - x^\top \theta^\star$ is σ_y^2-sub-Gaussian, conditionally on x, with σ_y a positive constant, that is, for all $\phi \in \mathbb{R}$

$$\mathbb{E}_y\left[\exp\left(\phi(y - x^\top \theta^\star)\right) \mid x\right] \leq \exp\left(\sigma_y^2 \phi^2 / 2\right)$$

(note that this implies that $\mathbb{E}[y \mid x] = x^\top \theta^\star$);

(A.6) for all x, the conditional variance of $y - x^\top \theta^\star$ is

$$\mathbb{E}_y\left[(y - x^\top \theta^\star)^2 \mid x\right] = \sigma^2$$

where σ is a positive constant.

We emphasize that σ_x, σ_y and σ are absolute constants, independent of all other problem parameters (n, p and Σ). All the constants going forward may depend on the value of these constants.

The assumptions stated above are satisfied in the case where u is generated from a mean-zero isotropic log-concave distribution with sub-Gaussian, independent entries and the noise $y - x^\top \theta^\star$ is independent and sub-Gaussian. We note that Assumptions A.(A.1)-A.(A.3), A.(A.5)-A.(A.6) are standard in the literature of benign overfitting in linear models [see, e.g., Bar+20]. We make an additional small-ball probability assumption (Assumption A.(A.4)) which allows us to derive a sharper multiplicative lower tail bound for the minimum eigenvalue of the submatrices of X (Lemma 4.5).

Given the training samples define the excess risk of an estimate $\theta \in \mathbb{R}^p$ to be

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

where x, y are independent test samples.

Define the shorthand

$$w := \frac{\theta(0)}{\sqrt{\|\theta(0)\|}},$$

so that the estimator described in equation (1) can be written as the solution to a constrained convex program given by

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \|\theta\|^{3/2} - w^\top\theta, \qquad \text{s.t., } \mathbf{y} = X\theta. \tag{2}$$

We let $UDV^\top = X$ be the singular value decomposition of X where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ are unitary matrices and $D \in \mathbb{R}^{n\times p}$ is a rectangular diagonal matrix with its eigenvalues in descending order. By Lemma A.1, we know that the rank of D is n. We let $D^\dagger \in \mathbb{R}^{p\times n}$ denote the pseudo-inverse of D. Since D has rank n, the bottom $p - n$ rows of $D^\dagger$ are identically zero.

Also define

$$\widetilde{\mathbf{y}} := D^\dagger U^\top \mathbf{y}, \qquad \widetilde{w} := V^\top w \qquad \text{and} \qquad \widetilde{\theta} := V^\top\hat{\theta}. \tag{3}$$

We will use the following definitions of the "effective rank" from [Bar+20].

Definition 2.1. *Given a subset $S \subseteq [p]$, define $s(S) := \sum_{i\in S}\lambda_i$, and define the following ranks of the covariance matrix Σ with eigenvalues $\lambda_1, \ldots, \lambda_p$:*

$$r(S) := \frac{s(S)}{\max_{i\in S}\lambda_i} \qquad \textit{and} \qquad R(S) := \frac{s(S)^2}{\sum_{i\in S}\lambda_i^2}.$$

Further given any $j \in [p]$, with some abuse of notation, define $s_j := \sum_{i>j}\lambda_i$ and

$$r_j := \frac{s_j}{\lambda_{j+1}} \qquad \textit{and} \qquad R_j := \frac{s_j^2}{\sum_{i>j}\lambda_i^2}.$$

The following lemma [Bar+20, Lemma 5] relates these different effective ranks.

Lemma 2.2. *For any subset $S \subseteq [p]$ the ranks defined above satisfy the following:*

$$r(S) \le R(S) \le r(S)^2.$$

We define the index k below. The value of k shall help determine what we consider the "tail" of the covariance matrix.

Definition 2.3. *For a large enough constant b (that will be fixed henceforth), define*

$$k := \min\{j \ge 0 : r_j \ge bn\},$$

where the minimum of the empty set is defined as ∞.

Finally we define ψ, which is a rescaling of w.

Definition 2.4. *Define*

$$\psi := \frac{2\sqrt{\sigma}n^{1/4}}{3s_k^{1/4}}w = \frac{2\sqrt{\sigma}n^{1/4}}{3s_k^{1/4}}\frac{\theta(0)}{\sqrt{\|\theta(0)\|}}.$$

3 Main results

In this section we present our main result, Theorem 3.1, which is an excess risk bound for the estimator $\hat{\theta}$. It is proved in Section 4.

Theorem 3.1. *Under Assumptions (A.1)-(A.6), there exist constants $c_0, \ldots, c_7$ such that for any $\delta \in (e^{-c_0\sqrt{n}}, 1-c_1e^{-c_2 n})$, if $p \geq c_3(n+k)$, $n \geq c_4 \max\{k, s_k\}$ and $\|\theta^\star\|, \|w\| \leq c_5$ then with probability at least $1 - c_6\delta$*

$$\mathsf{Risk}(\hat{\theta}) \leq \mathsf{Bias} + \mathsf{Variance} + \Xi,$$

where

$$\mathsf{Bias} \leq c_7 \left(\|(\theta^\star - \psi)_{1:k}\|^2_{\Sigma^{-1}_{1:k}} \left(\frac{s_k}{n}\right)^2 + \|(\theta^\star - \psi)_{k+1:p}\|^2_{\Sigma_{k+1:p}} \right) \leq \frac{2c_7\|\theta^\star - \psi\|^2 s_k}{n};$$

$$\mathsf{Variance} \leq c_7 \log(1/\delta) \left(\frac{k}{n} + \frac{n}{R_k}\right);$$

$$\Xi \leq c_7 \lambda_1 \|\psi\|^2 \left[\frac{n}{R_k} + \frac{n^2}{r_k^2} + \frac{s_k}{n} + \frac{\log(1/\delta)}{n} + \frac{k^2}{n^2}\right] \max\left\{\sqrt{\frac{r_0}{n}}, \frac{r_0}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}.$$

Note that Bias goes to zero as $\psi \to \theta^\star$. In the upper bounds on the excess risk for linear models with a standard (one-layer) parameterization [see TB20, Theorem 1], the corresponding term scales with the square of the norm of $\theta^\star$ rather than $(\theta^\star - \psi)$. If one has a "guess" $\widehat{\psi}$ for $\theta^\star$, then—given knowledge of σ, s_k and n—it is possible to set the initialization as follows:

$$\theta(0) = \frac{9\widehat{\psi}\|\widehat{\psi}\|}{4}\sqrt{\frac{s_k}{\sigma^2 n}};$$

which ensures that $\psi = \widehat{\psi}$. Very accurate prior guesses $\widehat{\psi}$ of θ^* are rewarded with a very small value of the Bias term.

Next, we note that the upper bound on $\mathsf{Variance}$ here is identical to the upper bound on the variance for the minimum ℓ_2-norm interpolant (the OLS estimator) [see Bar+20, Theorem 4]. The initialization $\theta(0)$ (through ψ) only affects the conditional bias of the estimator here, but leaves the conditional variance the same as the OLS solution. This is because, as we will show below in Lemma 4.1, $\hat{\theta}$ can be expressed as a perturbation to the OLS estimator in the subspace orthogonal to the row span of X. It turns out that the variance only depends on behavior of $\hat{\theta}$ in the subspace spanned by the data, where $\hat{\theta}$ and the OLS solution are identical.

As mentioned, $\hat{\theta}$ is a perturbation of the OLS estimator. In particular, it is perturbed by $\alpha^\star \mathsf{Proj}^{\perp}_X(w)$, where $\mathsf{Proj}^{\perp}_X(w)$ is the projection of w onto the subspace orthogonal to the row span of X and $\alpha^\star$ is a scalar random variable that depends on the data. We shall demonstrate in Lemma 4.3 that, under the setting specified by the theorem, $\alpha^\star$ concentrates around $\frac{2\sqrt{\sigma}n^{1/4}}{3s_k^{1/4}}$. The final term in the excess risk bound, Ξ, corresponds to the fluctuation of $\alpha^\star$. We might think of $\theta(0)$ (and hence w) as being constructed from ψ and an estimate of α^*; from this point of view, Ξ accounts for the contribution to the excess risk arising from the error in estimating $\alpha^\star$. Next we derive sufficient conditions for the excess risk to go to zero as $n, p \to \infty$. Consider the case where λ_1, $\|\theta^\star\|$, $\|\psi\|$ and $\log(1/\delta)$ are all bounded by constants. (In the case of $\|\psi\|$, this can be achieved by appropriately scaling $\theta(0)$.) For Bias to go to zero it suffices if

$$\frac{s_k}{n} \to 0.$$

For Variance to decrease to zero it suffices if

$$\frac{k}{n} \to 0 \quad \text{and} \quad \frac{n}{R_k} \to 0.$$

Finally, for Ξ to approach zero it suffices for

$$\frac{r_0}{n} \to 0$$

which also implies the condition $\frac{s_k}{n} \to 0$ needed to control the Bias term. (To see that $\frac{r_0}{n} \to 0$ suffices, recall that we have assumed that $\lambda_1, \log(1/\delta)$ and $\|\psi\|$ are constants. Further, the quantity in the square brackets of our bound on Ξ is at most a constant, which can be seen as follows. The definition of r_k implies that $r_k \geq bn$, and Lemma 2.2 gives $R_k \geq r_k$. Finally, we have assumed that $n \geq c \max\{k, s_k\}$.) To summarize, if $\frac{k}{n}, \frac{r_0}{n}, \frac{n}{R_k} \to 0$, the excess risk of this estimator approaches zero. Some discussion and examples of when this condition is satisfied are given in [Bar+20; TB20].

To develop intuition, we consider a special case of Theorem 3.1 defined as follows.

Definition 3.2 ((k, ε)-spike model)**.** *For $\varepsilon > 0$ and $k \in \mathbb{N}$, a (k, ε)-spike model is a setting where the eigenvalues of Σ are $\lambda_1 = \ldots = \lambda_k = 1$ and $\lambda_{k+1} = \ldots = \lambda_p = \varepsilon$.*

Instantiating Theorem 3.1 in the case of the (k, ε)-spike model, and removing some dominated terms, yields the following corollary.

Corollary 3.3. *Under Assumptions (A.1)-(A.6), there exist constants $c_0, \ldots, c_8$ such that in the (k, ε)-spike model for any $\delta \in (e^{-c_0\sqrt{n}}, 1-c_1 e^{-c_2 n})$, if $p > c_3(n+k)$, $n \geq c_4 \max\{k, \varepsilon p\}$ and $\|\theta^\star\|, \|w\| \leq c_5$ then with probability at least $1 - c_6\delta$*

$$\mathsf{Risk}(\hat{\theta}) \leq \mathsf{Bias} + \mathsf{Variance} + \Xi,$$

where

$$\mathsf{Bias} \leq c_7 \left(\|(\theta^\star - \psi)_{1:k}\|^2 \left(\frac{\varepsilon p}{n}\right)^2 + \varepsilon \|(\theta^\star - \psi)_{k+1:p}\|^2 \right) \leq c_8 \|\theta^\star - \psi\|^2 \left(\frac{\varepsilon p}{n}\right);$$

$$\mathsf{Variance} \leq c_7 \log(1/\delta) \left(\frac{k}{n} + \frac{n}{p} \right);$$

$$\Xi \leq c_7 \lambda_1 \|\psi\|^2 \left[\frac{n}{p} + \frac{\varepsilon p}{n} + \frac{\log(1/\delta)}{n} + \frac{k^2}{n^2} \right] \max\left\{ \sqrt{\frac{k + \varepsilon p}{n}}, \sqrt{\frac{\log(1/\delta)}{n}} \right\}.$$

Again, in the case where $\lambda_1, \|\psi\|$ and $\log(1/\delta)$ are bounded by constants, a sufficient condition for the excess risk to decrease to zero is when $\frac{\varepsilon p}{n}, \frac{k}{n}, \frac{n}{p} \to 0$.

Next we establish a lower bound. It is proved in Section 5.

Proposition 3.4. *If $a(0)$ and $W(0)$ are chosen randomly, independent of X and $\mathbf{y}$, so that the distribution of $a(0)^\top W(0)$ is symmetric about the origin, then*

$$\mathbb{E}_{a(0), W(0), X, \mathbf{y}} [\mathsf{Risk}(\hat{\theta})] \geq \mathbb{E}\left[\theta^{\star\top} B \theta^\star\right] + \sigma^2 \mathbb{E}\left[\mathrm{Tr}(C)\right],$$

where

$$B := \left(I - X^\top (XX^\top)^{-1} X\right) \Sigma \left(I - X^\top (XX^\top)^{-1} X\right) \quad \textit{and}$$
$$C := (XX^\top)^{-1} X \Sigma X^\top (XX^\top)^{-1}.$$

Remark 3.5. *For the distribution of $a(0)^\top W(0)$ to be symmetric about the origin, it suffices that $a(0)$ and $W(0)$ are chosen independently, and that either the distribution of $a(0)$ is symmetric about the origin, or the distribution of $W(0)$ is.*

Remark 3.6. *Bartlett et al. [Bar+20] proved that $\mathbb{E}[\mathrm{Tr}(C)] \geq c\left(\frac{k}{n}+\frac{n}{R_k}\right)$ for a constant c. Tsigler and Bartlett [TB20] proved a lower bound on $\mathbb{E}[\theta^{\star\top} B\theta^\star]$ under the assumption that the signs of the components of $\theta^\star$ are chosen uniformly at random. For the case that $\psi = 0$, their lower bound matches the upper bound on* Bias *from Theorem 3.1 of this paper under the assumptions of that theorem. However, there is a gap in the upper and lower bounds when $\psi \neq 0$.*

4 Proof details

The proof of Theorem 3.1 is built up in parts. First, in Lemma 4.1 we show that $\hat{\theta}$ can be viewed as a random perturbation of the ordinary least squares (OLS) solution in the subspace orthogonal to the row span of X. In Lemma 4.2, we show that the excess risk can be decomposed into two terms, one that can bounded above by $\mathsf{Variance}$ and the other that is upper bounded by $\mathsf{Bias} + \Xi$. The next piece is Lemma 4.3 which is crucial in helping us characterize the perturbation to the OLS solution. To do this we first present concentration inequalities in Section 4.1, then we establish upper and lower bounds on $\mathrm{Tr}((XX^\top)^{-1})$ in Section 4.2, and finally prove Lemma 4.3 in Section 4.3. We finish by combining all of these elements to prove the theorem in Section 4.4. Throughout this section we assume that the assumptions made in Theorem 3.1 are in force.

A note about constants. As mentioned earlier, we will not always provide specific constants. The constants $c_1, c_2, \ldots$ in our proofs are independent of the problem parameters, but they can depend on one another. It will not be hard to verify, however, that the constraints on their values are satisfiable. When we write "c_i is large enough", this should be understood to be relative to the constants previously introduced in the proof not including b, the constant used in the definition of r_k. Loosely speaking, b is chosen last: it should be taken to be large relative to all other constants.

We begin with the following lemma that provides a closed-form formula for $\hat{\theta}$ as a perturbation of the ordinary least squares solution.

Lemma 4.1. *The solution $\hat{\theta}$ can be expressed as follows:*

$$\hat{\theta} = \hat{\theta}_{\mathsf{OLS}} + \alpha^\star (I - X^\top (XX^\top)^{-1} X) w,$$

where $\hat{\theta}_{\mathsf{OLS}} = X^\top \left(XX^\top\right)^{-1} \mathbf{y}$ is the ordinary least squares solution (minimum ℓ_2-norm interpolant) and

$$\alpha^\star = \sqrt{\frac{8\|\widetilde{w}_{n+1:p}\|^2 + \sqrt{64\|\widetilde{w}_{n+1:p}\|^4 + 1296\|\widetilde{\mathbf{y}}\|^2}}{81}}. \tag{4}$$

where $\widetilde{w}$ and $\widetilde{\mathbf{y}}$ are defined in (3).

Proof. Recall that $X = UDV^\top$ is the singular value decomposition of X. Therefore,

$$\begin{aligned}
& \hat{\theta} \in \underset{\theta \in \mathbb{R}^p}{\arg\min}\ \|\theta\|^{3/2} - w^\top \theta, \qquad \text{s.t., } \mathbf{y} = X\theta \\
\iff & \hat{\theta} \in \underset{\theta \in \mathbb{R}^p}{\arg\min}\ \|\theta\|^{3/2} - w^\top \theta, \qquad \text{s.t., } D^\dagger U^\top \mathbf{y} = D^\dagger U^\top X\theta \\
\iff & \hat{\theta} \in \underset{\theta \in \mathbb{R}^p}{\arg\min}\ \|\theta\|^{3/2} - w^\top \theta, \qquad \text{s.t., } \widetilde{\mathbf{y}} = D^\dagger U^\top X\theta \\
\iff & V^\top\hat{\theta} \in \underset{\theta \in \mathbb{R}^p}{\arg\min}\ \|V^\top\theta\|^{3/2} - (w^\top V) V^\top \theta, \qquad \text{s.t., } \widetilde{\mathbf{y}} = D^\dagger U^\top XV(V^\top\theta) \\
& \qquad\qquad \text{(since the Euclidean norm is rotation invariant and } VV^\top = I_p) \\
\iff & \widetilde{\theta} \in \underset{\theta \in \mathbb{R}^p}{\arg\min}\ \|\theta\|^{3/2} - \widetilde{w}^\top \theta, \qquad \text{s.t., } \widetilde{\mathbf{y}} = D^\dagger U^\top XV\theta.
\end{aligned}$$

Since the bottom $p-n$ rows of $D^\dagger$ are identically zero, the vector $\widetilde{\mathbf{y}}$ has the form $(\widetilde{\mathbf{y}}_1, \ldots, \widetilde{\mathbf{y}}_n, 0, \ldots, 0)^\top$. Since the SVD of $X = UDV^\top$ we have that

$$D^\dagger U^\top XV\theta = D^\dagger U^\top UDV^\top V\theta = D^\dagger D\theta = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} \theta.$$

Hence, for the constraint to be satisfied, the first n coordinates of $\widetilde{\theta}$ are required to be equal to $(\widetilde{\mathbf{y}}_1, \ldots, \widetilde{\mathbf{y}}_n)$, and the remaining coordinates of $\widetilde{\theta}$ can be anything. That is, the constraints are satisfied when $\widetilde{\theta} = (\widetilde{\mathbf{y}}_1, \ldots, \widetilde{\mathbf{y}}_n, 0, \ldots, 0)^\top + \phi$, for some $\phi \in \mathbb{R}^p$ with its first n coordinates all equal to zero.

To find this optimal vector $\phi^\star$ we can now proceed to solve the following *unconstrained* optimization problem:

$$\phi^\star \in \underset{\phi \in \mathbb{R}^p}{\arg\min} \left(\|\widetilde{\theta}_{1:n}\|^2 + \|\phi\|^2 \right)^{3/4} - \sum_{j=n+1}^{p} \widetilde{w}_j \phi_j.$$

Since the first term in the objective function above is rotationally invariant, it must be the case that the minimizer has the form $\phi^\star = \alpha^\star \widetilde{w}_{n+1:p}$, for some $\alpha^\star > 0$. That is, it is positively aligned with the tail of the vector $\widetilde{w}$. (If $\phi^\star$ was not in the span of $\widetilde{w}_{n+1:p}$, removing the projection of $\phi^\star$ in the subspace orthogonal to this direction would improve the norm without affecting the second term, and if $\phi^\star \cdot \widetilde{w}_{n+1:p} < 0$, then $-\phi^\star$ would be a better solution than $\phi^\star$.) In particular, we have

$$\widetilde{w}_{n+1:p} = 0 \Rightarrow \phi^\star = 0.$$

Otherwise, $\phi^\star = \alpha^\star \widetilde{w}_{n+1:p}$ for the solution $\alpha^\star$ of the following one-dimensional problem:

$$\alpha^\star \in \underset{\alpha > 0}{\arg\min} \left(\|\widetilde{\theta}_{1:n}\|^2 + \alpha^2 \|\widetilde{w}_{n+1:p}\|^2 \right)^{3/4} - \alpha \|\widetilde{w}_{n+1:p}\|^2.$$

To simplify notation, let $\rho := \|\widetilde{\theta}_{1:n}\|^2$ and $\zeta := \|\widetilde{w}_{n+1:p}\|^2 > 0$. The first derivative of the objective function is as follows:

$$\frac{d}{d\alpha} \left[\left(\rho + \zeta\alpha^2\right)^{3/4} - \alpha\zeta \right] = \frac{3\zeta\alpha}{2\left(\rho + \zeta\alpha^2\right)^{1/4}} - \zeta.$$

Setting this first derivative equal to zero we get that

$$\begin{aligned}
&\frac{3\zeta\alpha}{2\left(\rho+\zeta\alpha^2\right)^{1/4}}-\zeta=0\\
\iff&\frac{3\alpha}{2\left(\rho+\zeta\alpha^2\right)^{1/4}}-1=0\\
\iff&\frac{3\alpha}{2}=\left(\rho+\zeta\alpha^2\right)^{1/4}\\
\iff&\frac{81\alpha^4}{16}=\rho+\zeta\alpha^2 \qquad\qquad \text{(because } \alpha>0 \text{ at the optimum)}\\
\iff&81\alpha^4-16\zeta\alpha^2-16\rho=0.
\end{aligned}$$

We can view this as a quadratic equation in α^2, so

$$\alpha^2=\frac{16\zeta\pm\sqrt{256\zeta^2+5184\rho}}{162}=\frac{16\zeta\left(1\pm\sqrt{1+\frac{81\rho}{4\zeta^2}}\right)}{162},$$

but the solution with the negative sign can be ignored since α^2 must be positive. Taking square roots we get that

$$\alpha=\pm\sqrt{\frac{8\zeta\left(1+\sqrt{1+\frac{81\rho}{4\zeta^2}}\right)}{81}}.$$

Again, we drop the negative solution since we know that $\alpha>0$ at the optimum. Thus we find that

$$\widetilde{\theta}=\widetilde{\mathbf{y}}+\alpha^\star\widetilde{w}_{n+1:p}$$

for

$$\begin{aligned}
\alpha^\star=\sqrt{\frac{8\zeta\left(1+\sqrt{1+\frac{81\rho}{4\zeta^2}}\right)}{81}}&=\sqrt{\frac{8\|\widetilde{w}_{n+1:p}\|^2\left(1+\sqrt{1+\frac{81\|\widetilde{\theta}_{1:n}\|^2}{4\|\widetilde{w}_{n+1:p}\|^4}}\right)}{81}}\\
&=\sqrt{\frac{8\|\widetilde{w}_{n+1:p}\|^2+\sqrt{64\|\widetilde{w}_{n+1:p}\|^4+1296\|\widetilde{\theta}_{1:n}\|^2}}{81}}\\
&=\sqrt{\frac{8\|\widetilde{w}_{n+1:p}\|^2+\sqrt{64\|\widetilde{w}_{n+1:p}\|^4+1296\|\widetilde{\mathbf{y}}\|^2}}{81}}.
\end{aligned}$$

Recall that by definition $\widetilde{\theta}=V^\top\hat{\theta}$, $\widetilde{\mathbf{y}}=D^\dagger U^\top\mathbf{y}$ and $\widetilde{w}=V^\top w$ and hence

$$\begin{aligned}
\hat{\theta}&=V\widetilde{\mathbf{y}}+\alpha^\star V\widetilde{w}_{n+1:p}\\
&=VD^\dagger U^\top\mathbf{y}+\alpha^\star V\widetilde{w}_{n+1:p}\\
&=X^\top\left(XX^\top\right)^{-1}\mathbf{y}+\alpha^\star V\widetilde{w}_{n+1:p}\\
&=\hat{\theta}_{\mathsf{OLS}}+\alpha^\star V\widetilde{w}_{n+1:p}\\
&=\hat{\theta}_{\mathsf{OLS}}+\alpha^\star V\begin{bmatrix}0_{n\times n} & 0_{n\times(p-n)}\\ 0_{(p-n)\times n} & I_{p-n}\end{bmatrix}V^\top w.
\end{aligned}$$

Recall that the SVD of X is $UDV^\top$, and the last $(p-n)$ columns of D are zero. Thus, the last $(p-n)$ rows of $V^\top$ span the null space of X.

Furthermore, $(I_p - X^\top(XX^\top)^{-1}X)$ represents the projection onto this null space of X. This can be seen as follows. First, any member u of this null space is mapped to itself (since $Xu = 0$). On the other hand, for each row x of X, $(I_p - X^\top(XX^\top)^{-1}X)x^\top = 0$, as

$$(I_p - X^\top(XX^\top)^{-1}X)X^\top = 0.$$

Recalling that the last $(p-n)$ rows of $V^\top$ span the null space of X, we have

$$V \begin{bmatrix} 0_{n\times n} & 0_{n\times(p-n)} \\ 0_{(p-n)\times n} & I_{p-n} \end{bmatrix} V^\top = I_p - X^\top(XX^\top)^{-1}X.$$

This wraps up our proof. ■

Armed with this formula for $\hat{\theta}$ we can now bound the excess risk.

Lemma 4.2. *The excess risk of $\hat{\theta}$ satisfies*

$$\mathsf{Risk}(\hat{\theta}) \le c(\theta^\star - \alpha^\star w)^\top B(\theta^\star - \alpha^\star w) + c\log(1/\delta)\mathrm{Tr}(C)$$

with probability at least $1-\delta$ over $\boldsymbol{\varepsilon}$, where

$$\begin{aligned} &B := \left(I - X^\top(XX^\top)^{-1}X\right)\Sigma\left(I - X^\top(XX^\top)^{-1}X\right) \quad \textit{and} \\ &C := (XX^\top)^{-1}X\Sigma X^\top(XX^\top)^{-1}. \end{aligned}$$

Proof. Since $\varepsilon = y - x^\top\theta$ is conditionally mean-zero given x,

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

Using the formula of $\hat{\theta}$ from Lemma 4.1, and because $\mathbf{y} = X\theta^\star + \boldsymbol{\varepsilon}$ we find that

$$\begin{aligned} \mathsf{Risk}(\hat{\theta}) &= \mathbb{E}_x\left[\left(x^\top\left(I - X^\top(XX^\top)^{-1}X\right)(\theta^\star - \alpha^\star w) - x^\top X^\top(XX^\top)^{-1}\boldsymbol{\varepsilon}\right)^2\right] \\ &\le 2\mathbb{E}_x\left[\left(x^\top\left(I - X^\top(XX^\top)^{-1}X\right)(\theta^\star - \alpha^\star w)\right)^2\right] + 2\mathbb{E}_x\left[\left(x^\top X^\top(XX^\top)^{-1}\boldsymbol{\varepsilon}\right)^2\right] \\ &= 2(\theta^\star - \alpha^\star w)^\top B(\theta^\star - \alpha^\star w) + 2\boldsymbol{\varepsilon}^\top C\boldsymbol{\varepsilon}. \end{aligned}$$

Now by [Bar+20, Lemma 19] we find that $2\boldsymbol{\varepsilon}^\top C\boldsymbol{\varepsilon} \le c'\sigma_y^2\log(1/\delta)\mathrm{Tr}(C) \le c\log(1/\delta)\mathrm{Tr}(C)$ with probability at least $1-\delta$. This completes the proof. ■

The following lemma provides upper and lower bounds on the value of $\alpha^\star$ that are tight up to the leading constant when p is large relative to $n+k$ and when n is sufficiently large relative to k and $s_k = \sum_{j>k}\lambda_j$.

Lemma 4.3. *There are constants $c_0, \dots, c_5$ such that for any $\delta \in (e^{-c_0\sqrt{n}}, 1)$, if $p \geq c_1(n+k)$, $n \geq c_2 \max\{k, s_k\}$ and $\|\theta^\star\|, \|w\| \leq c_3$ then with probability at least $1 - c_4\delta$,*

$$\left| \frac{\alpha^\star}{\frac{2\sqrt{\sigma}n^{1/4}}{3s_k^{1/4}}} - 1 \right| \leq c_5 \left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \sqrt{\frac{s_k}{n}} + \sqrt{\frac{\log(2/\delta)}{n}} + \frac{k}{n} \right].$$

Lemma 4.3 is proved over the next few subsections. As might be expected, for $\alpha^\star$ to reliably fall within a small interval, X must be well conditioned in some sense. We begin by establishing bounds on the singular values of submatrices of X in Sections 4.1, whose proofs are provided Appendix B. In Section 4.2, we show that the $\mathrm{Tr}((XX^\top)^{-1})$ is concentrated. Armed with these bounds, we build up our analysis of $\alpha^\star$ in stages in Section 4.3.

4.1 Bounds on the extreme singular values of submatrices of X

In this subsection, we will derive bounds on the largest and smallest singular value of a submatrix of X. Given a subset S of $[p]$, let $X_S \in \mathbb{R}^{n \times |S|}$ be a submatrix of X where only the columns with indices in S are included.

With this in place, we are now ready to prove our concentration results. We prove this lemma in Appendix B.1.

Lemma 4.4. *There exists a positive absolute constant c such that, for any subset $S \subseteq [p]$ and any $t \geq 0$, with probability at least $1 - 2e^{-t}$, for all $j \in \{1, \dots, \min(n, |S|)\}$*

$$\left| \mu_j(X_S X_S^\top) - s(S) \right| \leq cs(S) \left(\frac{t+n}{r(S)} + \sqrt{\frac{t+n}{R(S)}} \right).$$

This lemma provides an additive lower bound on the minimum singular value of submatrices of X. Next, we will provide a sharper multiplicative bound on the smallest singular value of such matrices. Its proof can be found in Appendix B.2.

Lemma 4.5. *There exist absolute positive constants $c_0, \dots, c_3$ such that given any subset $S \subseteq [p]$ if, $r(S) \geq c_0 n$ then for all $t < c_1 < 1$*

$$\mathbb{P}\left[\mu_n(X_S X_S^\top) \leq t \cdot s(S) \right] \leq (c_2 t)^{c_3 \cdot r(S)}.$$

This sharper multiplicative bound provides a much more refined lower tail probability estimate for the minimum eigenvalue than the previous additive bound in Lemma 4.4, especially when t is close to zero. This is useful in our analysis to control $\mathbb{E}[\mathrm{Tr}(XX^\top)^{-1}]$ which is in turn used to establish Lemma 4.6 that bounds $\mathrm{Tr}((XX^\top)^{-1})$.

4.2 Concentration of $\mathrm{Tr}((XX^\top)^{-1})$

In this subsection we shall prove the following lemma which shows that $\mathrm{Tr}((XX^\top)^{-1})$ concentrates.

Lemma 4.6. *There are positive constants $c_0, \dots, c_4$ such that, if $p \geq c_0(n+k)$ then with probability at least $1 - c_1 e^{-c_2 n}$*

$$\left| \mathrm{Tr}\left((XX^\top)^{-1} \right) - \frac{n}{s_k} \right| \leq \frac{c_3 n}{s_k} \left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \frac{k}{n} + e^{-c_4\sqrt{n}} \right].$$

The proof of Lemma 4.6 in turn requires some lemmas. To state them, we need some additional notation and definitions.

Recall that we have assumed without loss of generality that Σ is diagonal. Let $\lambda_1 \geq \ldots \geq \lambda_p$ be the elements of its diagonal, and define the random vectors

$$z_i := \frac{Xe_i}{\sqrt{\lambda_i}} \in \mathbb{R}^n.$$

These random vectors z_i have entries that are independent, σ_x^2-sub-Gaussian random variables [see Bar+20, Lemma 8]. Note that we can write the matrix

$$XX^\top = \sum_{i=1}^{p} \lambda_i z_i z_i^\top.$$

Definition 4.7. *Define the shorthand* $A := XX^\top$*, and define*

$$H := \sum_{i=1}^{k} \lambda_i z_i z_i^\top \qquad \text{and} \qquad T := \sum_{i=k+1}^{p} \lambda_i z_i z_i^\top.$$

Therefore $A = H + T$.

To prove Lemma 4.6 we shall prove the following four results:

- in Lemma 4.8, we show that $\mathrm{Tr}(A^{-1})$ is close to the $\mathrm{Tr}(T^{-1})$ with high probability;
- in Lemma 4.9, we show that $\mathbb{E}\left[\mathrm{Tr}(A^{-1})\right]$ is well approximated by $\mathbb{E}\left[\mathrm{Tr}(T^{-1})\right]$;
- in Lemma 4.10, we show that $\mathrm{Tr}(T^{-1})$ is close to its expectation with high probability;
- finally, in Lemma 4.11, we establish upper and lower bounds on $\mathbb{E}\left[\mathrm{Tr}(A^{-1})\right]$ that match up to leading constants.

By using these four results and the triangle inequality we shall demonstrate that $\mathrm{Tr}(A^{-1})$ is close to n/s_k with high probability and prove Lemma 4.6. Throughout this subsection we shall assume that the dimension $p \geq c_0(n+k)$, for a sufficiently large constant c_0. Under this condition, Lemma A.1 implies that the tail matrix T is full-rank and invertible.

4.2.1 $\mathrm{Tr}(A^{-1})$ is close to $\mathrm{Tr}(T^{-1})$

We begin by showing that $\mathrm{Tr}(A^{-1})$ is close to $\mathrm{Tr}(T^{-1})$ with high probability.

Lemma 4.8. *There exist positive constants* $c_0, \ldots, c_3$ *such that, for all* $\beta < c_0 < 1$*, with probability at least* $1 - 2\exp(-r_k/\beta^2) - (c_1\beta)^{c_2 \cdot r_k}$*,*

$$\left|\mathrm{Tr}(A^{-1}) - \mathrm{Tr}(T^{-1})\right| \leq \frac{c_3 k}{\beta^4 s_k}.$$

Proof. Recall that $A = XX^\top = \sum_{i=1}^k \lambda_i z_i z_i^\top + \sum_{i=k+1}^p \lambda_i z_i z_i^\top = H + T$. Let $u_1, \ldots, u_k \in \mathbb{R}^n$ be an orthonormal basis for the row span of H, and let $u_1, \ldots, u_n$ be an extension to a basis for $\mathbb{R}^n$. Write $U = [u_1, \ldots, u_n] = [E; F]$, where E is $n \times k$. Thus

$$\begin{aligned}
\mathrm{Tr}(A^{-1}) &\overset{(i)}{=} \mathrm{Tr}\left(U^\top A^{-1} U\right) \\
&= \mathrm{Tr}\left(\left[U^\top A U\right]^{-1}\right) \\
&= \mathrm{Tr}\left(\left[U^\top (H+T) U\right]^{-1}\right) \\
&= \mathrm{Tr}\left(\left[U^\top H U + U^\top T U\right]^{-1}\right) \\
&= \mathrm{Tr}\left(\left[\begin{pmatrix} E^\top H E & E^\top H F \\ F^\top H E & F^\top H F \end{pmatrix} + U^\top T U\right]^{-1}\right) \\
&\overset{(ii)}{=} \mathrm{Tr}\left(\left[\begin{pmatrix} E^\top H E & 0 \\ 0 & 0 \end{pmatrix} + U^\top T U\right]^{-1}\right) \\
&= \mathrm{Tr}\left(\left[\begin{pmatrix} E^\top \\ 0 \end{pmatrix} H \begin{pmatrix} E & 0 \end{pmatrix} + U^\top T U\right]^{-1}\right),
\end{aligned}$$

where (i) follows since U is a unitary matrix, (ii) follows since the columns of F are outside the span of H. Continuing, we apply the Sherman-Morrison-Woodbury identity to get that

$$\begin{aligned}
&\mathrm{Tr}(A^{-1}) \\
&= \mathrm{Tr}\left(\left[U^\top T U\right]^{-1}\right) \\
&\quad - \mathrm{Tr}\left(\left[U^\top T U\right]^{-1} \begin{pmatrix} E^\top \\ 0 \end{pmatrix} \left[H^\dagger + \begin{pmatrix} E & 0 \end{pmatrix} \left[U^\top T U\right]^{-1} \begin{pmatrix} E^\top \\ 0 \end{pmatrix}\right]^{-1} \begin{pmatrix} E & 0 \end{pmatrix} \left[U^\top T U\right]^{-1}\right) \\
&= \mathrm{Tr}\left(T^{-1}\right) \\
&\quad - \mathrm{Tr}\left(U^\top T^{-1} U \begin{pmatrix} E^\top \\ 0 \end{pmatrix} \left[H^\dagger + \begin{pmatrix} E & 0 \end{pmatrix} U^\top T^{-1} U \begin{pmatrix} E^\top \\ 0 \end{pmatrix}\right]^{-1} \begin{pmatrix} E & 0 \end{pmatrix} U^\top T^{-1} U\right) \\
&\overset{(i)}{=} \mathrm{Tr}\left(T^{-1}\right) - \mathrm{Tr}\left(U^\top T^{-1} E E^\top \left[H^\dagger + E E^\top T^{-1} E E^\top\right]^{-1} E E^\top T^{-1} U\right) \\
&= \mathrm{Tr}\left(T^{-1}\right) - \mathrm{Tr}\left(T^{-1} E E^\top \left[H^\dagger + E E^\top T^{-1} E E^\top\right]^{-1} E E^\top T^{-1}\right),
\end{aligned} \tag{5}$$

where (i) follows since $(E; 0) U^\top = (E; 0)(E; F)^\top = EE^\top$. Now

$$\begin{aligned}
0 &\leq \mathrm{Tr}\left(T^{-1} E E^\top \left[H^\dagger + E E^\top T^{-1} E E^\top\right]^{-1} E E^\top T^{-1}\right) \\
&\leq \mathrm{Tr}\left(T^{-1} E E^\top \left(E E^\top T^{-1} E E^\top\right)^{-1} E E^\top T^{-1}\right) \\
&= \mathrm{Tr}\left(T^{-1} E E^\top \left(E E^\top\right)^\dagger T \left(E E^\top\right)^\dagger E E^\top T^{-1}\right),
\end{aligned} \tag{6}$$

where the second inequality holds because

$$
\begin{aligned}
& H^{\dagger} \succeq 0 \\
& \Rightarrow \left(EE^{\top}T^{-1}EE^{\top}\right)^{-1} - \left(H^{\dagger} + EE^{\top}T^{-1}EE^{\top}\right)^{-1} \succeq 0 \\
& \Rightarrow T^{-1}EE^{\top}\left(\left(EE^{\top}T^{-1}EE^{\top}\right)^{-1} - \left(H^{\dagger} + EE^{\top}T^{-1}EE^{\top}\right)^{-1}\right)EE^{\top}T^{-1} \succeq 0 \\
& \Rightarrow T^{-1}EE^{\top}\left(EE^{\top}T^{-1}EE^{\top}\right)^{-1}EE^{\top}T^{-1} \succeq T^{-1}EE^{\top}\left(\left(H^{\dagger} + EE^{\top}T^{-1}EE^{\top}\right)^{-1}\right)EE^{\top}T^{-1}
\end{aligned}
$$

along with the fact that, for any symmetric positive semi-definite matrices Q and S such that $Q \succeq S$, for all i, $\mu_i(Q) \geq \mu_i(S) \geq 0$.

Thus combining equations (5) and (6) we get that

$$
\left|\mathrm{Tr}(A^{-1}) - \mathrm{Tr}\left(T^{-1}\right)\right| \leq \left|\mathrm{Tr}\left(T^{-1}EE^{\top}\left(EE^{\top}\right)^{\dagger}T\left(EE^{\top}\right)^{\dagger}EE^{\top}T^{-1}\right)\right|.
$$

The rank of $T^{-1}EE^{\top}\left(EE^{\top}\right)^{\dagger}T\left(EE^{\top}\right)^{\dagger}EE^{\top}T^{-1}$ is at most k, so

$$
\begin{aligned}
\left|\mathrm{Tr}(A^{-1}) - \mathrm{Tr}\left(T^{-1}\right)\right| &\leq k\left\|T^{-1}EE^{\top}\left(EE^{\top}\right)^{\dagger}T\left(EE^{\top}\right)^{\dagger}EE^{\top}T^{-1}\right\|_{op} \\
&\leq k\|T^{-1}\|_{op}^2\|EE^{\top}(EE^{\top})^{\dagger}\|_{op}^2\|T\|_{op} \\
&\leq \frac{k\mu_1(T)}{\mu_n(T)^2}. \qquad (7)
\end{aligned}
$$

Next by invoking Lemma 4.4, for any $t > 2n$, with probability at least $1 - 2e^{-t}$

$$
\begin{aligned}
\mu_1(T) &\leq s_k\left[1 + c\left(\frac{t+n}{r_k} + \sqrt{\frac{t+n}{R_k}}\right)\right] \\
&\leq s_k\left[1 + 3c\left(\frac{t}{r_k} + \sqrt{\frac{t}{R_k}}\right)\right] \\
&\leq c's_k\left[1 + \left(\frac{t}{r_k} + \sqrt{\frac{t}{R_k}}\right)\right].
\end{aligned}
$$

Recall that $r_k \geq bn$ by the definition of the index k in Definition 2.3. Given a $\beta < c_0 < 1$, where c_0 is small enough, set $t = \min\{r_k, R_k\}/\beta^2 = r_k/\beta^2$ (since $r_k \leq R_k$ by Lemma 2.2) to get that

$$
\mu_1(T) \leq c's_k\left[1 + \left(\frac{1}{\beta^2} + \frac{1}{\beta}\right)\right]
$$

with probability at least $1 - 2\exp(-r_k/\beta^2)$. Next, note that $p - k \geq \sum_{j>k}\lambda_j/\lambda_{k+1} = r_k \geq bn$. Therefore, by Lemma 4.5, for any $\beta < c_0 < 1$,

$$
\mathbb{P}\left[\mu_n(T) \leq \beta s_k\right] \leq (c_1\beta)^{c_2 \cdot r_k}.
$$

Combining the last two inequalities we find that, for any $\beta < c_0 < 1$

$$
\frac{\mu_1(T)}{\mu_n(T)^2} \leq \frac{c'}{s_k}\left[\frac{1 + \left(\frac{1}{\beta^2} + \frac{1}{\beta}\right)}{\beta^2}\right] \leq \frac{c_3}{\beta^4 s_k}
$$

with probability at least $1 - 2\exp(-r_k/\beta^2) - (c_1\beta)^{c_2 \cdot r_k}$. Combined with inequality (7) this completes our proof. ■

4.2.2 $\mathbb{E}\left[\mathrm{Tr}(A^{-1})\right]$ is close to $\mathbb{E}\left[\mathrm{Tr}(T^{-1})\right]$

Next, we show that $\mathbb{E}\left[\mathrm{Tr}(A^{-1})\right]$ is close to $\mathbb{E}\left[\mathrm{Tr}(T^{-1})\right]$.

Lemma 4.9. *There exists a positive constant c_0 such that*

$$\left|\mathbb{E}\left[\mathrm{Tr}(A^{-1})\right] - \mathbb{E}\left[\mathrm{Tr}(T^{-1})\right]\right| \leq \frac{c_0 k}{s_k}.$$

Proof. Given any β define

$$\omega = \frac{ck}{\beta^4 s_k} = \frac{ck}{\beta^4 r_k \lambda_{k+1}}.$$

By Lemma 4.8 for any $\omega > \frac{c_1 k}{s_k}$, where c_1 is a large enough constant

$$\mathbb{P}\left[\left|\mathrm{Tr}(A^{-1}) - \mathrm{Tr}(T^{-1})\right| > \omega\right] \leq 2\exp\left(-c' r_k^{3/2}\sqrt{\frac{\lambda_{k+1}}{k}}\sqrt{\omega}\right) + \left(\frac{c'' k}{\omega s_k}\right)^{\frac{c_2 \cdot r_k}{4}}.$$

Thus

$$\begin{aligned}
&\left|\mathbb{E}[\mathrm{Tr}(A^{-1})] - \mathbb{E}[\mathrm{Tr}(T^{-1})]]\right| \\
&\quad \leq \mathbb{E}\left[\left|\mathrm{Tr}(A^{-1}) - \mathrm{Tr}(T^{-1})\right|\right] \\
&\quad = \int_0^\infty \mathbb{P}\left[\left|\mathrm{Tr}(A^{-1}) - \mathrm{Tr}(T^{-1})\right| > \omega\right]\, \mathrm{d}\omega \\
&\quad = \int_0^{\frac{c_1 k}{s_k}} \mathbb{P}\left[\left|\mathrm{Tr}(A^{-1}) - \mathrm{Tr}(T^{-1})\right| > \omega\right]\, \mathrm{d}\omega + \int_{\frac{c_1 k}{s_k}}^\infty \mathbb{P}\left[\left|\mathrm{Tr}(A^{-1}) - \mathrm{Tr}(T^{-1})\right| > \omega\right]\, \mathrm{d}\omega \\
&\quad \leq \frac{c_1 k}{s_k} + \underbrace{\int_{\frac{c_1 k}{s_k}}^\infty 2\exp\left(-c' r_k^{3/2}\sqrt{\frac{\lambda_{k+1}}{k}}\sqrt{\omega}\right)\, \mathrm{d}\omega}_{=:\spadesuit} + \underbrace{\int_{\frac{c_1 k}{s_k}}^\infty \left(\frac{c'' k}{\omega s_k}\right)^{\frac{c_2 \cdot r_k}{4}}\, \mathrm{d}\omega}_{=:\clubsuit}. \qquad (8)
\end{aligned}$$

First we control $\spadesuit$ as follows:

$$\begin{aligned}
\spadesuit &= \int_{\frac{c_1 k}{s_k}}^\infty 2\exp\left(-c' r_k^{3/2}\sqrt{\frac{\lambda_{k+1}}{k}}\sqrt{\omega}\right)\, \mathrm{d}\omega \\
&= 4\exp\left(-c_3 r_k\right)\frac{c_3 r_k + 1}{\left(c' r_k^{3/2}\sqrt{\frac{\lambda_{k+1}}{k}}\right)^2} \qquad \left(\text{since } \textstyle\int \exp(-\sqrt{z}) = -2e^{-\sqrt{z}}(\sqrt{z}+1) + c\right) \\
&= \frac{4k}{s_k}\left[\exp\left(-c_3 r_k\right)\frac{c_3 r_k + 1}{(c' r_k)^2}\right] \qquad (\text{since } s_k = r_k \lambda_{k+1}) \\
&\leq \frac{c_4 k}{s_k}. \qquad (9)
\end{aligned}$$

Next we control ♣ as follows

$$\clubsuit = \int_{\frac{c_1 k}{s_k}}^{\infty} \left(\frac{c'' k}{\omega s_k}\right)^{\frac{c_2 \cdot r_k}{4}} \mathrm{d}\omega = \left(\frac{c'' k}{s_k}\right)^{\frac{c_2 \cdot r_k}{4}} \int_{\frac{c_1 k}{s_k}}^{\infty} \left(\frac{1}{\omega}\right)^{\frac{c_2 \cdot r_k}{4}} \mathrm{d}\omega$$

$$= \left(\frac{c'' k}{s_k}\right)^{\frac{c_2 \cdot r_k}{4}} \times \frac{1}{\frac{c_2 \cdot r_k}{4} - 1} \times \left(\frac{s_k}{c_1 k}\right)^{\frac{c_2 \cdot r_k}{4} - 1}$$

$$= \frac{c_5 k}{s_k} \left((c'')^{\frac{c_2 \cdot r_k}{4}} \times \frac{4}{c_2 \cdot r_k - 4} \times \left(\frac{1}{c_1}\right)^{\frac{c_2 \cdot r_k}{4} - 1} \right)$$

$$\leq \frac{c_5 k}{s_k} \tag{10}$$

where the last inequality follows because the constant c_1 large enough and because $r_k \geq bn$ for a large enough constant b. Combining inequalities (8), (9) and (10) we conclude that

$$\left|\mathbb{E}\left[\mathrm{Tr}(A^{-1})\right] - \mathbb{E}\left[\mathrm{Tr}(T^{-1})\right]\right| \leq \frac{c_1 k}{s_k} + \frac{c_4 k}{s_k} + \frac{c_5 k}{s_k} \leq \frac{c_0 k}{s_k},$$

wrapping up the proof. ■

4.2.3 $\mathrm{Tr}(T^{-1})$ concentrates around its mean

Finally, we shall show that $\mathrm{Tr}(T^{-1})$ is close to its expectation $\mathbb{E}\left[\mathrm{Tr}(T^{-1})\right]$ with high probability.

Lemma 4.10. *There exists a positive constant c_0 such that with probability at least $1 - 2e^{-n}$,*

$$\left|\mathrm{Tr}(T^{-1}) - \mathbb{E}\left[\mathrm{Tr}(T^{-1})\right]\right| \leq \frac{c_0 n}{s_k} \left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k}\right].$$

Proof. We use a symmetrization argument:

$$\left|\mathrm{Tr}(T^{-1}) - \mathbb{E}\left[\mathrm{Tr}(T^{-1})\right]\right| = \left|\sum_{i=1}^{n} \frac{1}{\mu_i(T)} - \mathbb{E}\left[\frac{1}{\mu_i(T)}\right]\right| \leq \sum_{i=1}^{n} \left|\frac{1}{\mu_i(T)} - \mathbb{E}\left[\frac{1}{\mu_i(T)}\right]\right|$$

$$= \sum_{i=1}^{n} \left|\frac{1}{\mu_i(T)} - \mathbb{E}_{T'}\left[\frac{1}{\mu_i(T')}\right]\right|,$$

where in the equation above the matrices T and T' are independent and identically distributed. Thus

$$\left|\mathrm{Tr}(T^{-1}) - \mathbb{E}\left[\mathrm{Tr}(T^{-1})\right]\right| \leq \sum_{i=1}^{n} \left|\mathbb{E}_{T'}\left[\frac{1}{\mu_i(T)} - \frac{1}{\mu_i(T')}\right]\right| \leq \sum_{i=1}^{n} \mathbb{E}_{T'}\left[\left|\frac{1}{\mu_i(T)} - \frac{1}{\mu_i(T')}\right|\right]$$

$$= \sum_{i=1}^{n} \mathbb{E}_{T'}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')}\right]. \tag{11}$$

By Lemma 4.4, with probability at least $1 - 2e^{-n}$, for all $i \in [n]$,

$$s_k \left[1 - c_1 \left(\frac{n}{r_k} + \sqrt{\frac{n}{R_k}}\right)\right] \leq \mu_i(T) \leq s_k \left[1 + c_1 \left(\frac{n}{r_k} + \sqrt{\frac{n}{R_k}}\right)\right]. \tag{12}$$

We will assume that the event described above, which controls the singular values of T, occurs going forward. (This determines the success probability in the statement of the lemma.) The game plan now is to evaluate the expectation with respect to T' in equation (11) by integrating tail bounds. Since (12) holds,

$$
\begin{aligned}
&\left|\mu_i(T) - \mu_i(T')\right| \\
&= \max\{\mu_i(T) - \mu_i(T'), \mu_i(T') - \mu_i(T)\} \\
&\le \max\left\{ s_k \left[1 + c_1 \left(\frac{n}{r_k} + \sqrt{\frac{n}{R_k}}\right)\right] - \mu_i(T'), \right. \\
&\qquad\qquad \left. \mu_i(T') - s_k \left[1 - c_1 \left(\frac{n}{r_k} + \sqrt{\frac{n}{R_k}}\right)\right] \right\} \\
&\le \max\left\{ s_k \left[1 + c_1 \left(\frac{n}{r_k} + \sqrt{\frac{n}{R_k}}\right)\right] - s_k \left[1 - c_2 \left(\frac{t+n}{r_k} + \sqrt{\frac{t+n}{R_k}}\right)\right], \right. \\
&\qquad \left. s_k \left[1 + c_2 \left(\frac{t+n}{r_k} + \sqrt{\frac{t+n}{R_k}}\right)\right] - s_k \left[1 - c_1 \left(\frac{n}{r_k} + \sqrt{\frac{n}{R_k}}\right)\right] \right\} \\
&\qquad\qquad\qquad \text{(by Lemma 4.4)} \\
&\le c_3 s_k \left(\frac{t+n}{r_k} + \sqrt{\frac{t+n}{R_k}}\right), \qquad (13)
\end{aligned}
$$

with probability $1 - 2e^{-t}$.

Next, by Lemma 4.5, we know that for all $\beta < c_4 < 1$

$$\mathbb{P}\left[\mu_n(T') \le \beta s_k\right] \le (c_5 \beta)^{c_6 \cdot r_k}. \qquad (14)$$

Combining equations (13) and (14), and because condition (12) holds, we get that

$$\mathbb{P}\left[\exists\, i \in [n] \;:\; \frac{|\mu_i(T) - \mu_i(T')|}{\mu_i(T)\mu_i(T')} \ge \frac{c_3\left(\frac{t+n}{r_k} + \sqrt{\frac{t+n}{R_k}}\right)}{\beta s_k \left[1 - c_1\left(\frac{n}{r_k} + \sqrt{\frac{n}{R_k}}\right)\right]}\right] \le 2e^{-t} + (c_5\beta)^{c_6 \cdot r_k}.$$

Now since $r_k \ge bn$ for a large enough constant b by the definition of k, and since $R_k > r_k$ by Lemma 2.2, we can simplify the denominator in the equation above to get that

$$\mathbb{P}\left[\exists\, i \in [n] \;:\; \frac{|\mu_i(T) - \mu_i(T')|}{\mu_i(T)\mu_i(T')} \ge \frac{c_7\left(\frac{t+n}{r_k} + \sqrt{\frac{t+n}{R_k}}\right)}{\beta s_k}\right] \le 2e^{-t} + (c_5\beta)^{c_6 \cdot r_k}.$$

Setting $t = n/\beta$ yields

$$\mathbb{P}\left[\exists\, i \in [n] \;:\; \frac{|\mu_i(T) - \mu_i(T')|}{\mu_i(T)\mu_i(T')} \ge \frac{c_7\left(\frac{n(\beta+1)}{\beta r_k} + \sqrt{\frac{n(\beta+1)}{\beta R_k}}\right)}{\beta s_k}\right] \le 2e^{-n/\beta} + (c_5\beta)^{c_6 \cdot r_k}.$$

Now since $\beta < c_4 < 1$, we find that

$$
\begin{aligned}
\mathbb{P}\left[\exists\, i \in [n] \;:\; \frac{|\mu_i(T) - \mu_i(T')|}{\mu_i(T)\mu_i(T')} \ge \frac{c_8}{s_k} \max\left\{\frac{n}{\beta^2 r_k}, \frac{\sqrt{n}}{\beta^{3/2}\sqrt{R_k}}\right\}\right]& \\
\le 2e^{-n/\beta} + (c_5\beta)^{c_6 \cdot r_k}.& \qquad (15)
\end{aligned}
$$

For every β define

$$\omega := \frac{c_8}{s_k} \max\left\{ \frac{n}{\beta^2 r_k}, \frac{\sqrt{n}}{\beta^{3/2}\sqrt{R_k}} \right\}.$$

Inverting the map from β to ω yields

$$\beta(\omega) = \begin{cases} \left(\frac{c_8\sqrt{n}}{\omega\sqrt{R_k}s_k}\right)^{2/3} & \text{if } \omega \le \omega_\tau := \frac{c_8}{s_k}\left(\frac{r_k^3}{R_k^2 n}\right), \\ \sqrt{\frac{c_8 n}{\omega r_k s_k}} & \text{otherwise.} \end{cases} \tag{16}$$

Let ω_0 be such that $\beta(\omega_0) = c_4$, and define

$$\omega_- := \min\{\omega_0, \omega_\tau\} \quad \text{and} \quad \omega_+ := \max\{\omega_0, \omega_\tau\}.$$

Applying inequality (15) we have that, for all $\omega \in (\omega_-, \omega_\tau]$

$$\begin{aligned} &\mathbb{P}\left[\exists\, i \in [n] \ :\ \frac{|\mu_i(T) - \mu_i(T')|}{\mu_i(T)\mu_i(T')} \ge \omega\right] \\ &\quad \le 2\exp\left(-c_9\left(\omega n\sqrt{R_k}s_k\right)^{2/3}\right) + \left(\frac{c_{10}\sqrt{n}}{\omega\sqrt{R_k}s_k}\right)^{c_{11}\cdot r_k}, \end{aligned} \tag{17}$$

and for $\omega > \omega_+$, we have

$$\mathbb{P}\left[\exists\, i \in [n] \ :\ \frac{|\mu_i(T) - \mu_i(T')|}{\mu_i(T)\mu_i(T')} \ge \omega\right] \le 2\exp\left(-c_{12}\left(\omega n r_k s_k\right)^{1/2}\right) + \left(\frac{c_{13} n}{\omega r_k s_k}\right)^{c_{14}\cdot r_k}. \tag{18}$$

Thus

$$\begin{aligned} \mathbb{E}_{T'}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')}\right] &= \int_0^\infty \mathbb{P}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')} \ge \omega\right] \mathrm{d}\omega \\ &= \int_0^{\omega_0} \mathbb{P}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')} \ge \omega\right] \mathrm{d}\omega \\ &\quad + \int_{\omega_-}^{\omega_\tau} \mathbb{P}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')} \ge \omega\right] \mathrm{d}\omega \\ &\quad + \int_{\omega_+}^{\infty} \mathbb{P}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')} \ge \omega\right] \mathrm{d}\omega \\ &\le \omega_0 + \underbrace{\int_{\omega_-}^{\omega_\tau} \mathbb{P}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')} \ge \omega\right] \mathrm{d}\omega}_{=:\spadesuit} \\ &\quad + \underbrace{\int_{\omega_+}^{\infty} \mathbb{P}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')} \ge \omega\right] \mathrm{d}\omega}_{=:\clubsuit}. \end{aligned} \tag{19}$$

Let us perform each of these two integrals $\spadesuit$ and $\clubsuit$ separately.

First,

$$
\begin{aligned}
&\spadesuit \\
&= \int_{\omega_-}^{\omega_\tau} \mathbb{P}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')} \geq \omega\right] \, \mathrm{d}\omega \\
&\leq \int_{\omega_-}^{\omega_\tau} \left[2\exp\left(-c_9\left(\omega n\sqrt{R_k}s_k\right)^{2/3}\right) + \left(\frac{c_{10}\sqrt{n}}{\omega\sqrt{R_k}s_k}\right)^{c_{11}\cdot r_k}\right] \mathrm{d}\omega \qquad \text{(by inequality (17))} \\
&\leq \mathbb{I}[\omega_- < \omega_\tau] \int_{\omega_-}^{\infty} \left[2\exp\left(-c_9\left(\omega n\sqrt{R_k}s_k\right)^{2/3}\right) + \left(\frac{c_{10}\sqrt{n}}{\omega\sqrt{R_k}s_k}\right)^{c_{11}\cdot r_k}\right] \mathrm{d}\omega.
\end{aligned}
$$

Now, for $\zeta := c_9\left(n\sqrt{R_k}s_k\right)^{2/3}$, we have that

$$
\begin{aligned}
&2\int_{\omega_-}^{\infty} \exp\left(-c_9\left(\omega n\sqrt{R_k}s_k\right)^{2/3}\right) \, \mathrm{d}\omega \\
&= 2\int_{\omega_-}^{\infty} \exp\left(-\zeta\omega^{2/3}\right) \, \mathrm{d}\omega \\
&= \frac{3\omega_-^{1/3}\exp(-\zeta\omega_-^{2/3})}{\zeta} + \frac{3\sqrt{\pi}\left(1 - \mathsf{erf}\left(\sqrt{\zeta}\omega_-^{1/3}\right)\right)}{2\zeta^{3/2}} \\
&\qquad\qquad\qquad\qquad \left(\text{since } \textstyle\int \exp(-z^{2/3}) = \frac{3}{4}\left(\sqrt{\pi}\mathsf{erf}(z^{1/3}) - 2e^{-z^{2/3}}z^{1/3}\right) + c\right) \\
&\leq \frac{3\omega_-^{1/3}\exp(-\zeta\omega_-^{2/3})}{\zeta} + \frac{3\exp(-\zeta\omega_-^{2/3})}{2\zeta^2\omega_-^{1/3}} \\
&= \frac{c_{15}\exp\left(-c_9\left(n\sqrt{R_k}s_k\right)^{2/3}\omega_-^{2/3}\right)}{(n\sqrt{R_k}s_k)^{2/3}} \left(\omega_-^{1/3} + \frac{1}{\left(n\sqrt{R_k}s_k\right)^{2/3}\omega_-^{1/3}}\right). \qquad (20)
\end{aligned}
$$

Continuing our work of bounding $\spadesuit$, we have that

$$
\begin{aligned}
\int_{\omega_-}^{\infty} \left(\frac{c_{10}\sqrt{n}}{\omega\sqrt{R_k}s_k}\right)^{c_{11}\cdot r_k} \mathrm{d}\omega &= \left(\frac{c_{10}\sqrt{n}}{\sqrt{R_k}s_k}\right)^{c_{11}\cdot r_k} \int_{\omega_-}^{\infty} \left(\frac{1}{\omega}\right)^{c_{11}\cdot r_k} \mathrm{d}\omega \\
&= \left(\frac{c_{10}\sqrt{n}}{\sqrt{R_k}s_k}\right)^{c_{11}\cdot r_k} \times \frac{1}{c_{11}\cdot r_k - 1}\left(\frac{1}{\omega_-}\right)^{c_{11}\cdot r_k - 1} \\
&\leq c_{16}\left(\frac{c_{10}\sqrt{n}}{\sqrt{R_k}s_k}\right)^{c_{11}\cdot r_k}\left(\frac{1}{\omega_-}\right)^{c_{11}\cdot r_k - 1}, \qquad (21)
\end{aligned}
$$

where the last inequality follows since $r_k \geq bn$ for a large enough constant b. By combining inequalities (20) and (21) we get the following bound on the integral $\spadesuit$:

$$
\begin{aligned}
\spadesuit \leq \mathbb{I}[\omega_- < \omega_\tau] &\frac{c_{15}\exp\left(-c_9\left(n\sqrt{R_k}s_k\right)^{2/3}\omega_-^{2/3}\right)}{(n\sqrt{R_k}s_k)^{2/3}} \left(\omega_-^{1/3} + \frac{1}{\left(n\sqrt{R_k}s_k\right)^{2/3}\omega_-^{1/3}}\right) \\
&+ \mathbb{I}[\omega_- < \omega_\tau]c_{16}\left(\frac{c_{10}\sqrt{n}}{\sqrt{R_k}s_k}\right)^{c_{11}\cdot r_k}\left(\frac{1}{\omega_-}\right)^{c_{11}\cdot r_k - 1}. \qquad (22)
\end{aligned}
$$

Let us now bound ♣

$$\clubsuit = \int_{\omega_+}^{\infty} \mathbb{P}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')} \geq \omega\right] \mathrm{d}\omega$$
$$\leq \int_{\omega_+}^{\infty} \left[2\exp\left(-c_{12}\left(\omega n r_k s_k\right)^{1/2}\right) + \left(\frac{c_{13} n}{\omega r_k s_k}\right)^{c_{14}\cdot r_k}\right] \mathrm{d}\omega \qquad \text{(applying inequality (18)).}$$

For $\zeta' := c_{12}\left(n r_k s_k\right)^{1/2}$, we have

$$2\int_{\omega_+}^{\infty} \exp\left(-c_{12}\left(\omega n r_k s_k\right)^{1/2}\right) \mathrm{d}\omega$$
$$= 2\int_{\omega_+}^{\infty} \exp\left(-\zeta'\omega^{1/2}\right) \mathrm{d}\omega$$
$$= \frac{4\exp(-\zeta'\sqrt{\omega_+})(\zeta'\sqrt{\omega_+}+1)}{\zeta'^2} \qquad \left(\text{since } \textstyle\int \exp(-\sqrt{z}) = -2e^{-\sqrt{z}}(\sqrt{z}+1) + c\right)$$
$$= \frac{c_{17}\exp\left(-c_{12}\left(n r_k s_k \omega_+\right)^{1/2}\right)\left[c_{12}\left(n r_k s_k \omega_+\right)^{1/2} + 1\right]}{n r_k s_k}. \tag{23}$$

We continue to bound the other integral in ♣ as follows

$$\int_{\omega_+}^{\infty} \left(\frac{c_{13} n}{\omega r_k s_k}\right)^{c_{14}\cdot r_k} \mathrm{d}\omega \leq c_{18}\left(\frac{c_{13} n}{r_k s_k}\right)^{c_{14}\cdot r_k}\left(\frac{1}{\omega_+}\right)^{c_{14}\cdot r_k - 1}, \tag{24}$$

where the bound follows by mirroring the logic used to arrive at inequality (21) above. Therefore, combining inequalities (23) and (24) we get that

$$\clubsuit \leq \frac{c_{17}\exp\left(-c_{12}\left(n r_k s_k \omega_+\right)^{1/2}\right)\left[c_{12}\left(n r_k s_k \omega_+\right)^{1/2} + 1\right]}{n r_k s_k} + c_{18}\left(\frac{c_{13} n}{r_k s_k}\right)^{c_{14}\cdot r_k}\left(\frac{1}{\omega_+}\right)^{c_{14}\cdot r_k - 1}. \tag{25}$$

Having controlled both ♠ and ♣ in (22) and (25) respectively, by using the decomposition in (19) we find that

$$\mathbb{E}_{T'}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')}\right]$$
$$\leq \omega_0 + \mathbb{I}[\omega_- < \omega_\tau]\frac{c_{15}\exp\left(-c_9\left(n\sqrt{R_k}s_k\right)^{2/3}\omega_-^{2/3}\right)}{(n\sqrt{R_k}s_k)^{2/3}}\left(\omega_-^{1/3} + \frac{1}{\left(n\sqrt{R_k}s_k\right)^{2/3}\omega_-^{1/3}}\right)$$
$$+ \mathbb{I}[\omega_- < \omega_\tau] c_{16}\left(\frac{c_{10}\sqrt{n}}{\sqrt{R_k}s_k}\right)^{c_{11}\cdot r_k}\left(\frac{1}{\omega_-}\right)^{c_{11}\cdot r_k - 1}$$
$$+ \frac{c_{17}\exp\left(-c_{12}\left(n r_k s_k \omega_+\right)^{1/2}\right)\left[c_{12}\left(n r_k s_k \omega_+\right)^{1/2} + 1\right]}{n r_k s_k}$$
$$+ c_{18}\left(\frac{c_{13} n}{r_k s_k}\right)^{c_{14}\cdot r_k}\left(\frac{1}{\omega_+}\right)^{c_{14}\cdot r_k - 1}. \tag{26}$$

We now consider two cases.

Case 1: $(\omega_0 < \omega_\tau)$. In this case, using the fact that $\beta(\omega_0) = c_4$ and the formula for β in equation (16) we get that

$$\omega_0 = \frac{c_8\sqrt{n}}{c_4^{3/2}\sqrt{R_k}s_k} = \frac{c_{19}\sqrt{n}}{\sqrt{R_k}s_k},$$

and that

$$\omega_- = \min\{\omega_0, \omega_\tau\} = \omega_0 = \frac{c_8\sqrt{n}}{c_4^{3/2}\sqrt{R_k}s_k},$$
$$\omega_+ = \max\{\omega_0, \omega_\tau\} = \omega_\tau = \frac{c_8 r_k^3}{R_k^2 n s_k}.$$

Also note that in this case since,

$$\omega_0 = \frac{c_8\sqrt{n}}{c_4^{3/2}\sqrt{R_k}s_k} < \frac{c_8 r_k^3}{R_k^2 n s_k} = \omega_\tau \quad \Rightarrow R_k \leq \frac{c_4 r_k^2}{n}$$

and so $\omega_+ \geq \frac{c_8 n}{c_4^2 r_k s_k}$.

Thus, substituting the above values of ω_0, ω_- in inequality (26), and, because the RHS of this inequality is a decreasing function in ω_+ (since the function $z \mapsto \exp(-z)(z+1)$ is a decreasing function for all positive z), replacing ω_+ with the above lower bound, we find that

$$\begin{aligned}
&\mathbb{E}_{T'}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')}\right] \\
&\quad \leq \frac{c_{19}\sqrt{n}}{\sqrt{R_k}s_k} + \frac{c_{20}\exp(-c_{21}n)}{\sqrt{nR_k}s_k} + \frac{c_{20}\exp(-c_{21}n)}{n^{3/2}\sqrt{R_k}s_k} \\
&\qquad\qquad + \frac{c_{16}c_8\sqrt{n}}{c_4^{3/2}\sqrt{R_k}s_k}\left(\frac{c_{10}c_4^{3/2}}{c_8}\right)^{c_{11}\cdot r_k} + \frac{c_{22}\exp(-c_{23}n)}{r_k s_k} + \frac{c_{18}c_8 n}{c_4^2 r_k s_k}\left(\frac{c_{13}c_4^2}{c_8}\right)^{c_{14}\cdot r_k} \\
&\quad \overset{(i)}{\leq} \frac{c_{19}\sqrt{n}}{\sqrt{R_k}s_k} + \frac{c_{20}\exp(-c_{21}n)}{\sqrt{nR_k}s_k} + \frac{c_{20}\exp(-c_{21}n)}{n^{3/2}\sqrt{R_k}s_k} + \frac{c_{16}c_8\sqrt{n}}{c_4^{3/2}\sqrt{R_k}s_k} + \frac{c_{22}\exp(-c_{23}n)}{r_k s_k} + \frac{c_{18}c_8 n}{c_4^2 r_k s_k} \\
&\quad \leq \frac{c_{24}\sqrt{n}}{\sqrt{R_k}s_k} + \frac{c_{25}n}{r_k s_k},
\end{aligned}$$

where (i) follows since c_4 is small enough. This combined with inequalities (11) and (12) proves the lemma in this case.

Case 2: $(\omega_0 \geq \omega_\tau)$. In this case, using the fact that $\beta(\omega_0) = c_4$ and the formula for β in equation (16) we get that

$$\omega_0 = \frac{c_8 n}{c_4^2 r_k s_k}$$

and that

$$\omega_- = \min\{\omega_0, \omega_\tau\} = \omega_\tau,$$
$$\omega_+ = \max\{\omega_0, \omega_\tau\} = \omega_0 = \frac{c_8 n}{c_4^2 r_k s_k}.$$

Now by applying inequality (26) we get that

$$\mathbb{E}_{T'}\left[\frac{|\mu_i(T') - \mu_i(T)|}{\mu_i(T)\mu_i(T')}\right] \leq \frac{c_8 n}{c_4^2 r_k s_k} + \frac{c_{26}\exp(-c_{27}n)}{r_k s_k} + \frac{c_{28}n}{r_k s_k}\left(\frac{c_{13}c_4^2}{c_8}\right)^{c_{14}\cdot r_k}$$
$$\stackrel{(i)}{\leq} \frac{c_8 n}{c_4^2 r_k s_k} + \frac{c_{26}\exp(-c_{27}n)}{r_k s_k} + \frac{c_{28}n}{r_k s_k} \leq \frac{c_{29}n}{r_k s_k},$$

where (i) follows since c_4 is small enough. Again, combining this inequality with inequalities (11) and (12) proves the lemma in this second case. ■

4.2.4 Bounds on $\mathbb{E}\left[\mathrm{Tr}(A^{-1})\right]$

To characterize $\mathrm{Tr}\left(A^{-1}\right)$ in terms of relevant problem parameters we will need to establish upper and lower bounds that are tight up to the leading constant on its expectation.

Lemma 4.11. *There are positive constants c_0 and c_1 such that*

$$\left|\mathbb{E}\left[\mathrm{Tr}\left(A^{-1}\right)\right] - \frac{n}{s_k}\right| \leq \frac{c_0 n}{s_k}\left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \frac{k}{n} + e^{-c_1\sqrt{n}}\right].$$

Proof. By Lemma 4.9 we know that

$$\mathbb{E}\left[\mathrm{Tr}(T^{-1})\right] - \frac{ck}{s_k} \leq \mathbb{E}\left[\mathrm{Tr}(A^{-1})\right] \leq \mathbb{E}\left[\mathrm{Tr}(T^{-1})\right] + \frac{ck}{s_k}. \tag{27}$$

Thus, we shall instead upper and lower bound $\mathbb{E}\left[\mathrm{Tr}(T^{-1})\right]$.

The lower bound: By definition

$$\mathbb{E}\left[\mathrm{Tr}(T^{-1})\right] = \mathbb{E}\left[\sum_{i=1}^{n}\frac{1}{\mu_i(T)}\right] \geq \mathbb{E}\left[\frac{n}{\frac{1}{n}\sum_{i=1}^{n}\mu_i(T)}\right] \qquad \text{(by the AM-HM inequality).} \tag{28}$$

By Bernstein's inequality (see Theorem B.5) we know that with probability at least $1 - 2e^{-t}$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mu_i(T) &= \frac{1}{n}\mathrm{Tr}(T) \\
&= \frac{1}{n}\sum_{i>k}\lambda_i \mathrm{Tr}(z_i z_i^\top) \\
&= \frac{1}{n}\sum_{i>k}\lambda_i \|z_i\|^2 \\
&\leq \sum_{i>k}\lambda_i + c_2\max\left\{t\lambda_{k+1}, \sqrt{t\sum_{i>k}\lambda_i^2}\right\} \\
&= s_k\left[1 + c_2\max\left\{\frac{t}{r_k}, \sqrt{\frac{t}{R_k}}\right\}\right] \qquad \text{(since } s_k = \textstyle\sum_{j>k}\lambda_j\text{)} \\
&\leq s_k\left[1 + c_2\max\left\{\frac{t}{\sqrt{R_k}}, \sqrt{\frac{t}{R_k}}\right\}\right],
\end{aligned}$$

since $r_k \geq \sqrt{R_k}$ by Lemma 2.2. Setting $t = \sqrt{n}$ implies that

$$\frac{1}{n}\sum_{i=1}^{n} \mu_i(T) \leq s_k \left[1 + 2c_2\sqrt{\frac{n}{R_k}}\right]$$

with probability at least $1 - 2e^{-\sqrt{n}}$. Thus by inequality (28)

$$\begin{aligned}
\mathbb{E}\left[\operatorname{Tr}(T^{-1})\right] &\geq \frac{n}{s_k\left(1 + 2c_2\sqrt{\frac{n}{R_k}}\right)} \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \mu_i(T) \leq s_k \left[1 + 2c_2\sqrt{\frac{n}{R_k}}\right]\right] \\
&\geq \frac{n}{s_k}\left[\frac{1 - 2e^{-\sqrt{n}}}{1 + 2c_2\sqrt{\frac{n}{R_k}}}\right]. \qquad (29)
\end{aligned}$$

Combined with the lower bound in inequality (27) we find that

$$\begin{aligned}
\mathbb{E}\left[\operatorname{Tr}(A^{-1})\right] &\geq \frac{n}{s_k}\left[\frac{1 - 2e^{-\sqrt{n}}}{1 + 2c_2\sqrt{\frac{n}{R_k}}}\right] - \frac{c_1 k}{s_k} \\
&\geq \frac{n}{s_k}\left[1 - \frac{2c_2\sqrt{\frac{n}{R_k}} + 2e^{-\sqrt{n}}}{1 + 2c_2\sqrt{\frac{n}{R_k}}} - \frac{c_1 k}{n}\right] \\
&\geq \frac{n}{s_k}\left[1 - c_0\left(\sqrt{\frac{n}{R_k}} + \frac{k}{n} + e^{-\sqrt{n}}\right)\right] \qquad (\text{since } R_k \geq r_k \geq bn). \qquad (30)
\end{aligned}$$

This proves the desired lower bound.

The upper bound: To obtain the upper bound we shall bound

$$\mathbb{E}\left[\operatorname{Tr}(T^{-1})\right] = \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{\mu_i(T)}\right] \leq n\mathbb{E}\left[\frac{1}{\mu_n(T)}\right]. \qquad (31)$$

We will upper bound the expected value of $1/\mu_n(T)$ again by integrating tail bounds. We have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\mu_n(T)}\right] &= \int_0^\infty \mathbb{P}\left[\frac{1}{\mu_n(T)} \geq \omega\right] \mathrm{d}\omega \\
&= \int_0^\infty \mathbb{P}\left[\mu_n(T) \leq \frac{1}{\omega}\right] \mathrm{d}\omega \\
&= \underbrace{\int_0^{\frac{1}{s_k\left[1 - c_3\left(\frac{n+\eta}{r_k} + \sqrt{\frac{n+\eta}{R_k}}\right)\right]}} \mathbb{P}\left[\mu_n(T) \leq \frac{1}{\omega}\right] \mathrm{d}\omega}_{=:\clubsuit} \\
&\quad + \underbrace{\int_{\frac{1}{s_k\left[1 - c_3\left(\frac{n+\eta}{r_k} + \sqrt{\frac{n+\eta}{R_k}}\right)\right]}}^{\frac{1}{c_4 s_k}} \mathbb{P}\left[\mu_n(T) \leq \frac{1}{\omega}\right] \mathrm{d}\omega}_{=:\spadesuit} \\
&\quad + \underbrace{\int_{\frac{1}{c_4 s_k}}^{\infty} \mathbb{P}\left[\mu_n(T) \leq \frac{1}{\omega}\right] \mathrm{d}\omega}_{=:\blacklozenge}, \qquad (32)
\end{aligned}$$

where

- c_3 is the constant c from Lemma 4.4,
- c_4 is smaller than the constant c_1^2 in Lemma 4.5, and
- η is small enough such that it satisfies $c_4 \leq 1 - c_3 \left(\frac{n+\eta}{r_k} + \sqrt{\frac{n+\eta}{R_k}}\right)$.

Below we will set η to scale linearly with n, thus, this condition will be satisfied since $R_k \geq r_k \geq bn$ for a large enough value of b.

The first term ♣ is positive because η scales linearly with n and $R_k \geq r_k \geq bn$ for suitably large b, and so it can be bounded as follows:

$$♣ \leq \frac{1}{s_k \left[1 - c_3 \left(\frac{n+\eta}{r_k} + \sqrt{\frac{n+\eta}{R_k}}\right)\right]}. \tag{33}$$

Next, consider the term ♠. Here we will use the additive concentration inequality (Lemma 4.4). By Lemma 4.4 we know that with probability at most $2e^{-t}$

$$\begin{aligned} \mu_n(T) &\leq s_k \left[1 - c_3 \left(\frac{t+n}{r_k} + \sqrt{\frac{t+n}{R_k}}\right)\right] \\ &\leq s_k \left[1 - c_3 \left(\frac{t+n}{r_k} + \sqrt{\frac{t+n}{r_k}}\right)\right] \qquad \text{(since } r_k \leq R_k \text{ by Lemma 2.2)} \\ &\leq s_k \left[1 - 2c_3 \max\left\{\frac{t+n}{r_k}, \sqrt{\frac{t+n}{r_k}}\right\}\right]. \end{aligned} \tag{34}$$

Also, the integral term ♠ is positive, because c_4 is chosen to be small enough, η scales linearly with n, and $R_k \geq r_k \geq bn$ for suitably large b. Thus,

$$\begin{aligned} ♠ &= \int_{\frac{1}{s_k\left[1-c_3\left(\frac{n+\eta}{r_k}+\sqrt{\frac{n+\eta}{R_k}}\right)\right]}}^{\frac{1}{c_4 s_k}} \mathbb{P}\left[\mu_n(T) \leq \frac{1}{\omega}\right] \, \mathrm{d}\omega \\ &\leq \int_{\frac{1}{s_k\left[1-c_5\sqrt{\frac{n+\eta}{r_k}}\right]}}^{\frac{1}{c_4 s_k}} \mathbb{P}\left[\mu_n(T) \leq \frac{1}{\omega}\right] \, \mathrm{d}\omega \\ &\leq 2 \int_{\frac{1}{s_k\left[1-c_5\sqrt{\frac{n+\eta}{r_k}}\right]}}^{\frac{1}{c_4 s_k}} \exp\left[-r_k \min\left\{\frac{\left(1-\frac{1}{\omega s_k}\right)}{2c_3}, \frac{\left(1-\frac{1}{\omega s_k}\right)^2}{4c_3^2}\right\} + n\right] \, \mathrm{d}\omega \\ &\qquad \text{(applying inequality (34), and by setting } 1/\omega \text{ equal to the RHS of (34) and solving for } t) \\ &\leq 2e^n \int_{\frac{1}{s_k\left[1-c_5\sqrt{\frac{n+\eta}{r_k}}\right]}}^{\frac{1}{c_4 s_k}} \exp\left[-c_6 r_k \left(1-\frac{1}{\omega s_k}\right)^2\right] \, \mathrm{d}\omega, \end{aligned}$$

where the last inequality follows since $\omega > 1/s_k$ and therefore the term in the round bracket is always smaller than 1. Thus, we get that

$$♠ \leq 2e^n \int_{\frac{1}{s_k\left[1-c_5\sqrt{\frac{n+\eta}{r_k}}\right]}}^{\frac{1}{c_4 s_k}} \exp\left[-c_6 r_k \left(1-\frac{1}{\omega s_k}\right)^2\right] \, \mathrm{d}\omega.$$

Now we set $\eta = c_7 n$, for a large enough constant c_7, and perform a change of variables, redefining $1 - \frac{1}{\omega s_k} \to \bar{\omega}$, to get

$$\begin{aligned}
\spadesuit &\leq \frac{2e^n}{s_k} \int_{c_5\sqrt{\frac{(c_7+1)n}{r_k}}}^{1-c_4} \frac{\exp\left(-c_6 r_k \bar{\omega}^2\right)}{(1-\bar{\omega})^2} \, \mathrm{d}\bar{\omega} \\
&\leq \frac{2\exp\left(-c_8 n\right)}{s_k} \int_{c_5\sqrt{\frac{(c_7+1)n}{r_k}}}^{1-c_4} \frac{1}{(1-\bar{\omega})^2} \, \mathrm{d}\omega \\
&= \frac{2\exp\left(-c_8 n\right)}{s_k} \left[\frac{1}{1 - c_5\sqrt{\frac{(c_7+1)n}{r_k}}} - \frac{1}{c_4} \right] \\
&\overset{(i)}{\leq} \frac{c_9 \exp\left(-c_8 n\right)}{s_k},
\end{aligned} \tag{35}$$

where (i) holds because $r_k \geq bn$ for a large value of b.

Finally, we turn our attention to the term ♦. By using Lemma 4.5 we know that

$$\begin{aligned}
\blacklozenge &= \int_{\frac{1}{c_4 s_k}}^{\infty} \mathbb{P}\left[\mu_n(T) \leq \frac{1}{\omega}\right] \, \mathrm{d}\omega \\
&\leq \int_{\frac{1}{c_4 s_k}}^{\infty} \left(\frac{c_{10}}{\omega s_k}\right)^{c_{11} r_k} \, \mathrm{d}\omega \\
&= \frac{1}{c_4 s_k (c_{11} r_k - 1)} \left(c_4 c_{10}\right)^{c_{11} r_k} \leq \frac{c_{12}}{r_k s_k},
\end{aligned} \tag{36}$$

where the last inequality follows since $r_k \geq bn$ and because c_4 is chosen to be small enough.

By combining inequalities (32), (33), (35) and (36) we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\mu_n(T)}\right] \\
&\leq \frac{1}{s_k\left[1 - c_3\left(\frac{n+\eta}{r_k} + \sqrt{\frac{n+\eta}{R_k}}\right)\right]} + \frac{c_9 \exp\left(-c_8 n\right)}{s_k} + \frac{c_{12}}{r_k s_k} \\
&\leq \frac{1}{s_k\left[1 - c_{13}\left(\frac{n}{r_k} + \sqrt{\frac{n}{R_k}}\right)\right]} + \frac{c_9 \exp\left(-c_8 n\right)}{s_k} + \frac{c_{12}}{r_k s_k} \qquad (\text{since } \eta = c_7 n) \\
&= \frac{1}{s_k}\left[1 + c_{14}\left(\frac{\sqrt{\frac{n}{R_k}} + \frac{n}{r_k}}{1 - c_{13}\left(\sqrt{\frac{n}{R_k}} + \frac{n}{r_k}\right)} + \exp(-c_8 n) + \frac{1}{r_k}\right)\right] \\
&\leq \frac{1}{s_k}\left[1 + c_{15}\left(\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \exp(-c_8 n)\right)\right],
\end{aligned}$$

where the last inequality follows since $R_k \geq r_k \geq bn$ with b being large enough. Hence by inequality (31)

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Tr}(T^{-1})\right] &\leq \frac{n}{s_k}\left[1 + c_{15}\left(\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \exp(-c_8 n)\right)\right] \\
&\leq \frac{n}{s_k}\left[1 + c_{15}\left(\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \exp(-c_8 \sqrt{n})\right)\right],
\end{aligned}$$

which combined with inequality (27) completes our proof. ■

4.2.5 Proof of Lemma 4.6

As mentioned previously, by using the previous four lemmas we will now show that the trace of A^{-1} is close to n/s_k with high probability. Recall the statement of the lemma from above.

Lemma 4.6. *There are positive constants $c_0, \ldots, c_4$ such that, if $p \geq c_0(n+k)$ then with probability at least $1 - c_1 e^{-c_2 n}$*

$$\left| \mathrm{Tr}\left((XX^\top)^{-1} \right) - \frac{n}{s_k} \right| \leq \frac{c_3 n}{s_k} \left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \frac{k}{n} + e^{-c_4\sqrt{n}} \right].$$

Proof. Recall that by definition $A = XX^\top$. By an application of the triangle inequality,

$$\begin{aligned} \left| \mathrm{Tr}(A^{-1}) - \frac{n}{s_k} \right| &\leq \left| \mathrm{Tr}(A^{-1}) - \mathrm{Tr}(T^{-1}) \right| + \left| \mathbb{E}\left[\mathrm{Tr}(A^{-1}) \right] - \mathbb{E}\left[\mathrm{Tr}(T^{-1}) \right] \right| \\ &\quad + \left| \mathrm{Tr}(T^{-1}) - \mathbb{E}\left[\mathrm{Tr}(T^{-1}) \right] \right| + \left| \mathbb{E}\left[\mathrm{Tr}(A^{-1}) \right] - \frac{n}{s_k} \right|. \end{aligned} \tag{37}$$

By Lemma 4.8 we know that

$$\left| \mathrm{Tr}(A^{-1}) - \mathrm{Tr}(T^{-1}) \right| \leq \frac{c_5 k}{s_k} \tag{38}$$

with probability at least

$$1 - 2\exp(-c_6 r_k) - (c_7)^{c_8 r_k} \geq 1 - c_9 \exp(-c_{10} r_k) \geq 1 - c_9 \exp(-c_{11} n),$$

where the last two inequalities follow since $r_k \geq bn$ for some large enough value of b. Next, by Lemma 4.9 we know that

$$\left| \mathbb{E}\left[\mathrm{Tr}(A^{-1}) \right] - \mathbb{E}\left[\mathrm{Tr}(T^{-1}) \right] \right| \leq \frac{c_{12} k}{s_k}. \tag{39}$$

By Lemma 4.10 we get that with probability at least $1 - 2e^{-n}$,

$$\left| \mathrm{Tr}(T^{-1}) - \mathbb{E}\left[\mathrm{Tr}(T^{-1}) \right] \right| \leq \frac{c_{13} n}{s_k} \left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} \right]. \tag{40}$$

Finally, by Lemma 4.11 we know that

$$\left| \mathbb{E}\left[\mathrm{Tr}(A^{-1}) \right] - \frac{n}{s_k} \right| \leq \frac{c_{14} n}{s_k} \left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \frac{k}{n} + e^{-c_{15}\sqrt{n}} \right]. \tag{41}$$

Combining the (37)-(41) establishes our claim. ■

4.3 Proof of Lemma 4.3

Armed with Lemmas 4.4, 4.5 and 4.6, we are ready to prove Lemma 4.3 and establish upper and lower bounds on $\alpha^\star$. This proof is further divided into a series of lemmas.

We prove bounds on $\alpha^\star$ in terms of $\|\widetilde{\mathbf{y}}\|$ and $\|w\|$ in Lemma 4.12. We in turn bound $\|\widetilde{\mathbf{y}}\|$ in terms of $\|\theta^\star\|$ and $\|D^\dagger U^\top \varepsilon\|$ in Lemma 4.13. Next, in Lemma 4.14 we show that, with high probability, $\|D^\dagger U^\top \varepsilon\|$ is close to $\sigma^2 \mathrm{Tr}\left((XX^\top)^{-1} \right)$. Recall that, in Section 4.2, we showed that $\mathrm{Tr}\left((XX^\top)^{-1} \right)$ concentrates around n/s_k.

The next lemma provides an upper and lower bound on $\alpha^\star$.

Lemma 4.12. *The scaling factor $\alpha^\star$ satisfies the following*

$$\frac{2\|\widetilde{\mathbf{y}}\|^{1/2}}{3} \leq \alpha^\star \leq \frac{2\|\widetilde{\mathbf{y}}\|^{1/2}}{3}\sqrt{\sqrt{1+\frac{4\|w\|^4}{81\|\widetilde{\mathbf{y}}\|^2}}+\frac{2\|w\|^2}{9\|\widetilde{\mathbf{y}}\|}}.$$

Proof. Recall the definition of $\alpha^\star$ from above

$$\alpha^\star = \sqrt{\frac{8\|\widetilde{w}_{n+1:p}\|^2 + \sqrt{64\|\widetilde{w}_{n+1:p}\|^4 + 1296\|\widetilde{\mathbf{y}}\|^2}}{81}}.$$

Note that $\|\widetilde{w}_{n+1:p}\| \geq 0$. This immediately leads to the lower bound. For the upper bound note that $\|\widetilde{w}_{n+1:p}\| \leq \|\widetilde{w}\| = \|V^\top w\| = \|w\|$, since V is a unitary matrix. ■

The following lemma provides high probability upper and lower bounds on the norm of $\widetilde{\mathbf{y}}$.

Lemma 4.13. *The squared norm of $\widetilde{\mathbf{y}}$ satisfies the following*

$$\|D^\dagger U^\top \varepsilon\|^2 \left(1 - \frac{2\|\theta^\star\|}{\|D^\dagger U^\top \varepsilon\|}\right) \leq \|\widetilde{\mathbf{y}}\|^2 \leq \|D^\dagger U^\top \varepsilon\|^2 \left(1 + \frac{\|\theta^\star\|}{\|D^\dagger U^\top \varepsilon\|}\right)^2.$$

Proof. Recall that $UDV^\top$ is the SVD of X, $\widetilde{\mathbf{y}} = D^\dagger U^\top \mathbf{y}$ and that $\mathbf{y} = X\theta^\star + \varepsilon$. Therefore

$$\begin{aligned}\widetilde{\mathbf{y}} = D^\dagger U^\top (X\theta^\star + \varepsilon) = D^\dagger U^\top (UDV^\top \theta^\star + \varepsilon) &= D^\dagger D V^\top \theta^\star + D^\dagger U^\top \varepsilon \\ &= \begin{bmatrix} I_n \\ 0_{(p-n)\times n}\end{bmatrix} V^\top \theta^\star + D^\dagger U^\top \varepsilon.\end{aligned}$$

Define $\widetilde{\theta}^\star := V^\top \theta^\star$ and so

$$\widetilde{\mathbf{y}} = \widetilde{\theta}^\star_{1:n} + D^\dagger U^\top \varepsilon.$$

Thus,

$$\|\widetilde{\mathbf{y}}\|^2 = \|\widetilde{\theta}^\star_{1:n}\|^2 + \|D^\dagger U^\top \varepsilon\|^2 + 2\left(\varepsilon^\top U D^{\dagger\top}\right)\left(\widetilde{\theta}^\star_{1:n}\right).$$

Now since $0 \leq \|\widetilde{\theta}^\star_{1:n}\| \leq \|\widetilde{\theta}^\star\| = \|V^\top \theta^\star\| = \|\theta^\star\|$ we get that

$$\|\widetilde{\mathbf{y}}\|^2 \geq \|D^\dagger U^\top \varepsilon\|^2 - 2\|D^\dagger U^\top \varepsilon\|\|\theta^\star\| = \|D^\dagger U^\top \varepsilon\|^2 \left(1 - \frac{2\|\theta^\star\|}{\|D^\dagger U^\top \varepsilon\|}\right)$$

and also that

$$\begin{aligned}\|\widetilde{\mathbf{y}}\|^2 &\leq \|D^\dagger U^\top \varepsilon\|^2 + 2\|D^\dagger U^\top \varepsilon\|\|\theta^\star\| + \|\theta^\star\|^2 \\ &= \left(\|D^\dagger U^\top \varepsilon\| + \|\theta^\star\|\right)^2 = \|D^\dagger U^\top \varepsilon\|^2 \left(1 + \frac{\|\theta^\star\|}{\|D^\dagger U^\top \varepsilon\|}\right)^2,\end{aligned}$$

which establishes our claim. ■

The next result upper and lower bounds $\|D^\dagger U^\top \varepsilon\|^2$ with high probability.

Lemma 4.14. *For any $t \geq 0$, with probability at least $1 - 2e^{-t}$*

$$\left|\|D^\dagger U^\top \varepsilon\|^2 - \sigma^2 \mathrm{Tr}\left((XX^\top)^{-1}\right)\right| \leq c \max\left\{\frac{t}{\mu_n(XX^\top)}, \sqrt{t\sum_{i=1}^n \frac{1}{\mu_i^2(XX^\top)}}\right\}.$$

Proof. Let $u_1, \dots, u_n$ be the columns of U. The matrix U is unitary so each column u_i has unit norm. So

$$\|D^\dagger U^\top \varepsilon\|^2 = \sum_{i=1}^n \frac{(u_i^\top \varepsilon)^2}{D_{ii}^2} = \sum_{i=1}^n \frac{(u_i^\top \varepsilon)^2}{\mu_i(XX^\top)}$$

and

$$\mathbb{E}_\varepsilon \left[\|D^\dagger U^\top \varepsilon\|^2 \mid X \right] = \sum_{i=1}^n \frac{\mathbb{E}\left[(u_i^\top \varepsilon)^2 \mid X\right]}{D_{ii}^2} = \sum_{i=1}^n \frac{\sigma^2}{D_{ii}^2} = \sigma^2 \mathrm{Tr}\left((XX^\top)^{-1}\right).$$

Since the components are ε are independent, σ_y^2-sub-Gaussian random variables, with variance σ^2, by invoking the Hanson-Wright inequality [see RV13, Theorem 1] we infer that

$$\begin{aligned}
\left| \|D^\dagger U^\top \varepsilon\|^2 - \sigma^2 \mathrm{Tr}((XX^\top)^{-1}) \right| &= \left| \varepsilon^\top \left(U D^{\dagger\top} D^\dagger U^\top \right) \varepsilon - \sigma^2 \mathrm{Tr}((XX^\top)^{-1}) \right| \\
&\le c_1 \sigma_y^2 \max \left\{ \frac{t}{\mu_n(XX^\top)}, \sqrt{t \cdot \sum_{i=1}^n \frac{1}{\mu_i^2(XX^\top)}} \right\} \\
&= c \max \left\{ \frac{t}{\mu_n(XX^\top)}, \sqrt{t \cdot \sum_{i=1}^n \frac{1}{\mu_i^2(XX^\top)}} \right\}
\end{aligned}$$

with probability at least $1 - 2e^{-t}$, completing the proof. ∎

With these lemmas in place we are now ready to prove Lemma 4.3. We restate it here.

Lemma 4.3. *There are constants $c_0, \dots, c_5$ such that for any $\delta \in (e^{-c_0\sqrt{n}}, 1)$, if $p \ge c_1(n+k)$, $n \ge c_2 \max\{k, s_k\}$ and $\|\theta^\star\|, \|w\| \le c_3$ then with probability at least $1 - c_4\delta$,*

$$\left| \frac{\alpha^\star}{\frac{2\sqrt{\sigma} n^{1/4}}{3 s_k^{1/4}}} - 1 \right| \le c_5 \left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \sqrt{\frac{s_k}{n}} + \sqrt{\frac{\log(2/\delta)}{n}} + \frac{k}{n} \right].$$

Proof. Using Lemma 4.6, with probability at least $1 - c_6 e^{-c_7 n}$,

$$\left| \mathrm{Tr}(XX^{-1}) - \frac{n}{s_k} \right| \le \frac{c_8 n}{s_k} \left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \frac{k}{n} + e^{-c_9\sqrt{n}} \right]. \tag{42}$$

Next, by Lemma 4.4, with probability at least $1 - 2e^{-\sqrt{n}}$, for all $i \in [n]$

$$\begin{aligned}
\mu_i(XX^\top) &\ge s_k \left[1 - c_{10} \left(\frac{n + \sqrt{n}}{r_k} + \sqrt{\frac{n+\sqrt{n}}{R_k}} \right) \right] \\
&\ge s_k \left[1 - c_{11} \left(\frac{n}{r_k} + \sqrt{\frac{n}{r_k}} \right) \right] && \text{(since } R_k \ge r_k \text{ by Lemma 2.2)} \\
&\ge c_{12} s_k && \text{(since } r_k \ge bn).
\end{aligned}$$

This, combined with Lemma 4.14, tells us that for any $\delta \in (e^{-c_0\sqrt{n}}, 1)$ with probability at least $1 - c_{13}\delta$

$$\begin{aligned}\left| \|D^\dagger U^\top \boldsymbol{\varepsilon}\|^2 - \sigma^2 \mathrm{Tr}((XX^\top)^{-1}) \right| &\le c_{14} \max\left\{ \frac{\log(2/\delta)}{s_k}, \frac{\sqrt{n\log(2/\delta)}}{s_k} \right\} \\ &\le \frac{c_{14}\sqrt{n\log(2/\delta)}}{s_k}.\end{aligned}$$

Combining this with inequality (42) and recalling that σ^2 is a constant, we infer that, with probability at least $1 - c_3\delta$,

$$\begin{aligned}\left| \|D^\dagger U^\top \boldsymbol{\varepsilon}\|^2 - \frac{\sigma^2 n}{s_k} \right| &\le \frac{c_{15} n}{s_k} \left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + e^{-c_9\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{n}} + \frac{k}{n} \right] \\ &\le \frac{c_{16} n}{s_k} \left[\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \sqrt{\frac{\log(2/\delta)}{n}} + \frac{k}{n} \right] && (43) \\ &\le \frac{c_{17} n}{s_k}, && (44)\end{aligned}$$

where the last inequality follows since $R_k \ge r_k \ge bn$, $n \ge c_2 k$ and since $\delta \ge e^{-c_0\sqrt{n}}$.

We shall assume that condition (43) holds going forward. (This determines the success probability in the statement of the lemma.) Now since $r_k \ge bn$ and $n \ge c_2 \max\{k, s_k\}$ for a large enough constants b and c_2, by invoking Lemma 4.13, we find that

$$\begin{aligned}\|\widetilde{\mathbf{y}}\|^2 &\le \|D^\dagger U^\top \boldsymbol{\varepsilon}\|^2 \left(1 + \frac{\|\theta^\star\|}{\|D^\dagger U^\top \boldsymbol{\varepsilon}\|} \right)^2 \\ &= \|D^\dagger U^\top \boldsymbol{\varepsilon}\|^2 \left(1 + \frac{\|\theta^\star\|^2}{\|D^\dagger U^\top \boldsymbol{\varepsilon}\|^2} + \frac{2\|\theta^\star\|}{\|D^\dagger U^\top \boldsymbol{\varepsilon}\|} \right) \\ &\overset{(i)}{\le} \frac{\sigma^2 n}{s_k} \left[1 + c_{16} \left(\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \sqrt{\frac{\log(2/\delta)}{n}} + \frac{k}{n} \right) \right] \\ &\qquad \times \left(1 + c_{18} \left(\frac{\|\theta^\star\|^2 s_k}{n} + \frac{\|\theta^\star\|\sqrt{s_k}}{\sqrt{n}} \right) \right) \\ &\overset{(ii)}{\le} \frac{\sigma^2 n}{s_k} \left[1 + c_{16} \left(\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \sqrt{\frac{\log(2/\delta)}{n}} + \frac{k}{n} \right) \right] \left(1 + c_{19}\sqrt{\frac{s_k}{n}} \right) && (45) \\ &\le \frac{c_{20} n}{s_k}, && (46)\end{aligned}$$

where (i) follows by applying inequalities (43) and (44), and also because σ^2 is a constant. The second inequality (ii) follows since $\|\theta^\star\| \le c_3$ and because $n \ge c_2 s_k$. Also, by Lemma 4.13, we get that

$$\begin{aligned}\|\widetilde{\mathbf{y}}\|^2 &\ge \|D^\dagger U^\top \boldsymbol{\varepsilon}\|^2 \left(1 - \frac{2\|\theta^\star\|}{\|D^\dagger U^\top \boldsymbol{\varepsilon}\|} \right) \\ &\ge \frac{\sigma^2 n}{s_k} \left[1 - c_{16} \left(\sqrt{\frac{n}{R_k}} + \frac{n}{r_k} + \sqrt{\frac{\log(2/\delta)}{n}} + \frac{k}{n} \right) \right] \left(1 - c_{21}\sqrt{\frac{s_k}{n}} \right) && (47) \\ &\ge \frac{c_{22} n}{s_k}, && (48)\end{aligned}$$

where the last two inequalities follow by repeating the logic from the previous equation block.

Now recall that, by Lemma 4.12,

$$\frac{2\|\widetilde{\mathbf{y}}\|^{1/2}}{3} \le \alpha^\star \le \frac{2\|\widetilde{\mathbf{y}}\|^{1/2}}{3}\sqrt{\sqrt{1+\frac{4\|w\|^4}{81\|\widetilde{\mathbf{y}}\|^2}}+\frac{2\|w\|^2}{9\|\widetilde{\mathbf{y}}\|}}. \tag{49}$$

Using the lower bound in the equation above combining with inequality (47) we find that

$$\alpha^\star \ge \frac{2\sqrt{\sigma}n^{1/4}}{3s_k^{1/4}}\left[1-c_{16}\left(\sqrt{\frac{n}{R_k}}+\frac{n}{r_k}+\sqrt{\frac{\log(2/\delta)}{n}}+\frac{k}{n}\right)\right]\left(1-c_{23}\sqrt{\frac{s_k}{n}}\right),$$

and since $n \ge c_2 s_k$ for a large enough constant c_2,

$$\frac{\alpha^\star}{\frac{2\sqrt{\sigma}n^{1/4}}{3s_k^{1/4}}}-1 \ge -c_{24}\left[\sqrt{\frac{n}{R_k}}+\frac{n}{r_k}+\sqrt{\frac{s_k}{n}}+\sqrt{\frac{\log(2/\delta)}{n}}+\frac{k}{n}\right]. \tag{50}$$

Now for the upper bound, since $\|w\| \le c_3$, by using (48) we have that $\|w\|^2/\|\widetilde{\mathbf{y}}\| \le 1/20$, since $n > c_2 s_k$, where c_2 is large enough. Thus, by (49),

$$\begin{aligned}
\alpha^\star &\le \frac{2\|\widetilde{\mathbf{y}}\|^{1/2}}{3}\left(1+\frac{c_{25}\|w\|}{\|\widetilde{\mathbf{y}}\|^{1/2}}\right) \\
&\le \frac{2\sqrt{\sigma}n^{1/4}}{3s_k^{1/4}}\left[1+c_{26}\left(\sqrt{\frac{n}{R_k}}+\frac{n}{r_k}+\sqrt{\frac{\log(2/\delta)}{n}}+\frac{k}{n}\right)\right]\left(1+c_{27}\sqrt{\frac{s_k}{n}}\right)
\end{aligned}$$

and so

$$\frac{\alpha^\star}{\frac{2\sqrt{\sigma}n^{1/4}}{3s_k^{1/4}}}-1 \le c_{28}\left[\sqrt{\frac{n}{R_k}}+\frac{n}{r_k}+\sqrt{\frac{s_k}{n}}+\sqrt{\frac{\log(2/\delta)}{n}}+\frac{k}{n}\right].$$

This combined with (50) completes the proof. ■

4.4 Proof of Theorem 3.1

Let us first restate the theorem.

Theorem 3.1. *Under Assumptions (A.1)-(A.6), there exist constants $c_0, \dots, c_7$ such that for any $\delta \in (e^{-c_0\sqrt{n}}, 1-c_1e^{-c_2 n})$, if $p \ge c_3(n+k)$, $n \ge c_4 \max\{k, s_k\}$ and $\|\theta^\star\|, \|w\| \le c_5$ then with probability at least $1-c_6\delta$*

$$\mathsf{Risk}(\hat{\theta}) \le \mathsf{Bias}+\mathsf{Variance}+\Xi,$$

where

$$\begin{aligned}
\mathsf{Bias} &\le c_7\left(\|(\theta^\star-\psi)_{1:k}\|^2_{\Sigma^{-1}_{1:k}}\left(\frac{s_k}{n}\right)^2+\|(\theta^\star-\psi)_{k+1:p}\|^2_{\Sigma_{k+1:p}}\right) \le \frac{2c_7\|\theta^\star-\psi\|^2 s_k}{n}; \\
\mathsf{Variance} &\le c_7\log(1/\delta)\left(\frac{k}{n}+\frac{n}{R_k}\right); \\
\Xi &\le c_7\lambda_1\|\psi\|^2\left[\frac{n}{R_k}+\frac{n^2}{r_k^2}+\frac{s_k}{n}+\frac{\log(1/\delta)}{n}+\frac{k^2}{n^2}\right]\max\left\{\sqrt{\frac{r_0}{n}}, \frac{r_0}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}.
\end{aligned}$$

Proof. By Lemma 4.2, we know that

$$\mathsf{Risk}(\hat{\theta}) \leq c_8(\theta^\star - \alpha^\star w)^\top B(\theta^\star - \alpha^\star w) + c_8 \log(1/\delta)\mathrm{Tr}(C)$$

with probability at least $1 - \delta$, where the matrices

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

Recall that $\psi = \frac{2\sqrt{\sigma}n^{1/4}w}{3s_k^{1/4}}$. Thus, with the same probability

$$\begin{aligned} \mathsf{Risk}(\hat{\theta}) &\leq c_8\left(\theta^\star - \psi - (\alpha^\star w - \psi)\right)^\top B\left(\theta^\star - \psi - (\alpha^\star w - \psi)\right) + c_8\log(1/\delta)\mathrm{Tr}(C) \\ &= c_8\|\theta^\star - \psi - (\alpha^\star w - \psi)\|_B^2 + c_8\log(1/\delta)\mathrm{Tr}(C) \\ &\leq 2c_8\|\theta^\star - \psi\|_B^2 + 2c_8\|\alpha^\star w - \psi\|_B^2 + c_8\log(1/\delta)\mathrm{Tr}(C) \\ &= \underbrace{2c_8\left(\theta^\star - \psi\right)^\top B\left(\theta^\star - \psi\right)}_{\text{“Bias”}} + \underbrace{c_8\log(1/\delta)\mathrm{Tr}(C)}_{\text{“Variance”}} + \underbrace{2c_8\|B\|_{op}\|\alpha^\star w - \psi\|^2}_{\text{“}\Xi\text{”}}. \end{aligned} \tag{51}$$

We shall bound each of the three terms in the inequality above to establish the theorem.

Recall the definition of the matrix $T = \sum_{j>k}\lambda_j z_j z_j^\top$ from Definition 4.7 above. Define $S := \{j : j > k\}$, and let $X_S \in \mathbb{R}^{n\times|S|}$ be the submatrix formed by the last $p - k$ columns of $X \in \mathbb{R}^{n\times p}$. It can be verified that $T = X_S X_S^\top$. By Lemma 4.4, with probability at least $1 - 2e^{-n} \geq 1 - c_9\delta$, (since $\delta \geq e^{-c_0\sqrt{n}}$)

$$\mu_1(T) \leq c_{10}\sum_{j>k}\lambda_j \quad \text{and} \quad \mu_n(T) \geq c_{11}\sum_{j>k}\lambda_j.$$

Therefore, the condition number of the matrix T is a constant with the same probability. Assuming this bound on the condition number holds we shall bound the first two terms in (51).

Bound on the bias and variance: Since the condition number of T is at most a constant, by invoking [TB20, Theorem 1] we get that with probability at least $1 - c_{12}\delta$

$$\mathsf{Bias} \leq c_7\left(\|(\theta^\star - \psi)_{1:k}\|_{\Sigma_{1:k}^{-1}}^2\left(\frac{s_k}{n}\right)^2 + \|(\theta^\star - \psi)_{k+1:p}\|_{\Sigma_{k+1:p}}^2\right) \tag{52}$$

and

$$\mathsf{Variance} \leq c_7\log(1/\delta)\left(\frac{k}{n} + \frac{n}{R_k}\right). \tag{53}$$

We simplify our upper bound on Bias by noting that under our choice of k as follows:

$$
\begin{aligned}
&c_7\left(\|(\theta^\star-\psi)_{1:k}\|^2_{\Sigma^{-1}_{1:k}}\left(\frac{s_k}{n}\right)^2+\|(\theta^\star-\psi)_{k+1:p}\|^2_{\Sigma_{k+1:p}}\right)\\
&\quad= c_7\sum_{i=1}^p\left[\mathbb{I}(i\le k)(\theta^\star_i-\psi_i)^2\frac{s_k^2}{n^2\lambda_i}+\mathbb{I}(i>k)\lambda_i(\theta^\star_i-\psi_i)^2\right]\\
&\quad= c_7\sum_{i=1}^p\lambda_i(\theta^\star_i-\psi_i)^2\left[\mathbb{I}(i\le k)\frac{s_k^2}{n^2\lambda_i^2}+\mathbb{I}(i>k)\right]\\
&\quad= c_7\sum_{i=1}^p\lambda_i(\theta^\star_i-\psi_i)^2\frac{(\frac{s_k}{n})^2}{(\frac{s_k}{n})^2+\lambda_i^2}\left[\mathbb{I}(i\le k)\left(1+\frac{1}{\lambda_i^2}\left(\frac{s_k}{n}\right)^2\right)+\mathbb{I}(i>k)\left(1+\lambda_i^2\left(\frac{n}{s_k}\right)^2\right)\right]\\
&\quad\overset{(i)}{\le} c_7\sum_{i=1}^p\lambda_i(\theta^\star_i-\psi_i)^2\frac{(\frac{s_k}{n})^2}{(\frac{s_k}{n})^2+\lambda_i^2}\left[\mathbb{I}(i\le k)\left(1+b^2\right)+\mathbb{I}(i>k)\left(1+\lambda_i^2\left(\frac{n}{s_k}\right)^2\right)\right]\\
&\quad\le c_7\sum_{i=1}^p\lambda_i(\theta^\star_i-\psi_i)^2\frac{(\frac{s_k}{n})^2}{(\frac{s_k}{n})^2+\lambda_i^2}\left[\mathbb{I}(i\le k)\left(1+b^2\right)+\mathbb{I}(i>k)\left(1+\lambda_{k+1}^2\left(\frac{n}{s_k}\right)^2\right)\right]\\
&\quad\le c_7\sum_{i=1}^p\lambda_i(\theta^\star_i-\psi_i)^2\frac{(\frac{s_k}{n})^2}{(\frac{s_k}{n})^2+\lambda_i^2}\left[\mathbb{I}(i\le k)\left(1+b^2\right)+\mathbb{I}(i>k)\left(1+\left(\frac{n}{r_k}\right)^2\right)\right]\\
&\quad\overset{(ii)}{\le} c_7\sum_{i=1}^p\lambda_i(\theta^\star_i-\psi_i)^2\frac{(\frac{s_k}{n})^2}{(\frac{s_k}{n})^2+\lambda_i^2}\left[\mathbb{I}(i\le k)\left(1+b^2\right)+\mathbb{I}(i>k)\left(1+\frac{1}{b^2}\right)\right]\\
&\quad\le c_{13}\sum_{i=1}^p\lambda_i(\theta^\star_i-\psi_i)^2\frac{(\frac{s_k}{n})^2}{(\frac{s_k}{n})^2+\lambda_i^2},
\end{aligned}
$$

where (i) follows since by definition $k=\min\{j\ge 0: r_j\ge bn\}$ and so for $i\le k$, $s_k/\lambda_i\le s_i/\lambda_i=r_i<bn$. Inequality (ii) follows since $r_k\ge bn$. Continuing we get that

$$
\begin{aligned}
&c_7\left(\|(\theta^\star-\psi)_{1:k}\|^2_{\Sigma^{-1}_{1:k}}\left(\frac{s_k}{n}\right)^2+\|(\theta^\star-\psi)_{k+1:p}\|^2_{\Sigma_{k+1:p}}\right)\\
&\quad\le c_{13}\sum_{i=1}^p\lambda_i(\theta^\star-\psi)_i^2\frac{\left(\frac{s_k}{n}\right)^2}{\left(\frac{s_k}{n}\right)^2+\lambda_i^2}\\
&\quad= c_{13}\left(\frac{s_k}{n}\right)^2\sum_{i=1}^p(\theta^\star-\psi)_i^2\frac{\lambda_i}{\left(\frac{s_k}{n}\right)^2+\lambda_i^2}\\
&\quad\le c_{13}\left(\frac{s_k}{n}\right)^2\|\theta^\star-\psi\|^2\max_{i\in[p]}\frac{\lambda_i}{\left(\frac{s_k}{n}\right)^2+\lambda_i^2} \qquad \text{(by Hölder's inequality)}\\
&\quad\le c_{13}\left(\frac{s_k}{n}\right)^2\|\theta^\star-\psi\|^2\max_{\zeta\ge 0}\frac{\zeta}{\left(\frac{s_k}{n}\right)^2+\zeta^2}=\frac{2c_{13}\|\theta^\star-\psi\|^2 s_k}{n}. \qquad (54)
\end{aligned}
$$

Bound on Ξ (the estimation error of $\alpha^\star$): By Lemma 4.3 with probability at least $1-c_{14}\delta$

$$
\left|\alpha^\star-\frac{2\sqrt{\sigma}n^{1/4}}{3s_k^{1/4}}\right|\le c_{15}\frac{2\sqrt{\sigma}n^{1/4}}{3s_k^{1/4}}\left[\sqrt{\frac{n}{R_k}}+\frac{n}{r_k}+\sqrt{\frac{s_k}{n}}+\sqrt{\frac{\log(2/\delta)}{n}}+\frac{k}{n}\right]
$$

and therefore,

$$\begin{aligned}\|\alpha^\star w-\psi\| &\le c_{16}\|\psi\|\left[\sqrt{\frac{n}{R_k}}+\frac{n}{r_k}+\sqrt{\frac{s_k}{n}}+\sqrt{\frac{\log(2/\delta)}{n}}+\frac{k}{n}\right]\\ &\le c_{17}\|\psi\|\left[\sqrt{\frac{n}{R_k}}+\frac{n}{r_k}+\sqrt{\frac{s_k}{n}}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{k}{n}\right] \quad (\text{since } \delta\le 1-c_1e^{-c_2n}). \quad (55)\end{aligned}$$

To control the operator norm of B, we first observe that

$$\begin{aligned}\|B\|_{op} &= \left\|\left(I-X^\top(XX^\top)^{-1}X\right)\Sigma\left(I-X^\top(XX^\top)^{-1}X\right)\right\|_{op}\\ &= \left\|\left(I-X^\top(XX^\top)^{-1}X\right)\left(\Sigma-\frac{X^\top X}{n}\right)\left(I-X^\top(XX^\top)^{-1}X\right)\right\|_{op}\\ &\le \left\|I-X^\top(XX^\top)^{-1}X\right\|_{op}^2\left\|\Sigma-\frac{X^\top X}{n}\right\|_{op}\\ &\le \left\|\Sigma-\frac{X^\top X}{n}\right\|_{op}.\end{aligned}$$

Thus, by invoking [KL17, Theorem 9] we get that with probability at least $1-\delta$

$$\begin{aligned}\|B\|_{op} &\le c_{18}\lambda_1\max\left\{\sqrt{\frac{r_0}{n}},\frac{r_0}{n},\sqrt{\frac{\log(1/\delta)}{n}},\frac{\log(1/\delta)}{n}\right\}\\ &\le c_{18}\lambda_1\max\left\{\sqrt{\frac{r_0}{n}},\frac{r_0}{n},\sqrt{\frac{\log(1/\delta)}{n}}\right\}, \quad (56)\end{aligned}$$

where the second inequality follows since $\delta\ge e^{-c_0\sqrt{n}}$.

Combining inequalities (55) and (56) we get that with probability at least $1-c_{19}\delta$

$$\begin{aligned}2c_8\|B\|_{op}\|\alpha^\star w-\psi\|^2 &\le c_{20}\lambda_1\|\psi\|^2\max\left\{\sqrt{\frac{r_0}{n}},\frac{r_0}{n},\sqrt{\frac{\log(1/\delta)}{n}}\right\}\\ &\qquad\times\left[\frac{n}{R_k}+\frac{n^2}{r_k^2}+\frac{s_k}{n}+\frac{\log(2/\delta)}{n}+\frac{k^2}{n^2}\right]. \quad (57)\end{aligned}$$

Combining inequalities (52), (53), (54) and (57) along with a union bound completes the proof. ■

5 Proof of Proposition 3.4

Recall the statement of the proposition.

Proposition 3.4. *If $a(0)$ and $W(0)$ are chosen randomly, independent of X and $\mathbf{y}$, so that the distribution of $a(0)^\top W(0)$ is symmetric about the origin, then*

$$\mathbb{E}_{a(0),W(0),X,\mathbf{y}}[\mathsf{Risk}(\hat\theta)]\ge\mathbb{E}\left[\theta^{\star\top}B\theta^\star\right]+\sigma^2\mathbb{E}\left[\mathrm{Tr}(C)\right],$$

where

$$\begin{aligned}B &:= \left(I-X^\top(XX^\top)^{-1}X\right)\Sigma\left(I-X^\top(XX^\top)^{-1}X\right) \quad \text{and}\\ C &:= (XX^\top)^{-1}X\Sigma X^\top(XX^\top)^{-1}.\end{aligned}$$

Proof. In the proof of Lemma 4.2, we showed that, for all $X, \mathbf{y}$, we have

$$\mathsf{Risk}(\hat{\theta}) = \mathbb{E}_x \left[\left(x^\top \left(I - X^\top (XX^\top)^{-1} X \right) (\theta^\star - \alpha^\star w) - x^\top X^\top (XX^\top)^{-1} \varepsilon \right)^2 \right].$$

Expanding the quadratic yields

$$\begin{aligned}\mathsf{Risk}(\hat{\theta}) = \mathbb{E}_x &\left[\left(x^\top \left(I - X^\top (XX^\top)^{-1} X \right) (\theta^\star - \alpha^\star w) \right)^2 \right] + \mathbb{E}_x \left[\left(x^\top X^\top (XX^\top)^{-1} \varepsilon \right)^2 \right] \\ &+ 2\mathbb{E}_x \left[\left(x^\top \left(I - X^\top (XX^\top)^{-1} X \right) \theta^\star \right) \left(x^\top X^\top (XX^\top)^{-1} \varepsilon \right) \right] \\ &- 2\mathbb{E}_x \left[\left(x^\top \left(I - X^\top (XX^\top)^{-1} X \right) (\alpha^\star w) \right) \left(x^\top X^\top (XX^\top)^{-1} \varepsilon \right) \right].\end{aligned}$$

Since the distribution of w is symmetric about the origin, and independent of X and $\mathbf{y}$, and since, for fixed X and $\mathbf{y}$, $\alpha^\star$ is determined as a function of w, after conditioning on X and $\mathbf{y}$, the distribution of $\alpha^\star w$ is symmetric about the origin, and therefore has zero mean. This, along with the fact that $\mathbb{E}[\varepsilon] = 0$, gives

$$\begin{aligned}\mathbb{E}[\mathsf{Risk}(\hat{\theta})] &= \mathbb{E} \left[\left(x^\top \left(I - X^\top (XX^\top)^{-1} X \right) (\theta^\star - \alpha^\star w) \right)^2 \right] + \mathbb{E}_x \left[\left(x^\top X^\top (XX^\top)^{-1} \varepsilon \right)^2 \right] \\ &= \mathbb{E}[(\theta^\star - \alpha^* w)^\top B (\theta - \alpha^* w)^\star] + \mathbb{E} \left[\left(x^\top X^\top (XX^\top)^{-1} \varepsilon \right)^2 \right] \\ &\geq \mathbb{E}[\theta^{\star\top} B \theta^\star] + \mathbb{E} \left[\left(x^\top X^\top (XX^\top)^{-1} \varepsilon \right)^2 \right] \\ &\geq \mathbb{E}[\theta^{\star\top} B \theta^\star] + \sigma^2 \mathbb{E} \left[\mathrm{Tr}(C) \right],\end{aligned}$$

completing the proof. ∎

6 Discussion

Despite the fact that parameterizing a linear model using a balanced, two-layer linear network has been shown in previous work to have a substantial effect on the inductive bias of gradient descent [Azu+21], it remains compatible with benign overfitting, and the initial weights also encode a potentially useful bias.

While Proposition 3.4 limits the prospects for improving our upper bounds, there still appears to be a gap between our upper and lower bounds.

Moving beyond the case where the initialization is balanced would be an interesting next step. We briefly note that, for the initial parameters to be balanced, it is necessary for the weight matrix in the first layer $W \in \mathbb{R}^{m \times p}$ to have rank one. In the case where there is a single neuron ($m = 1$), Theorem 2 by [Azu+21] characterizes the implicit bias of the final solution learnt by gradient flow on the squared loss. The techniques developed in this paper might perhaps be useful in bounding the excess risk of this solution.

Yet another interesting open question concerns characterizing the implicit bias of gradient flow with the squared loss in the case where a linear model is parameterized using a deeper representation than two layers, building on existing research [Gun+17; Gun+18b; Aro+19; Woo+20; GSD20; RC20; YKM21; Azu+21; JRG21]. It would also be interesting to prove corresponding excess risk bounds for such solutions, and to study the effect of depth on the generalization properties of the resulting models.

Acknowledgements

We would like to thank the anonymous reviewers for pointing out a bug in a previous version of the proof of Lemma 4.5, and for their many helpful suggestions.

A The design matrix has full rank (and more)

Lemma A.1. *Under Assumption (A.4), for any eigenvector v of Σ, and any sample size n, the projection of the rows of X onto the subspace of $\mathbb{R}^p$ orthogonal to v has rank n.*

Proof. Assume without loss of generality that Σ is diagonal and $v = (1, 0, 0, \ldots, 0)$. Let $x_1, \ldots, x_n$ denote the rows of X, and let $x'_1, \ldots, x'_n$ be obtained from $x_1, \ldots, x_n$ by replacing each of their first components with 0, thereby projecting them onto the subspace orthogonal to v. It suffices to prove that, almost surely, $x'_1, \ldots, x'_n$ are linearly independent.

We will prove this by induction. The base case, there $n = 1$, is straightforward. When $n > 1$, by the inductive hypothesis, $x'_1, \ldots, x'_{n-1}$ are linearly independent. Since $p > n$, Assumption (A.4) implies that the span of $x'_1, \ldots, x'_{n-1}$ has probability 0, so that, almost surely, x'_n is not a member this span, completing the proof. ■

B Concentration inequalities

For an excellent reference of sub-Gaussian and sub-exponential concentration inequalities we refer the reader to Vershynin [Ver18]. We begin by defining sub-Gaussian and sub-exponential random variables.

Definition B.1. *A random variable ϕ is sub-Gaussian if*

$$\|\phi\|_{\psi_2} := \inf \left\{t > 0 : \mathbb{E}[\exp(\phi^2/t^2)] < 2\right\}$$

is bounded. Further, $\|\phi\|_{\psi_2}$ is defined to be its sub-Gaussian norm.

Definition B.2. *A random variable ϕ is said to be sub-exponential if*

$$\|\phi\|_{\psi_1} := \inf \{t > 0 : \mathbb{E}[\exp(|\phi|/t) < 2]\}$$

is bounded. Further, $\|\phi\|_{\psi_1}$ is defined to be its sub-exponential norm.

Next we state a few well-known facts about sub-Gaussian and sub-exponential random variables.

Lemma B.3 (Vershynin 2018, Lemma 2.7.6)**.** *If a random variable ϕ is sub-Gaussian then ϕ^2 is sub-exponential with $\|\phi^2\|_{\psi_1} = \|\phi\|^2_{\psi_2}$.*

Lemma B.4 (Vershynin 2018, Lemma 2.7.10)**.** *If a random variable ϕ is sub-exponential then $\phi - \mathbb{E}[\phi]$ is sub-exponential with $\|\phi - \mathbb{E}[\phi]\|_{\psi_1} \leq c\|\phi\|_{\psi_1}$ for some positive constant c.*

We state Bernstein's inequality [see, e.g., Ver18, Theorem 2.8.1], a concentration inequality for a sum of independent sub-exponential random variables.

Theorem B.5. *For independent mean-zero sub-exponential random variables $\phi_1, \dots, \phi_m$, for every $\eta > 0$, we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{m} \phi_i\right| \geq \eta\right] \leq 2\exp\left(-c\min\left\{\frac{\eta^2}{\sum_{i=1}^{m}\|\phi_i\|_{\psi_1}^2}, \frac{\eta}{\max_i \|\phi_i\|_{\psi_1}}\right\}\right),$$

where c is a positive absolute constant.

Let us continue by defining an ε-net with respect to the Euclidean distance.

Definition B.6. *Let $S \subseteq \mathbb{R}^p$. A subset K is called an ε-net of S if every point in S is within Euclidean distance ε of some point in K.*

The following lemma bounds the size of a $1/4$-net of unit vectors in $\mathbb{R}^p$.

Lemma B.7. *Let S be the set of all unit vectors in $\mathbb{R}^p$. Then there exists a $1/4$-net of S of size 9^p.*

Proof. Follows immediately by invoking [Ver18, Corollary 4.2.13] with $\varepsilon = 1/4$. ■

B.1 Proof of Lemma 4.4

Let $\Sigma = \sum_{i=1}^{p} \lambda_i e_i e_i^\top$ be the spectral decomposition of the covariance matrix. Define the random vectors

$$z_i := \frac{Xe_i}{\sqrt{\lambda_i}} \in \mathbb{R}^n.$$

These random vectors z_i have entries that are independent, σ_x^2-sub-Gaussian random variables [see Bar+20, Lemma 8]. Note that we can write the matrix

$$X_S X_S^\top = \sum_{i \in S} \lambda_i z_i z_i^\top.$$

Further, its expected value is as follows:

$$\mathbb{E}\left[X_S X_S^\top\right] = \sum_{i\in S} \lambda_i \mathbb{E}\left[z_i z_i^\top\right] = \sum_{i\in S} \mathbb{E}\left[X e_i e_i^\top X^\top\right] = I_n \sum_{i \in S} \lambda_i = I_n s(S).$$

With this in place, we are now ready to prove our concentration results.

Lemma 4.4. *There exists a positive absolute constant c such that, for any subset $S \subseteq [p]$ and any $t \geq 0$, with probability at least $1 - 2e^{-t}$, for all $j \in \{1, \dots, \min(n, |S|)\}$*

$$\left|\mu_j(X_S X_S^\top) - s(S)\right| \leq cs(S)\left(\frac{t+n}{r(S)} + \sqrt{\frac{t+n}{R(S)}}\right).$$

Proof. We shall prove this bound in the case where the set $S = [p]$. The bound for any other subset S shall follow by exactly the same logic. First, note that by a standard ε-net argument [see, e.g, Bar+20, Lemma 25] to bound the operator norm we can use the following inequality:

$$\left\|XX^\top - I_n \sum_{i=1}^{p} \lambda_i\right\|_{op} \leq 2 \max_{v_j \in \mathcal{N}_{\frac{1}{4}}} \left|v_j^\top \left(XX^\top - I_n \sum_{i=1}^{p}\lambda_i\right) v_j\right|, \tag{58}$$

where $\mathcal{N}_{\frac{1}{4}}$ is a $1/4$-net of the unit sphere with respect to the Euclidean norm of size at most 9^n. (We know that such a net exists by Lemma B.7.) Consider an arbitrary unit vector $v \in \mathbb{S}^{n-1}$. Then

$$v^\top \left(XX^\top - I_n \sum_{i=1}^{p} \lambda_i \right) v = \sum_{i=1}^{p} \lambda_i \left((z_i^\top v)^2 - 1 \right). \tag{59}$$

By Lemmas B.3 and B.4 we know that the random variables $\lambda_i((z_i^\top v)^2 - 1)$ are $c_1 \lambda_i \sigma_x^2$-sub-exponential, for some positive constant c_1.Therefore we can use Bernstein's inequality (see Theorem B.5) to upper bound the sum in equation (59) to get that, with probability at least $1 - 2e^{-t}$,

$$\left| \sum_{i=1}^{p} \lambda_i \left((z_i^\top v)^2 - 1 \right) \right| \leq c_2 \sigma_x^2 \max \left\{ \lambda_1 t, \sqrt{t \sum_{j=1}^{p} \lambda_j^2} \right\}. \tag{60}$$

Next by a union bound over all the elements of the cover $\mathcal{N}_{\frac{1}{4}}$ we find that, with probability at least $1 - 2e^{-t}$, for all $v \in \mathcal{N}_{\frac{1}{4}}$,

$$\left| \sum_{i=1}^{p} \lambda_i \left((z_i^\top v)^2 - 1 \right) \right| \leq c_2 \sigma_x^2 \max \left\{ \lambda_1 \left(t + n \log(9) \right), \sqrt{\left(t + n \log(9) \right) \sum_{j=1}^{p} \lambda_j^2} \right\}.$$

Hence, by using inequality (58) we get that with probability at least $1 - 2e^{-t}$

$$\begin{aligned} \left\| XX^\top - I_n \sum_{i=1}^{p} \lambda_i \right\|_{op} &\leq c_3 \sigma_x^2 \max \left\{ \lambda_1 \left(t + n \log(9) \right), \sqrt{\left(t + n \log(9) \right) \sum_{j=1}^{p} \lambda_j^2} \right\} \\ &\leq c_4 \sigma_x^2 \left(\lambda_1 (t+n) + \sqrt{(t+n) \sum_{j=1}^{p} \lambda_j^2} \right). \end{aligned}$$

Recalling that σ_x is assumed to be a positive constant, this implies that the greatest and least eigenvalues of $XX^\top$ are within $c_5 \left(\lambda_1 (t+n) + \sqrt{(t+n) \sum_{j=1}^{p} \lambda_j^2} \right)$ of $\sum_{i=1}^{p} \lambda_i$, which in turn implies

$$\begin{aligned} \left| \mu_j(XX^\top) - \sum_{i=1}^{p} \lambda_i \right| &\leq c_5 \left(\lambda_1 (t+n) + \sqrt{(t+n) \sum_{j=1}^{p} \lambda_j^2} \right) \\ &= c_5 \left(\sum_{i=1}^{p} \lambda_i \right) \left(\frac{t+n}{r_0} + \sqrt{\frac{t+n}{R_0}} \right), \end{aligned}$$

completing the proof. ■

B.2 Proof of Lemma 4.5

We begin by proving an auxiliary lemma that relates the minimum singular value of a matrix to its approximation over an ε-net under the assumption that its operator norm is bounded. Recall that $X_S \in \mathbb{R}^{n \times |S|}$.

Lemma B.8. *Let $\mathcal{N}_\varepsilon$ be an ε-net of the unit sphere in $\mathbb{R}^n$ with respect to the Euclidean norm. For any $a, b \geq 0$, if*

$$\inf_{z \in \mathbb{S}^{n-1}} \|X_S^\top z\| \leq a - \varepsilon b \quad \textit{and} \quad \|X_S^\top\|_{op} \leq b$$

then $\inf_{z \in \mathcal{N}_\varepsilon} \|X_S^\top z\| \leq a$.

Proof. Let ζ be a function that maps any unit vector z to its nearest neighbour (with respect to the Euclidean norm) in the net $\mathcal{N}_\varepsilon$. Therefore, if $\|X_S^\top\|_{op} \leq b$ then

$$\begin{aligned}
\inf_{z \in \mathbb{S}^{n-1}} \|X_S^\top z\| &= \inf_{z \in \mathbb{S}^{n-1}} \|X_S^\top (z - \zeta(z)) + X_S^\top \zeta(z)\| \\
&\geq \inf_{z \in \mathbb{S}^{n-1}} \|X_S^\top \zeta(z)\| - \inf_{z \in \mathbb{S}^{n-1}} \|X_S^\top (z - \zeta(z))\| \\
&= \inf_{z \in \mathcal{N}_\varepsilon} \|X_S^\top z\| - \inf_{z \in \mathbb{S}^{n-1}} \|X_S^\top (z - \zeta(z))\| \\
&\geq \inf_{z \in \mathcal{N}_\varepsilon} \|X_S^\top z\| - \|X_S^\top\|_{op} \inf_{z \in \mathbb{S}^{n-1}} \|z - \zeta(z)\| \\
&\geq \inf_{z \in \mathcal{N}_\varepsilon} \|X_S^\top z\| - \varepsilon b.
\end{aligned}$$

Further if $\inf_{z \in \mathbb{S}^{n-1}} \|X_S^\top z\| \leq a - \varepsilon b$ then, due to the inequality above, $\inf_{z \in \mathcal{N}_\varepsilon} \|X_S^\top z\| \leq a$ which completes the proof. ■

With this lemma in place let us prove our result.

Lemma 4.5. *There exist absolute positive constants $c_0, \ldots, c_3$ such that given any subset $S \subseteq [p]$ if, $r(S) \geq c_0 n$ then for all $t < c_1 < 1$*

$$\mathbb{P}\left[\mu_n(X_S X_S^\top) \leq t \cdot s(S)\right] \leq (c_2 t)^{c_3 \cdot r(S)}.$$

Proof. To reduce notational burden in the proof we shall present a proof in the case where the $S = [p]$, and therefore $X_S = X$. For any other subset S the proof shall proceed in exactly the same manner.

In the proof we shall prove bounds on the smallest singular value of X, $s_{\min}(X)$. This immediately leads to a bound on $\mu_n(XX^\top) = s^2_{\min}(X)$.

Recall that $s_{\min}(X) = s_{\min}(X^\top)$. So we will instead prove a bound on the smallest singular value of $X^\top$ to simplify our calculations. For some parameter $h \geq c_4 \geq 1$ that will be set in the sequel, decompose the probability into

$$\begin{aligned}
&\mathbb{P}\left[s_{\min}(X^\top) \leq t\sqrt{\sum_{j=1}^p \lambda_j}\right] \\
&= \mathbb{P}\left[\left\{s_{\min}(X^\top) \leq t\sqrt{\sum_{j=1}^p \lambda_j}\right\} \cap \left\{\|X\|_{op} \leq h\sqrt{\lambda_1 p}\right\}\right] \\
&\qquad + \mathbb{P}\left[\left\{s_{\min}(X^\top) \leq t\sqrt{\sum_{j=1}^p \lambda_j}\right\} \cap \left\{\|X\|_{op} > h\sqrt{\lambda_1 p}\right\}\right] \\
&\leq \mathbb{P}\left[\left\{s_{\min}(X^\top) \leq t\sqrt{\sum_{j=1}^p \lambda_j}\right\} \cap \left\{\|X\|_{op} \leq h\sqrt{\lambda_1 p}\right\}\right] + \mathbb{P}\left[\|X\|_{op} > h\sqrt{\lambda_1 p}\right]. \qquad (61)
\end{aligned}$$

Now we will control each of these probabilities separately. First, let us control the second probability

$$\mathbb{P}\left[\|X\|_{op} > h\sqrt{\lambda_1 p}\right] = \mathbb{P}\left[\|X\Sigma^{-1/2}\Sigma^{1/2}\|_{op} > h\sqrt{\lambda_1 p}\right] \\ \leq \mathbb{P}\left[\|X\Sigma^{-1/2}\|_{op} > h\sqrt{p}\right] \leq e^{-c_4 h^2 p}, \tag{62}$$

by invoking Proposition 2.4 by Rudelson and Vershynin [RV09].

To control the first probability in inequality (61) we need the following definition. Given a random vector $\xi \in \mathbb{R}^p$ define the Lévy concentration function

$$\mathcal{L}(\xi; t) := \sup_{w \in \mathbb{R}^p} \mathbb{P}\left[\|\xi - w\| \leq t\right].$$

Let $\phi \in \mathbb{S}^{n-1}$ be a fixed unit vector. By Assumption (A.4) we know that for any $a \leq b \in \mathbb{R}$:

$$\mathbb{P}\left[(\Sigma^{-1/2}X^\top \phi)_i \in [a, b]\right] \leq c|b - a|. \tag{63}$$

Using this fact we find that for any $i \in [p]$:

$$\mathcal{L}((\Sigma^{-1/2}X^\top \phi)_i\,; 2t) = \sup_{w \in \mathbb{R}} \mathbb{P}\left[|(\Sigma^{-1/2}X^\top \phi)_i - w| \leq 2t\right] \\ = \sup_{w \in \mathbb{R}} \mathbb{P}\left[(\Sigma^{-1/2}X^\top \phi)_i \in [w - 2t, w + 2t]\right] \leq 4c_5 t.$$

Next by invoking Theorem 1.5 in [RV15] we infer that

$$\mathcal{L}\left(X^\top \phi; 2t\sqrt{\sum_{i=1}^p \lambda_i}\right) \leq (ct)^{c' r_0}.$$

This implies that

$$\mathbb{P}\left[\|X^\top \phi\| \leq 2t\sqrt{\sum_{i=1}^p \lambda_i}\right] \leq \sup_{w \in \mathbb{R}^p} \mathbb{P}\left[\|X^\top \phi - w\| \leq 2t\sqrt{\sum_{i=1}^p \lambda_i}\right] \\ = \mathcal{L}\left(X^\top \phi; 2t\sqrt{\sum_{i=1}^p \lambda_i}\right) \leq (ct)^{c' r_0}. \tag{64}$$

This establishes a *small-ball* probability (anti-concentration) for a fixed unit vector ϕ. We will now proceed by using an ε-net argument. For some $\varepsilon \in \left(0, \frac{2t}{h}\sqrt{\frac{\sum_{i=1}^p \lambda_i}{\lambda_1 p}}\right)$ let $\mathcal{N}_\varepsilon$ be an ε-net of the unit vectors in $\mathbb{R}^n$ with respect to the Euclidean norm of size at most $\left(\frac{2}{\varepsilon} + 1\right)^n$ (such a net exists, see, e.g., Corollary 4.2.13 in [Ver18]). By a union bound over the elements of the net

$$\mathbb{P}\left[\min_{\phi \in \mathcal{N}_\varepsilon} \|X^\top \phi\| \leq 2t\sqrt{\sum_{i=1}^p \lambda_i}\right] \leq (ct)^{c' r_0} \cdot \left(\frac{2}{\varepsilon} + 1\right)^n. \tag{65}$$

Next by Lemma B.8 we know that

$$
\begin{aligned}
&\mathbb{P}\left[\left\{s_{\min}(X^\top) \le 2t\sqrt{\sum_{i=1}^p \lambda_i - \varepsilon h\sqrt{\lambda_1 p}}\right\} \cap \left\{\|X\|_{op} \le h\sqrt{\lambda_1 p}\right\}\right] \\
&\quad = \mathbb{P}\left[\left\{\inf_{z\in\mathbb{S}^{n-1}} \|X^\top z\| \le 2t\sqrt{\sum_{i=1}^p \lambda_i - \varepsilon h\sqrt{\lambda_1 p}}\right\} \cap \left\{\|X\|_{op} \le h\sqrt{\lambda_1 p}\right\}\right] \\
&\quad \le \mathbb{P}\left[\min_{z\in\mathcal{N}_\varepsilon} \|X^\top z\| \le 2t\sqrt{\sum_{i=1}^p \lambda_i}\right] \\
&\quad \le (ct)^{c' r_0} \cdot \left(\frac{2}{\varepsilon} + 1\right)^n.
\end{aligned}
$$

Setting $\varepsilon = \frac{t}{h}\sqrt{\frac{\sum_{i=1}^p \lambda_i}{\lambda_1 p}} = \frac{t}{h}\sqrt{\frac{r_0}{p}}$ we get that

$$
\mathbb{P}\left[\left\{s_{\min}(X^\top) \le t\sqrt{\sum_{i=1}^p \lambda_i}\right\} \cap \left\{\|X\|_{op} \le h\sqrt{\lambda_1 p}\right\}\right] \le (ct)^{c' r_0} \cdot \left(\frac{2h}{t}\sqrt{\frac{p}{r_0}} + 1\right)^n.
$$

This combined with inequalities (61) and (62) above yields

$$
\mathbb{P}\left[s_{\min}(X^\top) \le t\sqrt{\sum_{i=1}^p \lambda_i}\right] \le (ct)^{c' r_0} \cdot \left(\frac{2h}{t}\sqrt{\frac{p}{r_0}} + 1\right)^n + e^{-c_4 h^2 p}.
$$

Finally set $h = \frac{1}{t}\sqrt{\frac{r_0}{p}}$ to obtain the bound

$$
\mathbb{P}\left[s_{\min}(X^\top) \le t\sqrt{\sum_{i=1}^p \lambda_i}\right] \le (ct)^{c' r_0} \cdot \left(\frac{c''}{t^2}\right)^n + e^{-c_5 r_0/t^2} \overset{(i)}{\le} (c_2 t)^{c_3 \cdot r_0}
$$

where (i) follows since $r_0 > c_0 n$ for a large enough constant c_0 and because $t < c_1$ for a small enough constant c_1. ∎

References

[Aro+19] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. "Implicit regularization in deep matrix factorization". In: *Advances in Neural Information Processing Systems*. 2019, pp. 7413–7424 (Cited on page 35).

[Azu+21] Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E. Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. "On the implicit bias of initialization shape: beyond infinitesimal mirror descent". In: *International Conference on Machine Learning*. 2021, pp. 468–477 (Cited on pages 2, 35).

[Bar+20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* (2020) (Cited on pages 2–8, 11, 13, 37).

[BMR21] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. "Deep learning: a statistical viewpoint". In: *arXiv preprint arXiv:2103.09177* (2021) (Cited on page 2).

[Bel21] Mikhail Belkin. "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation". In: *arXiv preprint arXiv:2105.14368* (2021) (Cited on page 2).

[Bel+19] Mikhail Belkin, Daniel J Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854 (Cited on page 2).

[BHM18] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2300–2311 (Cited on page 2).

[BSW20] Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp. "Interpolation under latent factor regression models". In: *arXiv e-prints* (2020), arXiv–2002 (Cited on page 2).

[CL21] Niladri S Chatterji and Philip M Long. "Finite-sample analysis of interpolating linear classifiers in the overparameterized regime". In: *Journal of Machine Learning Research* 22.129 (2021), pp. 1–30 (Cited on page 2).

[CL20] Geoffrey Chinot and Matthieu Lerasle. "On the robustness of the minimum ℓ_2 interpolator". In: *arXiv preprint arXiv:2003.05838* (2020) (Cited on page 2).

[CLG20] Geoffrey Chinot, Matthias Löffler, and Sara van de Geer. "On the robustness of minimum-norm interpolators". In: *arXiv preprint arXiv:2012.00807* (2020) (Cited on page 2).

[Cor+09] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT Press, 2009 (Cited on page 3).

[DKT19] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. "A model of double descent for high-dimensional binary linear classification". In: *arXiv preprint arXiv:1911.05822* (2019) (Cited on page 2).

[GSD20] Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. "The implicit bias of depth: how incremental learning drives generalization". In: *International Conference on Learning Representations*. 2020 (Cited on page 35).

[Gun+18a] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nathan Srebro. "Characterizing implicit bias in terms of optimization geometry". In: *International Conference on Machine Learning*. 2018, pp. 1832–1841 (Cited on page 2).

[Gun+18b] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nathan Srebro. "Implicit bias of gradient descent on linear convolutional networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 9461–9471 (Cited on pages 2, 35).

[Gun+17] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathna Srebro. "Implicit regularization in matrix factorization". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6151–6159 (Cited on pages 2, 35).

[Has+19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. "Surprises in high-dimensional ridgeless least squares interpolation". In: *arXiv preprint arXiv:1903.08560* (2019) (Cited on page 2).

[HMX21] Daniel Hsu, Vidya Muthukumar, and Ji Xu. "On the proliferation of support vectors in high dimensions". In: *International Conference on Artificial Intelligence and Statistics*. 2021, pp. 91–99 (Cited on page 2).

[JRG21] Meena Jagadeesan, Ilya Razenshteyn, and Suriya Gunasekar. "Inductive Bias of Multi-Channel Linear Convolutional Networks with Bounded Weight Norm". In: *arXiv preprint arXiv:2102.12238* (2021) (Cited on page 35).

[JT19] Ziwei Ji and Matus Telgarsky. "The implicit bias of gradient descent on nonseparable data". In: *Conference on Learning Theory*. 2019, pp. 1772–1798 (Cited on page 2).

[KLS20] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. "The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization". In: *Journal of Machine Learning Research* 21 (2020), pp. 1–16 (Cited on page 2).

[Koe+21] Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. "Uniform convergence of interpolators: Gaussian width, norm bounds, and benign overfitting". In: *arXiv preprint arXiv:2106.09276* (2021) (Cited on page 2).

[KL17] Vladimir Koltchinskii and Karim Lounici. "Concentration inequalities and moment bounds for sample covariance operators". In: *Bernoulli* 23.1 (2017), pp. 110–133 (Cited on page 34).

[LZG21] Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. "Towards an understanding of benign overfitting in neural networks". In: *arXiv preprint arXiv:2106.03212* (2021) (Cited on page 2).

[LR20] Tengyuan Liang and Alexander Rakhlin. "Just interpolate: Kernel "ridgeless" regression can generalize". In: *Annals of Statistics* 48.3 (2020), pp. 1329–1347 (Cited on page 2).

[LRZ20] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. "On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels". In: *Conference on Learning Theory*. 2020, pp. 2683–2711 (Cited on page 2).

[LS20] Tengyuan Liang and Pragya Sur. "A precise high-dimensional asymptotic theory for boosting and min-ℓ_1-norm interpolated classifiers". In: *arXiv preprint arXiv:2002.01586* (2020) (Cited on page 2).

[MM19] Song Mei and Andrea Montanari. "The generalization error of random features regression: Precise asymptotics and the double descent curve". In: *Communications on Pure and Applied Mathematics* (2019) (Cited on page 2).

[Mon+19] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. "The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime". In: *arXiv preprint arXiv:1911.01544* (2019) (Cited on page 2).

[Mut+20a] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel J Hsu, and Anant Sahai. "Classification vs regression in overparameterized regimes: Does the loss function matter?" In: *arXiv preprint arXiv:2005.08054* (2020) (Cited on page 2).

[Mut+20b] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. "Harmless interpolation of noisy data in regression". In: *IEEE Journal on Selected Areas in Information Theory* (2020) (Cited on page 2).

[NSS19] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. "Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate". In: *International Conference on Artificial Intelligence and Statistics*. 2019, pp. 3051–3059 (Cited on page 2).

[NDR20] Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. "In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors". In: *International Conference on Machine Learning*. 2020, pp. 7263–7272 (Cited on page 2).

[NTS15] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. "In search of the real inductive bias: on the role of implicit regularization in deep learning." In: *International Conference on Learning Representations (Workshop)*. 2015 (Cited on page 2).

[RC20] Noam Razin and Nadav Cohen. "Implicit regularization in deep learning may not be explainable by norms". In: *Advances in Neural Information Processing Systems* 33 (2020) (Cited on page 35).

[RV09] Mark Rudelson and Roman Vershynin. "Smallest singular value of a random rectangular matrix". In: *Communications on Pure and Applied Mathematics* 62.12 (2009), pp. 1707–1739 (Cited on page 40).

[RV10] Mark Rudelson and Roman Vershynin. "Non-asymptotic theory of random matrices: extreme singular values". In: *Proceedings of the International Congress of Mathematicians 2010*. 2010, pp. 1576–1602 (Cited on page 3).

[RV13] Mark Rudelson and Roman Vershynin. "Hanson-Wright inequality and sub-Gaussian concentration". In: *Electronic Communications in Probability* 18 (2013), pp. 1–9 (Cited on page 29).

[RV15] Mark Rudelson and Roman Vershynin. "Small ball probabilities for linear images of high-dimensional distributions". In: *International Mathematics Research Notices* 2015.19 (2015), pp. 9594–9617 (Cited on page 40).

[Sou+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. "The implicit bias of gradient descent on separable data". In: *Journal of Machine Learning Research* 19.1 (2018), pp. 2822–2878 (Cited on page 2).

[TB20] Alexander Tsigler and Peter L Bartlett. "Benign overfitting in ridge regression". In: *arXiv preprint arXiv:2009.14286* (2020) (Cited on pages 2, 3, 6–8, 32).

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018 (Cited on pages 36, 37, 40).

[WT21] Ke Wang and Christos Thrampoulidis. "Benign overfitting in binary classification of gaussian mixtures". In: *International Conference on Acoustics, Speech and Signal Processing*. 2021 (Cited on page 2).

[Woo+20] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. "Kernel and rich regimes in overparametrized models". In: *Conference on Learning Theory*. 2020, pp. 3635–3673 (Cited on page 35).

[WX20] Denny Wu and Ji Xu. "On the optimal weighted ℓ_2 regularization in overparameterized linear regression". In: *Advances in Neural Information Processing Systems* 33 (2020) (Cited on page 2).

[Yan+20] Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. “Rethinking bias-variance trade-off for generalization of neural networks”. In: *International Conference on Machine Learning*. 2020, pp. 10767–10777 (Cited on page 2).

[YKM21] Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. “A unifying view on implicit bias in training linear neural networks”. In: *International Conference on Learning Representations*. 2021 (Cited on pages 2, 35).

[Zha+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. “Understanding deep learning requires rethinking generalization”. In: *International Conference on Learning Representations*. 2017 (Cited on page 2).